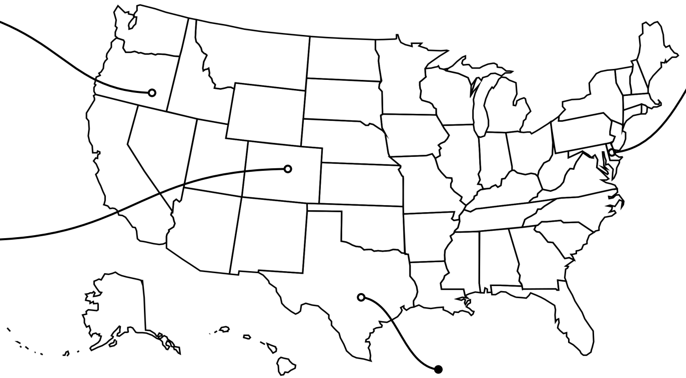
Multimodal Modeling for Political Science Research

@xxxx: F*** Trump. #FamiliesBelongTogether

@fams2gether: Nobody, no matter who you are, where you are from, you should spend an hour like this!



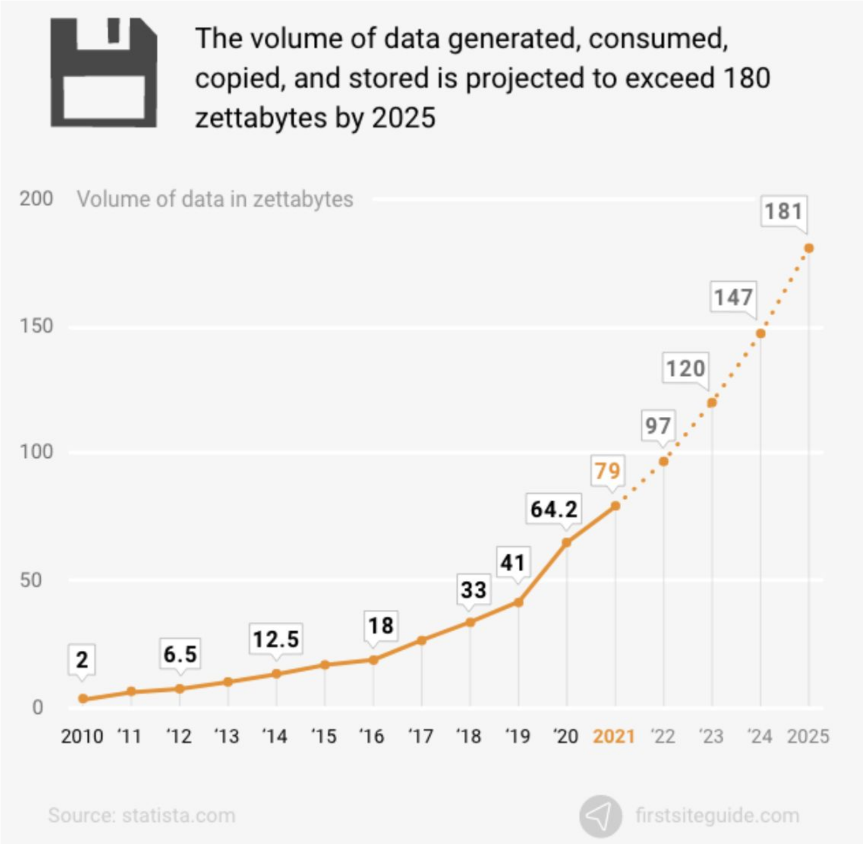
@xxxx: #FamiliesBelongTogether in Denver this morning.

@ResistSnow: #FreeTheChildren #FamiliesBelongTogether Our children long for realistic maps of the future that they can be…

