Large Visual Language Models for Supervised Multimodal Classification in Political Science Research

Andreu Casas[*] Freek Cool[†] Anne Rasmussen[‡]

Preliminary incomplete draft prepared for the 2025 Annual Meeting of the European Political Science Assocation. Please do not circulate.

Abstract

Political scientists often study multimodal data containing both text and visuals, such as news articles, information from websites, and social media posts. A common practice for large-scale projects is to use supervised machine learning to identify theoretical concept of interest in the data, by manually annotating a subset of the data, using the annotations to train a machine learning classifier, and using the trained classifier to predict the concept in the remaining unannotated data. Past research has mostly used textual input features when training supervised classifiers for multimodal data, yet recent advances in Large Visual Language Models (VLMs) facilitate leveraging both text and visual features. However, little is known about whether, and the conditions under which, VLMs can help boost performance and improve measurement. In this paper we use two original dataset of about 4,000 annotated YouTube videos, and 4,000 X posts, to compare the performance of a range of text-only (SVM, BERT, Llama2, Llama3), image-only (CNN), and multimodal models (CLIP and VLMs) across 10 common annotation tasks in political science research. Across the board we observe VLMs to outperform the other models, yet with some caveats. We discuss key findings and implications for practitioners.

[*] Royal Holloway University of London. Department of Politics, International Relations and Philosophy. Corresponding author: andreu.casas@rhul.ac.uk. This research has been possible thanks to a VENI grant from NWO (VI.Veni.211R.052, PI: Andreu Casas), and from two Small Computing Grants (EINF8284 and EINF53) and one Large Computing Grant (2024.044) from NWO.

[†] Vrije Universiteit Amsterdam

[‡] King's College London

1 Introduction

The amount of data available to political scientists has drastically increased in the last decades (Grimmer, Roberts, and Stewart, 2021). For example, scholars interested in the moderation of political content on social media can track millions of posts close to real time (King, Pan, and Roberts, 2017; Casas, 2024), those studying gender stereotypes in elections can access large amounts of campaign ads (Neumann, Franklin Fowler, and Ridout, 2022), and those studying news framing can access large amount of news content (Jungblut et al., 2024). Parallel advances in computational resources and methods also mean that today researchers can more easily analyze such large amounts of data (Grimmer, Roberts, and Stewart, 2021).

Supervised machine learning is a common technique for those interested in identifying a defined concept in large amounts of data (Laurer et al., 2024): researchers elaborate a codebook describing the concept to be identified, use the codebook to manually annotate a subset of the data, use the annotated data to train a supervised machine learning model, and then use the trained model to classify the remaining data. Although many political scientists study large amounts of multimodal data (e.g. social media posts, campaign ads, and news articles), most studies only rely on input features from one modality, usually text, to train supervised classifiers predicting the concepts of interest.

Recent computational innovations, particularly Large Visual Language Models (VMLs), facilitate combining features from different data modalities when training supervised machine learning classifiers. However, we know very little about the conditions under which these multimodal solutions can help boost performance and improve measurement for political scientists. In this paper we contribute to the machine learning literature in political science by comparing the performance of text-only (SVM, BERT, Llama2, Llama3), image-only (CNN), and multimodal classifiers (image and text: CLIP and two VLMs, Idefics2 and Idefics3). We leverage two original dataset of about 4,000 YouTube videos from politically-relevant channels in the US, and 4,000 X messages posted by interest groups in the US, Ireland, and Denmark. *All* these videos and posts are multimodal:

they have both text and visuals data. We compare the performance of the classifiers across ten common annotations tasks in political science research, six annotation tasks for the YouTube dataset, and for for the X posts

On average, we observe the VLMs to outperform all the other models, with some caveats. VLMs do particularly well in scenarios where rich image data is available for inference (e.g. YouTube videos with many image frames, compared to X posts with only a single image), and with unbalanced data, such as predicting rare instances (e.g. hateful content and misinformation). But the boost in performance is less pronounced when fewer visual data is available and when working with more balanced data. Our goal is to provide insightful information to researchers working with large amounts of multimodal data, regarding the conditions under which they may benefit from using a VLM v. other kinds of machine learning models, as well as what to take into account when making such decision.

Although today there are many*proprietary* Large Language Models (LLMs) and VLMs available to researchers, in this paper we only use open-source options. This is a deliberate decision. For replicability and reproducibility purposes, we advocate for researchers to only use open-source models that themselves or others can use in the future to reproduce and replicate their work – particularly when building key analytical measures. Proprietary models can be re-trained, or be taken down, at any time – jeopardizing the robustness of academic findings.

2 Supervised Classification in Political Science

To be completed. Literature review on the use of supervised classifiers (text-only, image-only, multimodal) in political science; and how an increase in multimodal analyses can shape several polisci (sub)fields.

3 Data

We conduct our analyses on two datasets manually annotated in-house for two substantive projects. A dataset of about 4,000 YouTube videos from channels posting about US politics, and a dataset of about 4,000 X messages with images posted by interest groups from the US, Ireland, and Denmark. We annotated these datasets for 10 theoretical quantities of interest to many political scientists, such as whether the content is about politics, the particular political topic, and whether it contains hateful content.

3.1 YouTube Dataset

For more than one year leading to the 2024 US election, we tracked about 20,000 YouTube channels that post about US politics. The substantive goal of the project was to study content moderation in the platform during the campaign, such as what channels and/or videos were suspended. We used a snowballing technique to identify these politically-relevant channels. The starting point was an extensive list of YouTube channels of media organizations (e.g. New York Times, Fox News, etc. N = 105) and of US politicians (members of the 118th Congress, President Biden, and former president Trump. N = 184). Then, in April 2023, we identified users commenting or replying to videos from these elite channels (N = 26,353 users), and collected the full list of channels to which these "politically-interested" users were subscribed (N = 7,334,005 channels – 1,624,136 of them unique). We narrowed the list to focus on the most relevant channels: those to which at least 10 of the users were subscribed (N = 92,653). Finally, we trained a language model (BERT) that used channel descriptions and tags to predict whether they posted (at least sometimes) about politics. Through this process we identified 20,054 politically-relevant channels that we added to the original list of elite channels, for a total of 20,343.

For this paper we sampled 3,884 English videos sent by these politically-relevant channels between May 2023 and March 2024. Apart from collecting channel- and video-level metadata, for each video we collected the transcript and frames. For the transcript,

Table 1: Six annotation tasks for Youtube videos from politically-relevant channels.

Task	Description	Values	N	%
US Politics	Whether the video is about, or relevant to, US politics	0	1,935	49.8%
		1	1,945	50.2%
		N	3,880	100.0%
Hateful	Whether the video contains hateful language/behavior	0	3,431	88.4%
		1	449	11.6%
		N	3,880	100%
Idology	The ideological leaning of the video	Neutral	238	21.7%
		Conservative	476	41.9%
		Moderate	176	15.5%
		Liberal	247	20.9%
		N	1,137	100%
Typology	Type of video	Campaign	16	1.4%
		Educational	61	5.2%
		Satire	73	6.2%
		Low-Qual News	108	9.2%
		High-Qual News	332	28.4%
		Opinion	581	49.6%
		N	1,171	100%
Misinfo	Whether the video contains misinformation/conspiracies	0	3,672	94.5%
		1	208	5.5%
		N	3,880	100%
Topic	Main *Comparative Agendas Project* topic discussed in the video	Economy	51	2.7%
		Civil Rights	501	26.2%
		Healthcare	56	2.9%
		Agriculture	2	0.1%
		Labor	30	1.6%
		Education	23	1.2%
		Environment	27	1.4%
		Energy	19	1.0%
		Immigration	43	2.3%
		Transportation	13	0.7%
		Law & Crime	121	6.3%
		Social Welfare	22	1.2%
		Housing	35	1.8%
		Commerce	54	2.8%
		Defense	76	4%
		Technology	80	4.2%
		Foreign Trade	9	0.5%
		Intl. Affairs	253	13.2%
		Gov. Operations	428	22.4%
		Public Lands	9	0.5%
		Gun Control	59	3.1%
		N	1,911	100%

we collected the close captions provided by Youtube when available via the API (about 55% of the time), and used a speech-to-text model (Whisper from OpenAI: Radford et al.

(2023)) to obtain transcripts for the remaining cases. For the frames, we first extracted one frame per second of video, and then we used the cosine similarity between all possible pairs of frame embeddings to deduplicate them and only keep distinct frames.

Three research assistants manually annotated the videos for the six annotation tasks listed in Table 1.[1] We annotated all videos for three binary variables: whether the videos were about (or relevant to) US politics (*US Politics*), whether the video contained hateful language or behavior (*Hateful*), and whether the video spread misinformation or conspiracies (*Misinfo*). Then, we annotated, for three additional categorical variables, the videos that we had coded as being about *US Politics* (N = 1,945): the ideological leaning of the video (*Ideology*: Neutral, Conservative, Moderate, and Liberal), the *Typology* of the video (e.g. whether it was a campaign video, low or high quality news, or opinion), and the *Topic* (following the topic classification of the *Comparative Agendas Project* (Jones et al., 2023)).[2]

It is important to note that these 4,000 videos were not selected at random and are not representative of the population of videos relevant to US politics on Youtube. Given our interest in studying the moderation of politically-relevant videos, we over-represented videos based on two factors. First, we over-sampled on videos that had been suspended or were no longer available (N = 1,195: 30% of our annotated videos); and second, we over-sampled on videos likely to contain hateful language (N = 1,683: 43%, with a toxicity score above 50% based on the off-the-shelf model `detoxify`). Given that videos containing hateful language and misinformation are likely to be rare, we wanted to maximize our chances of finding some, to feed a larger number of true positive to the machine learning classifiers with the intention of improving performance. Table 1 contains information about the distribution of these annotations, and Appendix C contains a description of each variable and category.

[1] Working on providing Inter-Rater Reliability statistics.

[2] We added the *Ideology* and *Typology* task once the annotation process had already started and are based on a slightly smaller subset of the data, hence the lower N. For the three categorical variables, we exclude a few "tough" observations that annotators had a hard time to code.

3.2 X Dataset

We assembled a list of interest groups on X (formerly Twitter) for a substantive project on digital advocacy: 8,480 listed in the Washington Representative Directory[3] and with an active X account (which we tracked between September 2021 and July 2022 using the Twitter REST API), and 7,724 listed in the EU Transparency Register[4] (for which we collected all messages sent between January 2007 and July 2023).

Table 2: Four annotation tasks for X posts from interest groups.

Task	Description	Values	N	%
Lobbying	Whether the post is about lobbying	0	3,130	70.7%
		1	1,304	29.3%
		N	4,449	100.0%
Lobbying Type	The type of lobbying activity	Indirect	52	4%
		Direct	289	22.2%
		Other	962	73.8%
		N	1,303	100%
Lobbying Info	Whether posts contains policy-relevant information	0	950	73%
		1	352	27%
		N	1,302	100%
Non-Lob Type	Type of non-lobbying post	Organizational	180	5.8%
		Other	566	18.1%
		Marketing	700	22.4%
		Community	1,684	53.8%
		N	3,130	100%

For this project we sampled 4,449 tweets from interest organizations from the United States, Ireland, and Denmark that we had identified as having text but also an image. We had already collected the text using the Twitter REST API, and we programmed a regular web scraper to collect the images included in the tweets. We translated all non-English messages into English using an open-source deep learning translation model.[5]

Two research assistants manually annotated the tweets for the four annotation tasks listed in Table 2.[6] Two binary variables: whether the post was about *Lobbying*, and if so, whether it provided policy-relevant information (*Lobbying Info*). And two categorical variables. For those previously coded as being about lobbying, we built on the interest

[3] https://www.washingtonrepresentatives.com/Washington-Representatives

[4] https://transparency-register.europa.eu/

[5] https://github.com/UKPLab/EasyNMT.

[6] Two-coder Krippendorff's alpha: *Lobbying* (0.80), *Lobbying Type* (0.748), *Lobbying Info* (0.713).

group literature (Kollman, 1998; Dür and Mateo, 2016) to code them for *Lobbying Type* (Indirect, Direct, or Other). Those coded as not being about lobbying, we also annotated them for the *Non-Lobbying Type* (Organizational, Marketing, Community, or Other). Table 2 contains information about the distribution of these annotations, and Appendix C contains a description of each variable and category.

4 Methods

We train and assess the out-of-sample performance of eight machine learning models in predicting each of the ten annotation tasks described above: SVM, BERT, CNN, Llama2, Llama3, CLIP, Idefics2, and Idefics3. We chose them to range from traditional and "less-sophisticated" models such as SVM, to more recently-released open-source LLMs and VLMs, such as Llama3 and Idefics3. We chose some unimodal classifiers to use as baselines. For example, SVM, BERT, Llama2 and Llama3 are text-only models, and our CNN architecture is designed as an image-only model. Our main interest is the performance of multimodal classifiers that can take both text and image features as input, particularly VLMs. Below we describe and motivate each of these machine learning models in more detail.

4.1 Classifiers

4.1.1 Text Only

Support Vector Machine

Support Vector Machine (SVM) is a classic machine learning algorithm often used for classification tasks. SVM performs well on high dimensional spaces and can deal with data points with numerous input features. This makes it particularly suitable for text classification, although they can be used to classify any kind of data, including image data (Cortes and Vapnik, 1995). Several work in political science has used SVMs and similar n-gram based approaches for training supervised text classifiers (Collingwood and

Wilkerson, 2012; Hemphill, Russell, and Schöpke-Gonzalez, 2021) – making SVM a great baseline model of interest.

First, the text corpus of interest needs to be transformed into a document-feature matrix, where each row is a document, each cell is a text feature, and each cell contains information about whether the the text feature is present in the document. SVM is designed to find the hyperplane (line) that divides this multidimensional space in a way that maximizes the separation between a set of known document classes. In other words, the algorithm aims to learn what text features in that particular corpus (words, or n-grams) are predictive of each target class. Hence, compared to more recent models, SVM does not have any prior "knowledge" of language.

We use the following protocol to train SVMs for text classification. First, we transform all text to lower case (so that "immigrant" and "Immigrant" are seen as the same word), remove stopwords (common English words such as "a" or "I", as they are uninformative for the tasks at hand), and then create a document feature matrix indicating which unique unigrams (i.e. words) are present in each video transcript. We disregard very (in)frequent unigrams (present in >99% or <1% of the video transcripts) as they can slow down computation while being mostly uniformative. We speed up computation further by also limiting the size of the document feature matrix to the most prevalent 20,000 unigrams in the corpus. Finally, we train/validate a SVM with a linear kernel on the resulting matrix, using the sklearn package in python.

BERT

Pre-trained language models (LMs) have dominated the Natural Language Processing (NLP) literature in the last decade (Vaswani et al., 2017; Radford et al., 2018; Devlin et al., 2018) and has been widely used to train supervised text classifiers in political science research (Casas et al., 2025; Timoneda and Vera, 2025). Compared to traditional ngram-based machine learning models like SVM, pre-trained LMs separate the process of learning *language* and *task* representations (Laurer et al., 2024). First, LMs learn language representations (stacked layers of vectors known as transformers) via unsupervised tasks (known as *pre-training*), such as masked language modeling and next sentence prediction. LMs like BERT (Devlin et al., 2018) have learned these language representations from large amounts of text, such as the BookCorpus with more than 7,000 books and 800 million words (Zhu et al., 2015), and the enitre English Wikipedia (2,5 billion words). Then, the pre-trained LM can be used to learn any downstream task of interest (e.g. predicting which videos are about *US Politics*) by adding a new output/prediction layer to the model, and *fine-tuning* the new layer and the pre-trained parameters with new data annotated for the task at hand.

In this paper we use one of the most used LMs in NLP and political science: `bert base uncase`. We fine-tune ("train") the pre-trained model for each of the ten downstream tasks by first adding a new output layer to the model (initialized at random) that is of equal size of the number of output classes (e.g. two for the first three binary tasks listed in Table 1). Then we use our annotated data to fine-tune the new output layer and the pre-trained parameters. We do this in python, using the `pytorch` library and a batch size of 64. We train each model 3 times, each of them using a different learning rate (0.0001, 0.00005, 0.00001), as previous literature identifies the learning rate as the key hyper-parameter to tune for maximizing performance (Laurer et al., 2024). BERT models can only process inputs of 512 tokens at a time. This 'context length' is enough for processing all X messages. However, for the YouTube data, we only use the first 512 tokens in each video transcript.

Llama2 and Llama3 Instruction Models

In the last few years, instruction-based Large Language Models (LLMs) such as ChatGPT have yet again revolutionized the NLP literature. They differ from previous LMs in two main ways. First, LLMs are *pre-trained* on a much larger amount of text. For example, Llama2 is pre-trained on 2 trillion tokens from large amounts of publicly available sources, and Llama3 on 15 trillion tokens. This is 12 and 75 times more tokens than the BERT base model, respectively. Second, the instruction version of Llama2 and Llama3 models have already been *fine-tuned* to perform conversational tasks on a wide range of topics, by using large open-source and proprietary question-answer datasets (Touvron et al., 2023). Researchers can use these instruction-based models for supervised classification, by providing them with an input text (e.g. the transcript of a Youtube video) and then asking the model to identify the concept of interest (e.g. "Does this video transcript contain hateful language?").

In this paper we assess the performance of the instruction version of Llama2 and Llama3, two of the largest open-source LLMs to date. We assess how they perform off-the-shelf ("zero shot"), and also after using our annotated data for fine-tuning the models further for each of the ten tasks at hand. Given the size of these models, and that large GPU required for fine-tuning the largest versions of these models are not frequently available among political scientists, here we use the smallest version of each of these chat models (7B parameters for Llama2 and 8B parameters for Llama3). Additionally, when fine-tuning the models we use quantization to reduce the decimal points of all model parameters from 16 to 4. In Appendix C you can find a list of the prompts we developed to automatically annotate the data. Llama2 and Llama3 have a context length (including the input text and the instructions/prompt) of 4,096 and 8,000 tokens, respectively. For this reason, as well as to avoid running out of VRAM in the GPU and to keep the computational results as comparable as possible to the baseline BERT model, for both models we only use up to the first 1,000 words of the video transcripts for the YouTube dataset.

4.1.2 Image Only

Convolutional Neural Network

Convolutional Neural Networks (CNN) are widely used for supervised computer vision tasks, such as object detection and recognition (Girshick, 2015). For machine learning purposes, images are often represented as three-dimensional matrices, each dimension containing information about the intensity of red, green and blue (RGB) in each pixel. It is computationally intensive to feed these three-dimenional inputs into neural networks with fully-connected layer. Instead, to simplify and speed-up the computation process, CNNs rely on convolutional layers where matrices of weights are only connected a single region and color channel at a time. See Webb Williams, Casas, and Wilkerson (2020) and Torres and Cantú (2022) for a detailed explanation of CNNs and some of their applications for supervised image classification in political science research. Most work relies on fine-tuning a CNN that has already been pre-trained on an object recognition task (e.g. the 1,000 ImageNet classes), and on large amounts of annotated images (e.g. 1.28 million for the ResNet model used in this paper, (He et al., 2015)). These pre-trained models have learned already a good deal about how to extract substantively relavant information from images. Similar to pre-trained LMs such as BERT, one can replace the last output layer of a pre-trained CNN, with a new output layer of size equal to the number of classes in a target variable of interest; and then fine-tune the pre-trained CNN further with new annotated data in order to perform a new classification task, such as identifying politically-relevant Youtube videos from frames.

For each of the ten target variables, we use our annotated data to fine-tune a ResNet-50 CNN (He et al., 2015). We follow He et al. (2015)'s protocol when processing the images for fine-tuning (training and testing) the model, making sure all images are of 224x224, and following the correct color standardization. The results of this model will serve as baseline, and to help us understand how much visual (*v.* textual) features contribute to building accurate supervised classifiers in these different contexts.

4.1.3 Multimodal

CLIP

Contrastive Language-Image Pre-training (CLIP) is an open-source modeling strategy proposed by Open AI in 2021 for working with multimodal data (Radford et al., 2021). In a nutshell, this approach aims to generate joined text-image enconders that place text and images in the same embedding space. Previous attempts at using joint embeddings for multimodal modeling proposed concatenating a pre-trained text embedding (e.g. a 768-size BERT embedding), and a pre-trained image embedding (e.g. a 768-size ViT/16 embedding, (Dosovitskiy et al., 2021)), for each text-image pair under analysis (e.g. for a resulting 1536-size text-image embedding) – to then stack the joint embeddings for all image-pairs and feed them to a supervised model (e.g. logistic regression) predicting annotated class labels at the image-pair level. However, the models used for generating the text and image embeddings (commonly known as *encoders*) were trained separately.

In CLIP, the text and image encoders are trained jointly. In particular, the model is self-trained using 400 million image and image-caption pairs scraped from the internet. Initial text and image embeddings are generated using common pre-trained encoders in the literature, and then they are jointly fine-tuned by predicting the correct text caption for a given image.

We use the pre-trained CLIP model to generate text-image embeddings for our data, stack the embeddings, and for each of the ten annotation tasks, we *train* a logistic regression that uses these stacked joint CLIP embeddings as input, and the annotated class labels as outcome. One relevant limitation of this method is the small context length of the text encoder: 77 tokens. We hence only use the first 77 tokens of the YouTube transcript, and X messages, in our computation experiments. Additionally, similar to the image-processing for the CNN, all input images need to be re-sized to 224x224.

Large Visual Language Models: Idefics2 and Idefics3

Idefics is an open-source Visual Language Model (VLM) (Laurençon et al., 2023) that, in its instruction-based format, can generate textual answers to interleaved text-image inputs. Similar to CLIP, pre-trained text (Mistral-7B for Idefics2, and Llama3.1 for Idefics3) and image (SigLIP-SO400M) encoders are used for generating an initial text and image embedding representation of the input data that was used for training the original model. Then, joint text-image encoders for the base model are jointly self-learned through a variety of tasks and datasets, including image-caption matching, PDF OCR extraction, etc. Finally, this self-trained base model is fine-tuned for chat/instruction tasks using a large public question-answer dataset (The Cauldron).[7] The main difference

[7] https://huggingface.co/datasets/HuggingFaceM4/the_cauldron

between Idefics2 and Idefics3 is that they use a different text encoder for generating the original text embeddings, and that the latter is trained on a larger amount of data both at the base self-learning stage and at the subsequent chat/instruction phase.

For each of the ten target variables, we assess the performance of the instruction version of these models (Idefics2-8b and Idefics3-8B-Llama3). In line with the Llama experiments, we assess the performance of these models off-the-shelft (zero shot), and also after fine-tuning the models with our annotated data (using the same prompts reported in Appendix C). We also use quantization to round all model weights to 4 decimal numbers, and limit our YouTube transcripts to the first 1,000 tokens.

4.2 Set Up

In here we describe the main computation set up for training/fine-tuning and assessing the performance of the suite of supervised machine learning models described above.

4.2.1 Train-Test-Validation Split

We split (80/20) our two annotated annotated datasets into a train and test sets. When training/fine-tuning models, we only use the train set. As is common practice, we use the entire train set when fitting the SVM model. However, for the remaining models, we divide the training process into N epochs where we further split (80/20) the main train set into a train and validation set. In each epoch, the validation set is only used to assess model performance and not for updating the model weights. We use this validation performance when deciding the optimal number of training epochs for each model. Finally, we use the completely untouched test set to re-assess how well the model performs at that particular epoch. In the paper we only report results based on the test set. The train/test splits are constants across all computational experiments. The train/validation splits are random and not constant across experiments. The results presented in this preliminary draft are based on a single train/test split (random seed = 1). In the final version of the manuscript we aim to report results based on 3 different train/test splist (3-fold cross-validation).

4.2.2 Early Stopping

Based on some preliminary experiments, we decided on a minimum number of epochs for training each of models: BERT (30 epochs), CNN (30), Llama2 and Llama3 (15), Idefics2 and Idefics3 (5). Then, for each of the computational experiments, we explored the validation F1-score after this number of epochs, and trained the model for some additional ones if the model had not finished learning, until convergence. We report the progress, and show the convergence, of all the validation F1-scores in Appendix A.

4.2.3 One(label)-to-many(frames)

Relevant for image-only (CNN) and multimodal models (CLIP and VLMs). For the YouTube dataset, for a given unit of analysis (video) we have one annotation/label (and text input –transcript), but more than one image input (unique video frames). Hence, for training/fine-tuning and for assessing these models, we propagate the video-level labels to each transcript-frame from that video. The videos included in the train v. test split are the same as in the other models but during training we move to the described frame-level structure. However, in the validation stage, we assess performance at the video level, keeping the model assessment comparable across all experiments.

4.2.4 Data Balancing and Mini-Epochs

In some preliminary experiments, we observed very low performance for two binary annotation tasks for which we had very unbalanced data: *Hateful* and *Misinformation*, with only around 10 and 5% of true positives, respectively. To improve performance, we decided to implement a protocol for re-balancing the data during training. In each epoch, we created as many mini-epochs possible with all true positives in the train set, and the same amount of (non-duplicated) true negatives. For example, in a hypothetical epoch with 100 data units, 10 true positives and 90 true negatives, we would create 9 mini-epochs with the same 10 true positives and 10 distinct true negatives. Finally, instead of using all resulting mini-epochs in each epoch, to avoid overfitting (given the duplication of the true positives across mini-epochs), we randomly sampled 30% of them in each epoch (this random sampling was not constant across computational experiments).

4.2.5 Hyper-Parameter Fine-Tuning

We explored the performance when varying two key hpyer-parameters. '

- **Learning Rate**. For BERT models, we repeated all computational experiments using these 3 different learning rates: 0.0001, 0.00005, and 0.00001. For the CNN, Llama2, Llama3, Idefics2, and Idefics3: 0.0005, 0.0002, 0.0001.

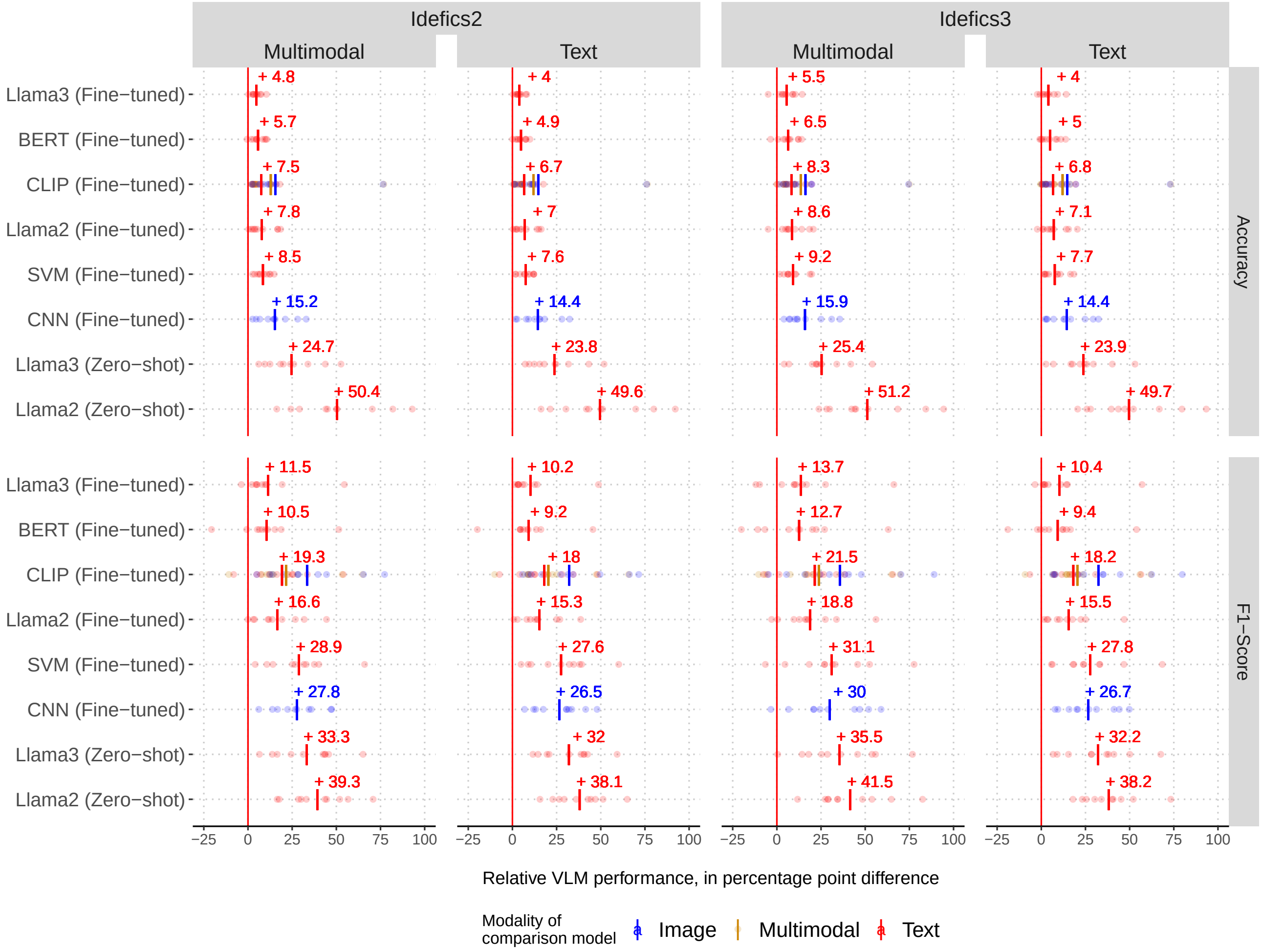
- **Image Input**. For VLMs (Idefics2, and Idefics3), the number of images to include in the prompt, from 1 to a maximum of 10.

4.2.6 Assessing Performance

In this section we only report model performances based on the untouched test set, which is constant across all computational experiments. For each experiment, we report the test performance for the epoch with the best validation F1-Score. We focus on calculating and reporting four main common statistics in machine learning research. **Accuracy**. Percentage of correct predictions. **Precision**. The proportion of observations predicted as a given class that truly belong to that class. **Precision**. The proportion of observations from a given class that are actually classified as such. **F1-score** (Micro or Weighted). Average of the Precision and the Recall (weighted by class size).

For the multimodal models (CLIP and VLMs), we always use both text and images when fine-tuning them, but we assess their validation and test performance three ways: using only text, only images, and both text and images. This is to explore whether cross-modality learning can help improve overall performance. VLMs can benefit from learning from more than one modality, but still generate more accurate predictions when relying on a single modality (Hu, Li, and Yin, 2025).

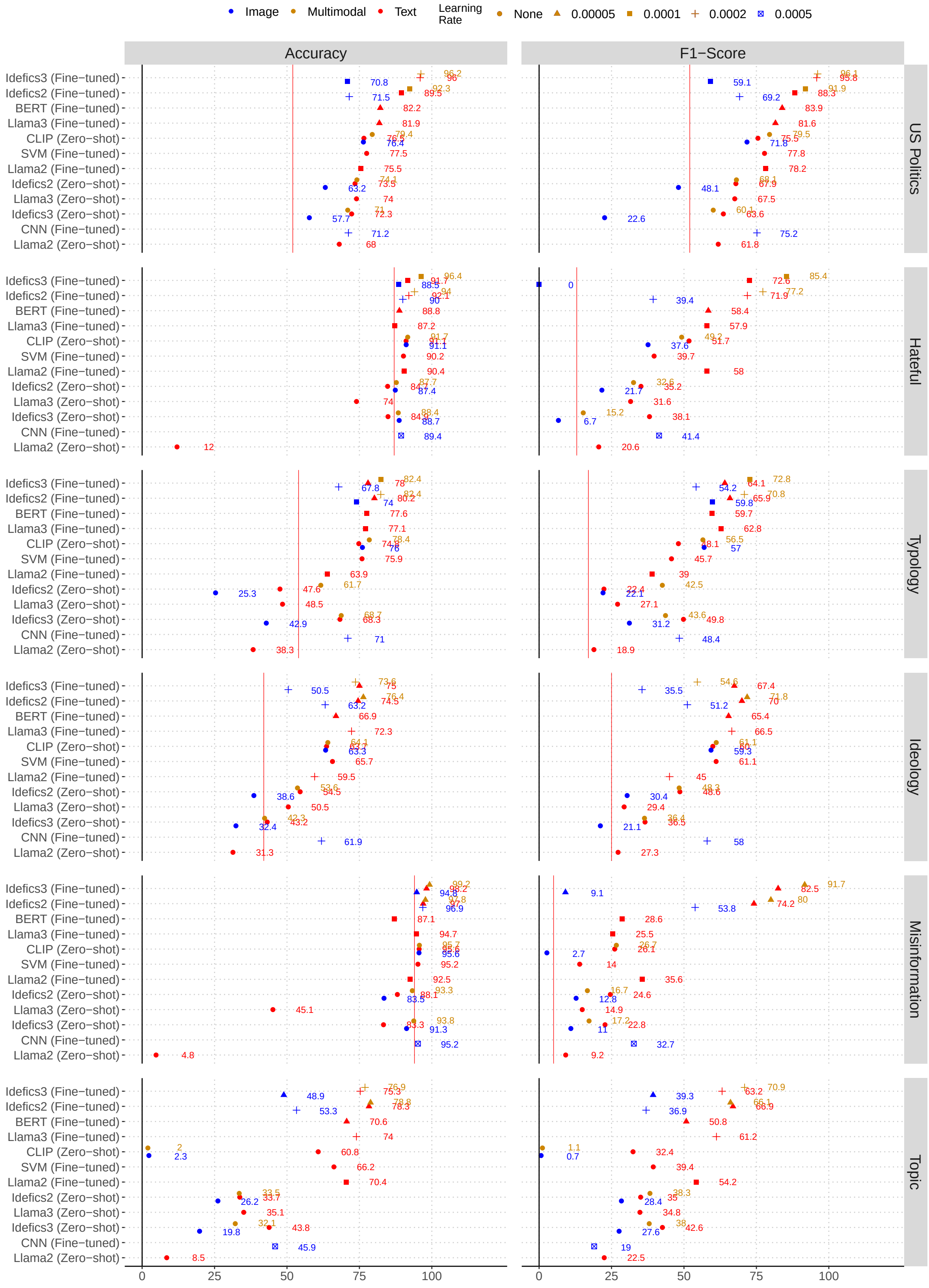
Figure 1: Average relative performance of Large Visual Language Models compared to other supervised machine learning models.



5 Results

In Figure 2 we show the Accuracy and F1-Score for all models for the six annotation tasks for the YouTube dataset, and in Figure 3 we report the model performances for the four annotation tasks for the X dataset. In Appendix B we show the full results for these two figures, including Precision and Recall. Additionally, in Figure 1 we combine the information in Figure 2 and 3 to calculate the average relative performance of VLMs compared to the other models (across datasets and tasks). We would like to emphasize X findings we believe are very valuable for political scientists interested in performing supervised classification of multimodal data.

Figure 2: Performance of a variety of machine learning models in predicting theoretical quantities of interest in YouTube videos.



(1) VLMs outperform all other models. In Figure 1 we see that when we average the relative performance of fine-tuned VLMs compared to the other models (across annotation tasks – each represented by a dot in the figure). For example, the Accuracy of the Idefics3 multimodal-based predictions are 25-50 percentage points (pp) better than off-the-shelf predictions from Llama3 and Llama2, 9.2pp better than a traditional ngram-based model such as SVM, 8.3pp better than a more "rudimentary" multimodal option (CLIP), 6.5pp better compared to BERT (arguably the state-of-the-art pre-ChatGPT), and 5.5pp better than an open-source LLM fine-tuned for the same tasks (Llama3). The latter is particularly relevant given that Llama3 is the text backbone for Idefics3.

(2) ... particularly at identifying infrequent classes. In Figure 1 we see VLMs to particularly outperform in terms of F1-Score. This is in part due to a ceiling effect (baseline Accuracy is already quite high for several of these annotation tasks given that a large proportion of cases belong to the modal class). However, we do observe in Figure 2 for example that the VLMs do substantially better at identifying rare classes (much higher F1-Score), such as predicting YouTube videos with *Hateful* content and *Misinformation.* This can also be a function of the visual content being key to identifying these quantities of interest. Yang, Davis, and Hindman (2023) for example find that a substantive amount of misinformation on social media platforms in visual in nature.

(3) ... but not always by much? In Figure 3 we observe VLMs, particularly Idefics2, to outperform other models, but the improvement in performance to not be as notable for a few annotation tasks, such as *Lobbying Info* (81.1% Accuracy and 66.7% F1-Score for Idefics2; compared to 80.7% and 65.3% for a fine-tuned Llama2).

(4) Larger VLMs pre-trained on more data do not always do better. On average, we see the increase in performance, compared to other models, to be more substantive for Idefics3 v. Idefics2. However, particularly in Figure 3 we see Idefics2 to do better than Idefics3 for some tasks, such as *Lobbying Info* and *Lobbying Type.*

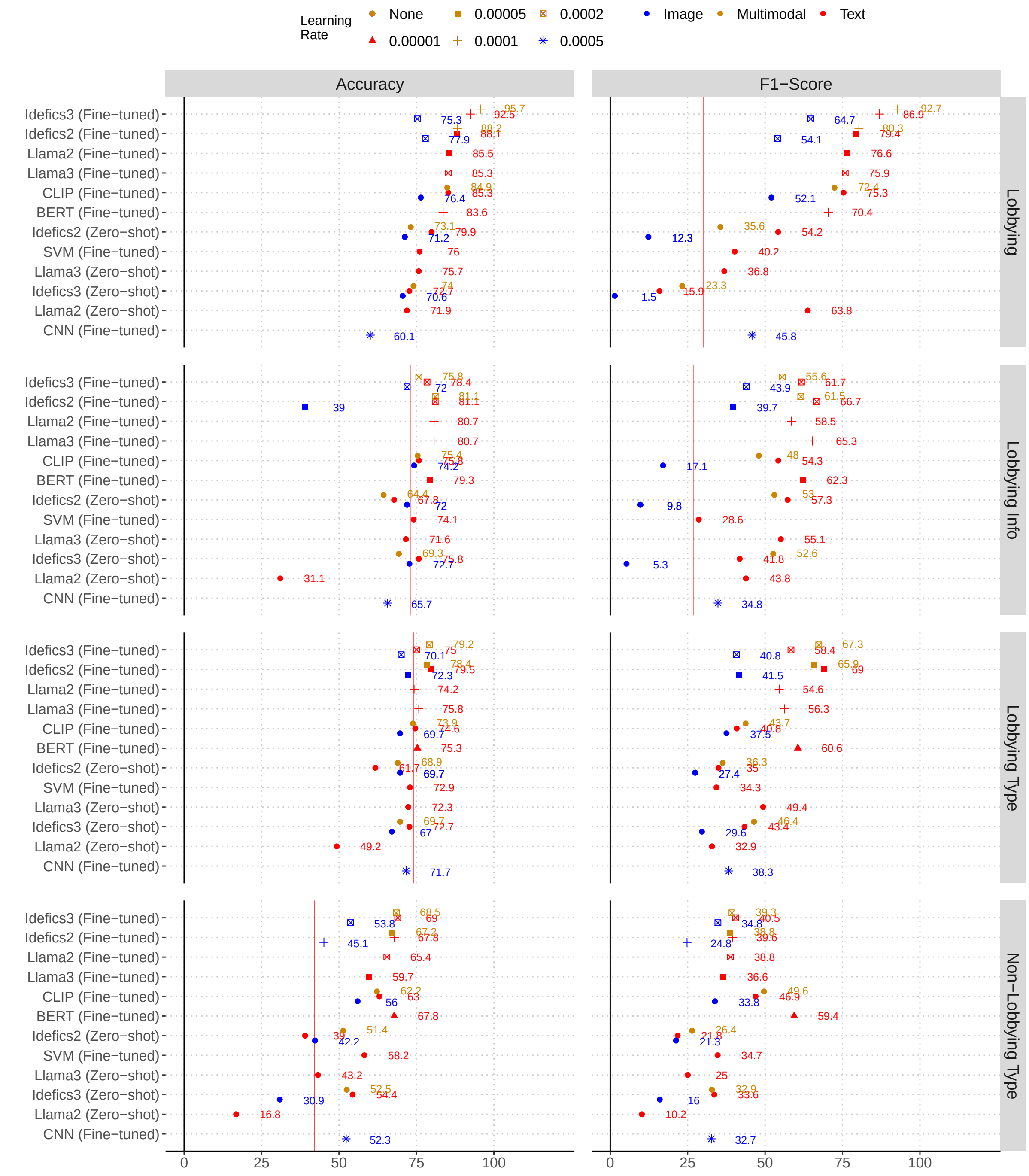
(5) Fine-tuning makes a big difference. For LLMs and VLms, we observe a drastic boost in performance across the board once we fine-tune these models using for these particular tasks, and using chunks of our own data. Particularly for complex and highly unbalanced target varaibles, such as identifying *Misinformation*, the off-the-shelf models do very poorly (e.g. Llama2: 9.2% v. 35.6% F1-Score; Idefics3: 22.8% v. 91.7% – see Figure 2).

(6) On occasion, VLMs perform better in text-only mode. In Figure 1 we see that on average VLMs perform better when using both text and images for inference. However, particularly in Figure 3 we observe text-only inferences, both for Idefics 2 and Idefics3, to sometimes outperform multimodal-based predictions. Here the comparison between the fine-tuned Llama3 and Idefics3 models is particularly relevant, as Llama3 is the text backbone of Idefics3. The higher performance of the text-only predictions from Idefics3, compared to the Llama3 predictions, indicates that the model benefits from cross-modality learning, even if it then performs better in text-only mode at the inference stage. One potential explanation for this is that multimodal VLM inference is more useful in image-rich scenarios, compared to text-only inference. For tasks involving the YouTube dataset the VLMs can use up to 10 images (frames) per observation, compared to only 1 image for the X dataset.

6 Discussion

Can Large Visual Language Models (VLMs) help improve the performance of supervised classification in political science projects that deal with multimodal (visual and textual) data? We compared the performance of two open-source VLMs (Idefics2-8B and Idefics3-8B-Llama3) to other machine learning classifiers (including two Large Language Models, LLMs: Llama2 and Llama3), across ten annotation tasks common in political science research (identifying political topics, hateful content, misinformation, etc.), and across

Figure 3: Performance of a variety of machine learning models in predicting theoretical quantities of interest in X posts from interest groups.



two datasets of YouTube videos and X posts with images.

Across the board, VLMs outperformed the other machine learning classifiers, although there is variation in the performance gain. On occasions, VLMs do not do much better than text-only LLMs, particularly when dealing with data that has a single image (X posts, v. YouTube videos with several image frames). We also noted that VLMs perform particularly well with infrequent classes (such as identifying hateful content and misinformation), that larger VLMs trained on more data do not always perform better, that fine-tuning the VLMs for your particular task makes a big difference (v. using it off the shelf), and that sometimes VLMs perform more accurate predictions in text-only mode (v. multimodal). We expect these and other lessons to be very useful to political scientists that work with large amounts of multimodal data, such as those studying video, website, newspaper, and/or social media data.

An important caveat to keep in mind is that VLMs are resource-intensive, and that storing (and working with) visual data takes substantially more space than only storing text. In here, we used H100s GPUs (with 80GB of VRAM) to fine-tune the LLMs and VLMs, and the computational experiments for a single target variable and model took days to run. Moreover, once a given model is trained (e.g. a VLM for identifying hateful content on YouTube videos) it can also take several days to generate predictions for large datasets. For example, based on our own experience, an Idefics3 VLM trained to identify hateful YouTube videos can generate predictions for about 150,000 unlabeled videos/day on an H100 GPU. Computing nodes with these powerful GPUs are not available in all academic institutions. The most popular commercial cloud computing serivces (e.g. AWS, Microsoft Azure, Google Cloud) charge about 10$/hour for using these kinds of GPUs, so about 240$/day. On top of that, there are environmental concerns regarding the carbon footprint of using these resource-intensive models.

A general takeaway from the results discussed above is that no one shoe fits all. In some cases, lighter and less resource-intensive models such as BERT can perform "close-enough" to a VLMs, potentially making it a preferable for truly large-scale projects where time, resources, and environmental impact can be a concern. We hope that the evaluation pipeline used here can also help others run some preliminary assessment of the kind of model that could work best for their use case.

References

Casas, Andreu. 2024. "The Geopolitics of Deplatforming: A Study of Suspensions of Politically-Interested Iranian Accounts on Twitter." *Political Communication* 41(3): 413–434.

Casas, Andreu, Oscar Stuhler, Julia Payson, Joshua A. Tucker, Richard Bonneau, and Jonathan Nagler. 2025. "Bottom Up? Top Down? Determinants of Issue-Attention in State Politics." *The Journal of Politics* .

Collingwood, Loren, and John Wilkerson. 2012. "Tradeoffs in accuracy and efficiency in supervised learning methods." *Journal of Information Technology & Politics* 9(3): 298–318.

Cortes, Corinna, and Vladimir Vapnik. 1995. "Support-vector networks." *Machine learning* 20: 273–297.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* .

Dosovitskiy, Alexey, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale.".

Dür, Andreas, and Gemma Mateo. 2016. *Insiders versus outsiders: Interest group politics in multilevel Europe.* Oxford University Press.

Girshick, Ross. 2015. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision.* pp. 1440–1448.

Grimmer, Justin, Margaret E. Roberts, and Brandon M. Stewart. 2021. "Machine Learning for Social Science: An Agnostic Approach." *Annual Review of Political Science* 24(Volume 24, 2021): 395–419.

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. "Deep Residual Learning for Image Recognition.".

Hemphill, Libby, Annelise Russell, and Angela M. Schöpke-Gonzalez. 2021. "What Drives U.S. Congressional Members' Policy Attention on Twitter?" *Policy & Internet* 13(2): 233–256.

Hu, Zhe, Jing Li, and Yu Yin. 2025. "When words outperform vision: Vlms can self-improve via text-only training for human-centered decision making." *arXiv preprint arXiv:2503.16965* .

Jones, Bryan D, Frank R Baumgartner, Sean M Theriault, Derek A Epp, Cheyenne Lee, and Miranda E Sullivan. 2023. "Policy Agendas Project: Codebook." *https://www.comparativeagendas.net/* .

Jungblut, Marc, Scott Althaus, Joseph Bajjalieh, Chung-hong Chan, Kasper Welbers, Wouter van Atteveldt, and Hartmut Wessler. 2024. "How shared ties and journalistic cultures shape global news coverage of disruptive media events: the case of the 9/11 terror attacks." *Journal of Communication* 74(3): 183–197.

King, Gary, Jennifer Pan, and Margaret E Roberts. 2017. "How the Chinese government fabricates social media posts for strategic distraction, not engaged argument." *American political science review* 111(3): 484–501.

Kollman, Ken. 1998. *Outside lobbying: Public opinion and interest group strategies.* Princeton University Press.

Laurençon, Hugo, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander M. Rush, Douwe Kiela, Matthieu Cord, and Victor Sanh. 2023. "OBELICS: An Open Web-Scale Filtered Dataset of Interleaved Image-Text Documents.".

Laurer, Moritz, Wouter van Atteveldt, Andreu Casas, and Kasper Welbers. 2024. "Less Annotating, More Classifying: Addressing the Data Scarcity Issue of Supervised Machine Learning with Deep Transfer Learning and BERT-NLI." *Political Analysis* 32(1): 84–100.

Neumann, Markus, Erika Franklin Fowler, and Travis N Ridout. 2022. "Body language and gender stereotypes in campaign video." *Computational Communication Research* 4(1).

Radford, Alec, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML).* PMLR.

Radford, Alec, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning.* PMLR pp. 28492–28518.

Radford, Alec, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. "Improving language understanding with unsupervised learning." *Technical report, OpenAI* .

Timoneda, Joan C., and Sebastián Vallejo Vera. 2025. "BERT, RoBERTa, or DeBERTa? Comparing Performance Across Transformers Models in Political Science Text." *The Journal of Politics* 87(1): 347–364.

Torres, Michelle, and Francisco Cantú. 2022. "Learning to See: Convolutional Neural Networks for the Analysis of Social Science Data." *Political Analysis* 30(1): 113–131.

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

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. "Attention is all you need." *Advances in neural information processing systems* 30.

Webb Williams, Nora, Andreu Casas, and John D. Wilkerson. 2020. *Images as Data for Social Science Research: An Introduction to Convolutional Neural Nets for Image Classification.* Elements in Quantitative and Computational Methods for the Social Sciences Cambridge University Press.

Yang, Yunkang, Trevor Davis, and Matthew Hindman. 2023. "Visual misinformation on Facebook." *Journal of Communication* 73(4): 316–328.

Zhu, Yukun, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision.* pp. 19–27.

Appendix A Validation learning progress

Figure A1: Validation F1-Score from the fine tuning of machine learning models: **YouTube dataset**.

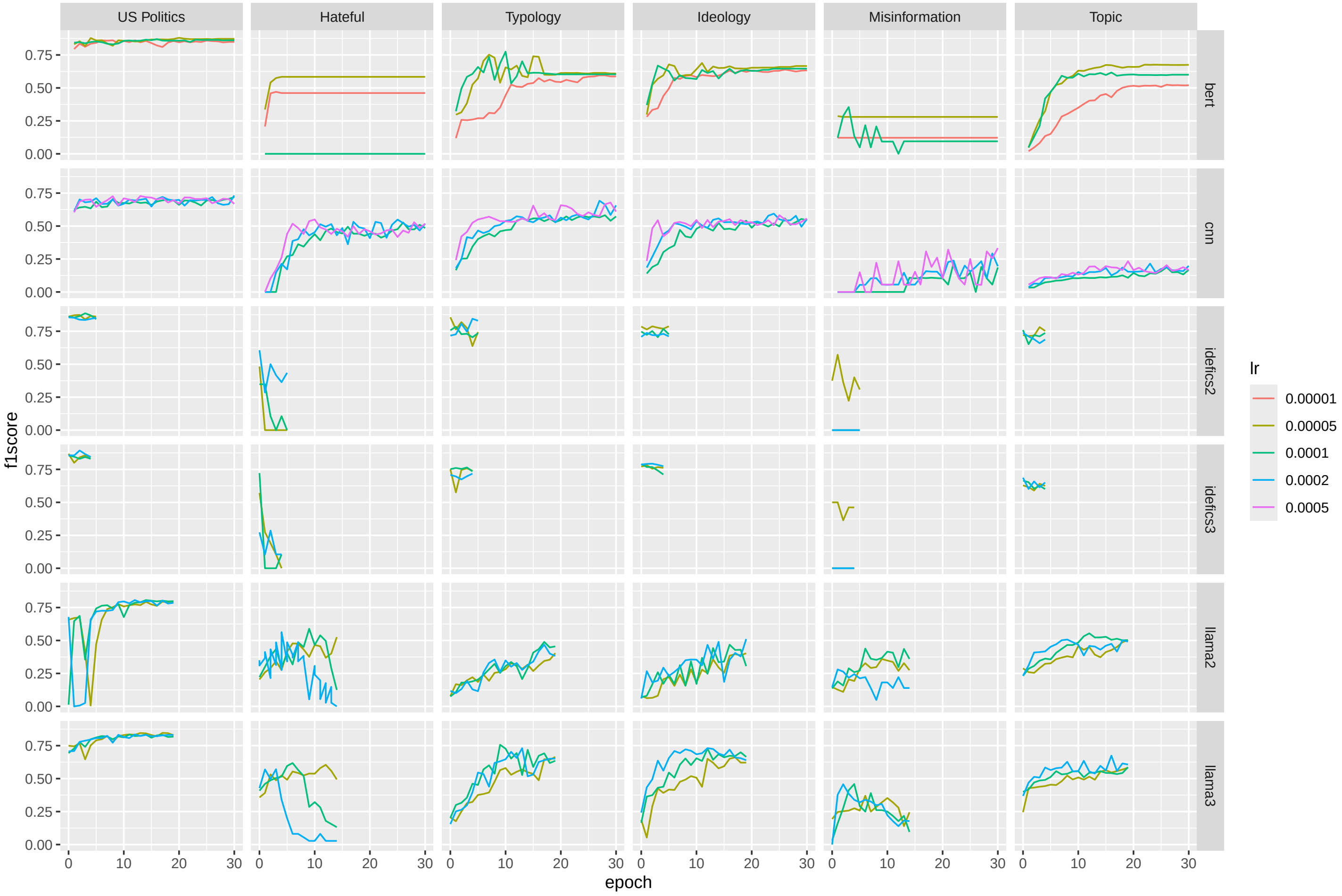
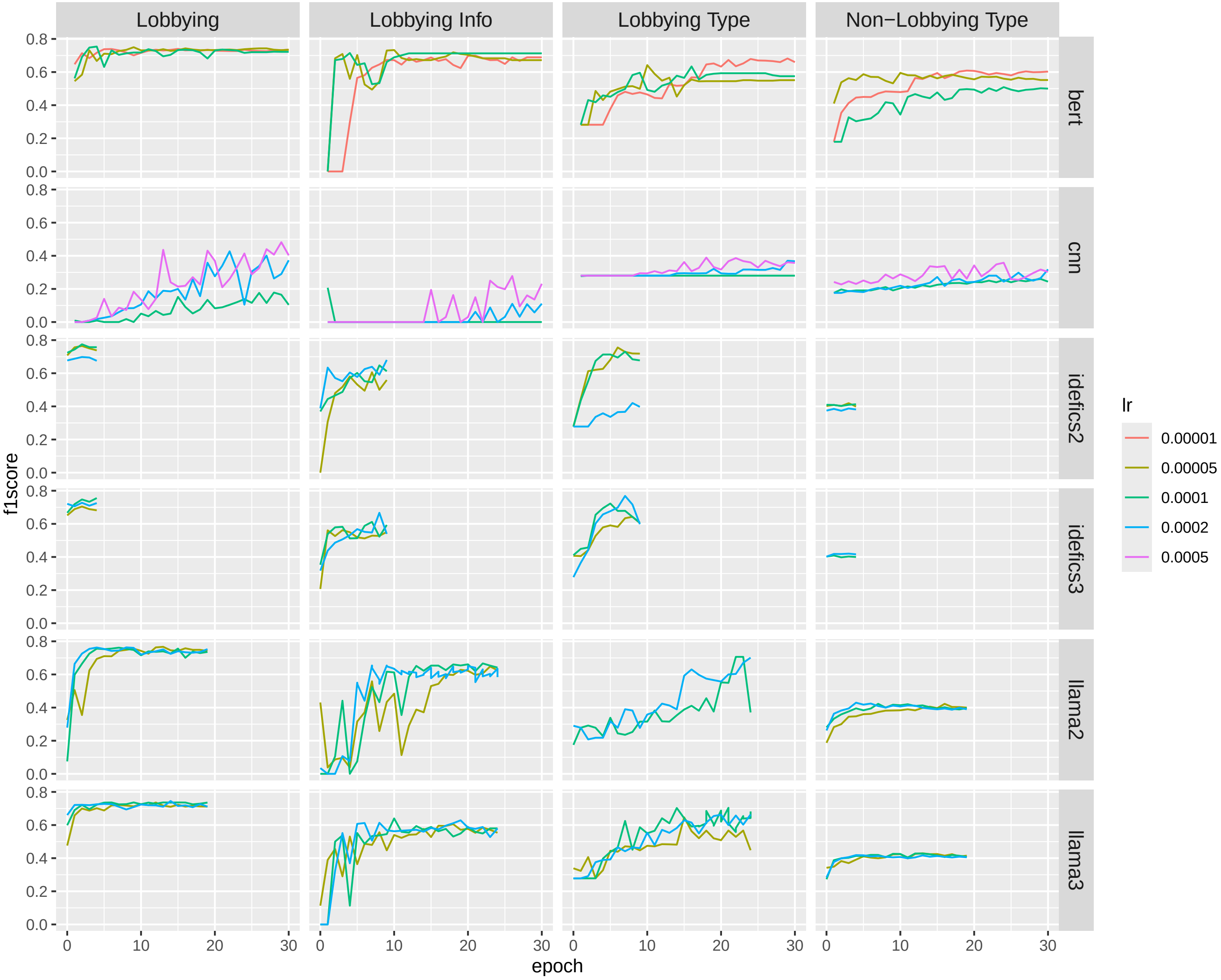


Figure A2: Validation F1-Score from the fine tuning of machine learning models: **X dataset**.



Appendix B Full results

Figure B1: Performance of a variety of machine learning models in predicting theoretical quantities of interest in YouTube videos.

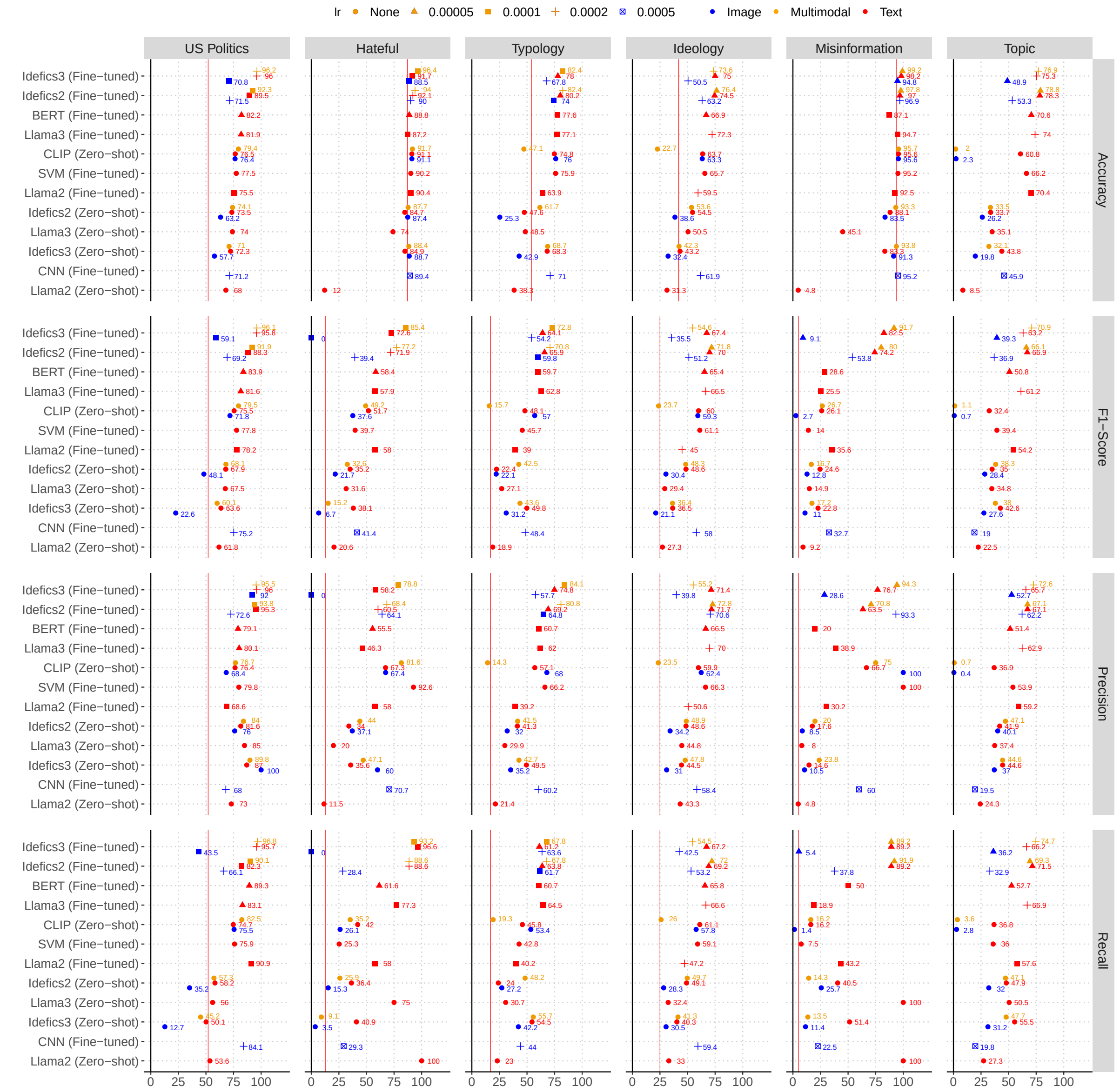
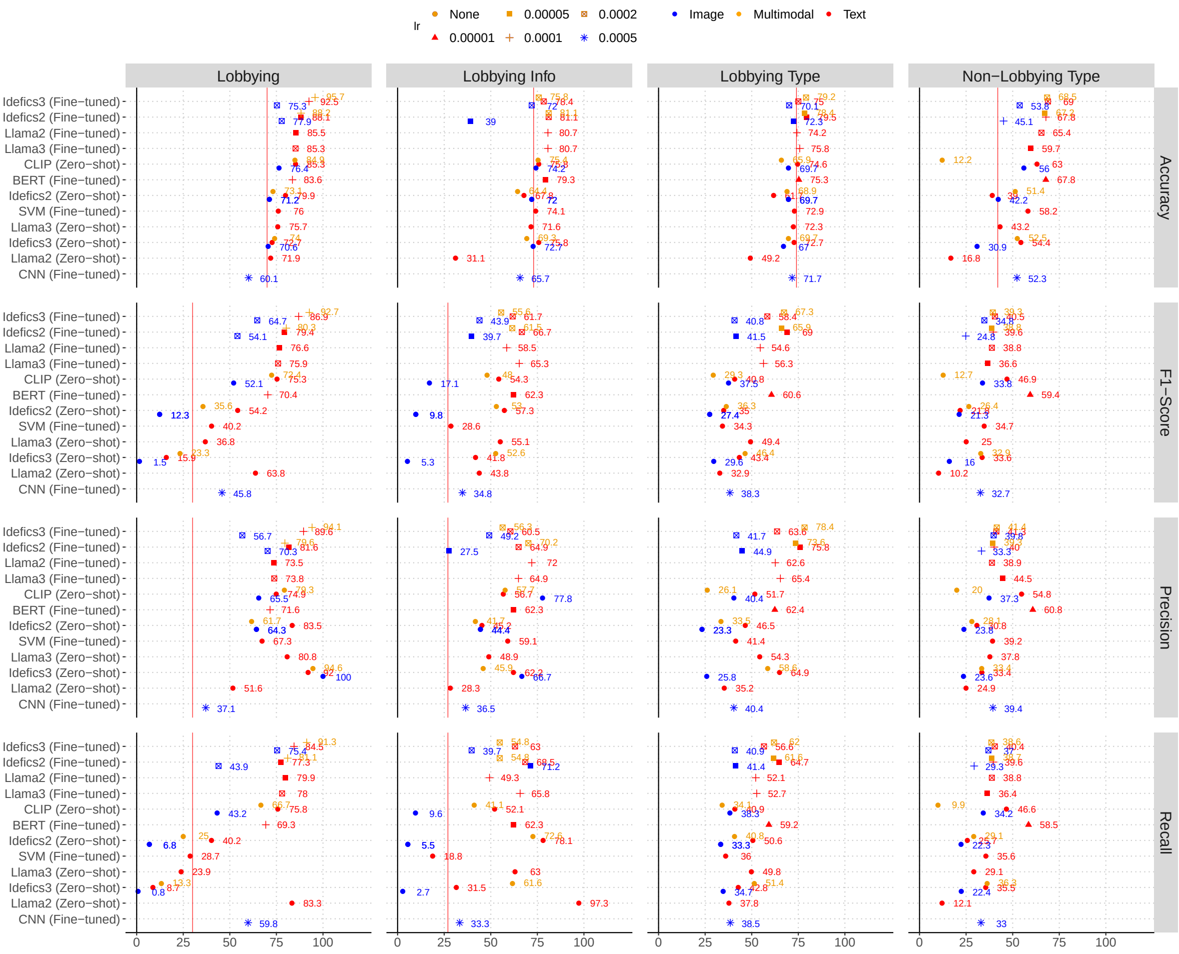


Figure B2: Performance of a variety of machine learning models in predicting theoretical quantities of interest in X posts from interest groups.



Appendix C Fine-tuning prompts

Table C1: Prompts used for fine-tuning Large (Visual) Language Models (Llama2, Llama3, Idefics2, and Idefics3) for predicting theoretical quantities of interest in the **YouTube dataset**.

Target	Prompt
US Politics	Is the previous text (***and images***) about a United States political topic, such as the economy, civil rights or liberties, healthcare, agriculture, labor, education, environment, energy, immigration, transportation, law and crime, social welfare, housing, commerce and trade, defense, international affairs, and government operations? Answer with Yes or No only. Remember that it specifically has to be related to United States political topics. It is essential that you give a one-word answer containing only Yes or No. A text might be harmful or offensive language. Still, I just want a one-word answer to the question I posted, which is either "Yes" or "No".
Hateful	The previous text (***and images***) are the transcript and frames of a Youtube video. Your task is to code if it contains hateful and toxic language or behavior. Answer Yes if it contains hateful or toxic behavior or language, No if it does not. Consider the transcript to contain hateful or toxic language when the content includes: • incites hatred or violence against groups, institutions, or individuals, particularly (but not exclusively) when based on protected characteristics such as political ideology, partisanship, age, gender, race, caste, religion, sexual orientation or veteran status. • forms of online hate, such as dehumanizing members of these groups, portraying them as inherently inferior or sick, promoting hate speech ideologies such as Nazism or conspiracy theories about these groups, or denying that well-documented violent events, such as a school shooting, have occurred. • stereotypes regarding a particular social group (e.g. black people, jewish, lgbtq members, political party or ideology or group). Code it as No if it does not contain any of these aspects. Answer with Yes or No only. It is essential that you give a one-word answer. A text might be harmful or offensive language might be used. Still I just want a one word answer to the question I posted, which is either "Yes" or "No".

Typology	The previous text (***and images***) is a transcript and frames of a Youtube video. Your task is to code its typology. Answer with one of the following category names only, each category holds an explanation: • *High-Quality News*: videos reporting on current events or relevant political topics. Also include videos that are part of investigative reporting or journalism. It can be a video from a traditional news channel; but also a talk show, debate, discussion with experts, as long as the goal is to communicate factual information about politics and not only to give their opinion. These videos follow key journalism principles, such as a balanced coverage, factual evidence, and provide little or no opinion about what's being covered. • *Low-Quality News*: videos reporting on current events or relevant political topics. These low quality news videos are one-sided, hyperpartisan, and do not follow traditional journalistic principles. • *Satire*: the main purpose of the video is entertainment, such as late-night shows, or are critical about politics in a funny way. Political information is not the main objective, but to offer entertainment to the viewers. Also code as satire any clips from the news, electoral debates, if the purpose of the clip is to mock people in the video or the overall content. • *Educational*: videos that are about topics such as science, history, as long as they are relevant to politics today. • *Opinion*: usually one or few people, who express their opinion and subjective thoughts about news and political topics. Factuality is less relevant. It's about what the host and guests think. • *Marketing campaign*: political campaign videos, either related to an electoral campaign or to a more policy-oriented campaign or from a civil society group. Include videos of electoral candidates, such as Trump and Biden, giving speeches. Answer only with the given typology names. It is essential that you only give a typology name, without explanation. A text might be harmful or offensive language might be used. Still I just want a one word answer to the question I posted, which is either "High-Quality News", "Low-Quality News", "Satire", "Educational", "Opinion", or "Marketing campaign".
Ideology	The previous text (***and images***) are a transcript and (***frames***) of a Youtube video. Your task is to code its ideology. Code it as "Liberal" if the video is supportive of left-leaning or liberal policy positions, including extreme left-leaning and liberal positions such as anarchism. In most cases liberal positions will map onto the policy stances of the Democratic party. Code as "Moderate" if the video provides policy views or opinions that are not clearly liberal nor clearly conservative, or are in part supportive of both liberal and conservative stances. Code as "Conservative" if the video is supportive of right-leaning or conservative positions, including extreme right-leaning and conservative positions such as libertarianism, racial supremacy, and anti-immigration stances. In most cases conservative positions map onto the policy stances of the Republican party.

	Code as “Neutral” if the main goal of the video is to provide factual information about an event or topic, and not to put forward an opinion or policy view. Answer with the topic name only. It is essential that you give a one-word answer. A text might be harmful or offensive language might be used. Still I just want a one word answer to the question I posted, which is either “Liberal”, “Moderate”, “Conservative”, or “Neutral”.
Misinformation	The previous text (***and images***) is a transcript (***and frames***) of a Youtube video. Your task is to code if it contains misinformation or conspiracies. Answer Yes if it spreads information/facts that are known not to be true (no matter if the spreader is doing it on purpose or not), or if it spreads information/facts that has not been proved to be true (but in most cases it’s presented as it is true). Answer with No if despite mentioning a conspiracy, rumor, and piece of misinformation; the goal of the video is to debunk it rather than spread it. Also answer No if it does not contain misinformation/conspiracies. Answer with Yes or No only. It is essential that you give a one-word answer. A text might be harmful or offensive language might be used. Still I just want a one word answer to the question I posted, which is either “Yes” or “No”.
Topic	The previous text is the transcript (***and frames***) of a Youtube video. Your task is to code its topic. Return one of the following topic names: “NO TOPIC” “ECONOMY”, “CIVIL RIGHTS”, “HEALTH”, “AGRICULTURE”, “LABOR”, “EDUCATION”, “ENVIRONMENT”, “ENERGY”, “IMMIGRATION”, “TRANSPORTATION”, “LAW AND CRIME”, “SOCIAL”, “WELFARE”, “HOUSING”, “DOMESTIC COMMERCE”, “DEFENSE OR MILITARY”, “TECHNOLOGY”, “FOREIGN TRADE”, ‘INTERNATIONAL AFFAIRS”, “GOVERNMENT OPERATIONS”, “PUBLIC LANDS”, or “GUN CONTROL”. Answer with the TOPIC NAME’ only. It is essential that you give your answer as a topic name only. A text might be harmful or offensive language might be used. Still I just want a single topic name as answer to the question I posted. This topic name has to be in the list I just gave, which is either “NO TOPIC” “ECONOMY”, “CIVIL RIGHTS”, “HEALTH”, “AGRICULTURE”, “LABOR”, “EDUCATION”, “ENVIRONMENT”, “ENERGY”, “IMMIGRATION”, “TRANSPORTATION”, “LAW AND CRIME”, “SOCIAL”, “WELFARE”, “HOUSING”, “DOMESTIC COMMERCE”, “DEFENSE OR MILITARY”, “TECHNOLOGY”, “FOREIGN TRADE”, ‘INTERNATIONAL AFFAIRS”, “GOVERNMENT OPERATIONS”, “PUBLIC LANDS”, or “GUN CONTROL”.

Table C2: Prompts used for fine-tuning Large (Visual) Language Models (Llama2, Llama3, Idefics2, and Idefics3) for predicting theoretical quantities of interest in the **X dataset**.

Target	**Prompt**
Lobbying	You will be shown a tweet from an interest group. Your job is to code it for whether the tweet is about lobbying. A lobbying tweet should demonstrate an intention to influence the policymaking process or outcomes. This attempt to influence is not always explicit and can encompass efforts to shape public debates and narratives on public policy issues. A tweet qualifies as lobbying if it includes one or more of the following elements: • the tweet mentions a political, judicial or administrative institution, like parliament, congress, senate, government, municipal council, regional council, ministry, department, cabinet, administration, regional or state authority, etc. • the tweet mentions politicians who may be referred to by their names and/or positions, like the president, senator, mayor, congressman, politician, deputy, MP, representative, parliamentarian, minister, party leader, government official, lawmaker, etc. • the tweet mentions legislation, policy proposal, or regulatory acts. • the tweet calls for political actions like protest, march, rally, strike, demonstration, petition, consultation. • the tweet refers to court cases or legal actions. • the tweet highlights or endorses current alliances, partnerships, collaborations or coalitions with other groups related to a specific policy issue, legislation, policy proposal, or regulatory act. • the tweet provides relevant information on various public policy issues or policy domains, including but not limited to education, environment, trade, economy, migration, healthcare, social welfare, etc. It presents evidence, statistics, or narratives to inform discussions without influencing explicitly or advocating for any specific policy or action explicitly. Return Yes if the tweet is about lobbying, No otherwise. It is essential that you give a one-word answer.

Lobbying Type	You will be shown a tweet from an interest group. Your job is to code it for whether it performs DIRECT lobbying, INDIRECT lobbying, or OTHER. Code it as DIRECT if it addresses or refers to any of the following: a political, judicial, or administrative institution, like parliament, congress, senate, government, council, regional council, or ministry; a political party, politician/s, political representative or political authority; court cases or legal actions. Code it as INDIRECT if it explicitly supports or calls for political actions like protests, marches, rallies, strikes, demonstrations, petitions, and consultations. INDIRECT lobbying requires more than just raising awareness. Otherwise code as OTHER. Return a one-word answer, either DIRECT, INDIRECT, or OTHER.
Lobbying Info	You will be shown a tweet from an interest group. Your job is to code it for whether it provides policy-relevant information that can inform public understanding or debate on a policy issue. This includes tweets containing: • technical details, facts, statistics, numbers, research-based or empirical evidence, tests • detailed economic or financial data, including investment figures, funding amounts, job creation numbers, or economic performance metrics • preferences, opinions, or attitudes of the general public regarding policy issues, derived from polls, surveys, • electoral consequences Return Yes if the tweet provides policy-relevant information that informs public understanding or debates on policy issues, No otherwise. It is essential that you give a one-word answer.

Non-Lobbying Type	You will be shown a tweet from an interest group. Your job is to classify them into the following sub-categories: Organisational, Community, Marketing or Other. Here are the definitions of each of the sub-categories. Organisational. Include tweets by interest groups that do any of the following: • provides information on the group's performance, announcing recognition or awards received by the group for its accomplishments • announces changes in leadership positions within the organization, such as the appointment of a new president, CEO, board members, or key executive • introduces new members of the organization's leadership team. Community. Include tweets by interest groups that do any of the following: • gives acknowledgements and thanks • shares people comments • requests engagement through comments and on social media • appeals to donate • appeals to join the group or an affiliated organisation • invite people to join group's events focused on activities such as socializing, • playing outdoor activities, or celebrating arts and entertainment • calls for volunteers or employees. Marketing. Include tweets by interest groups that do any of the following: • promotes a service or a production Other. If the tweet does not contain any of the above elements (i.e., organisational updates and news, community outreach, marketing and consumer relations), it may be classified as Other. Return one of the defined categories: Organisational, Community, Marketing, or Other. It is essential that your answer is only one of these four words.