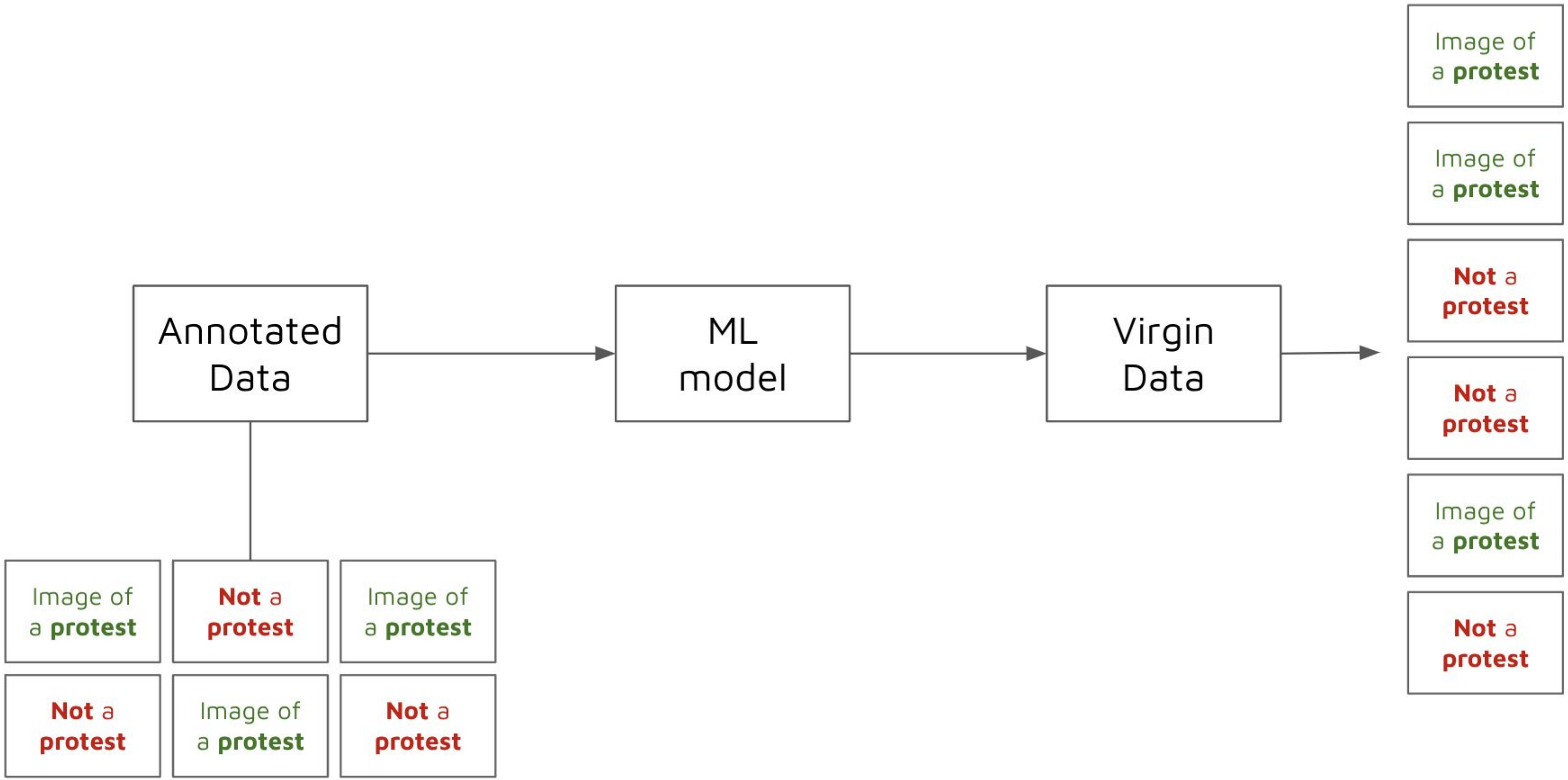
Andreu Casas
Royal Holloway Univ. of London

Freek Cool
Vrije Universiteit Amsterdam

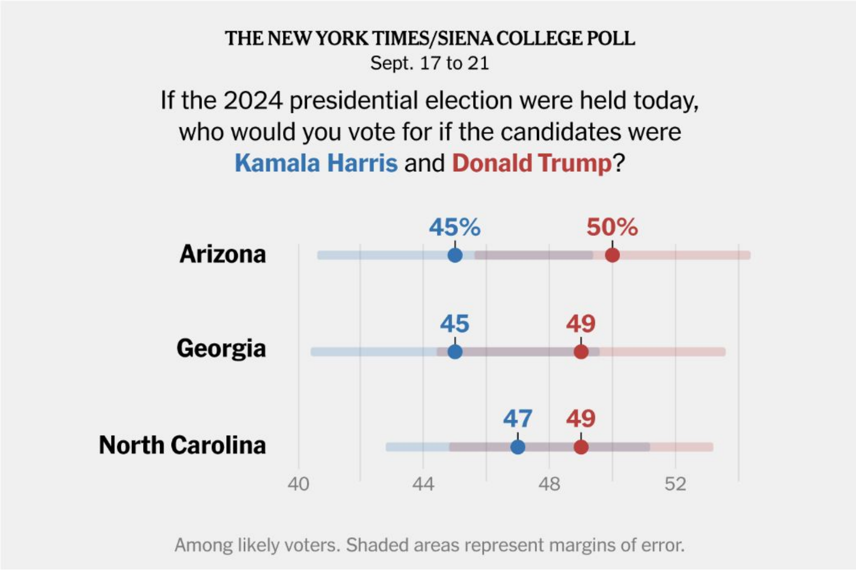
(Political) scientists have an **increasing amount of (digital) data available** for addressing their questions of interest



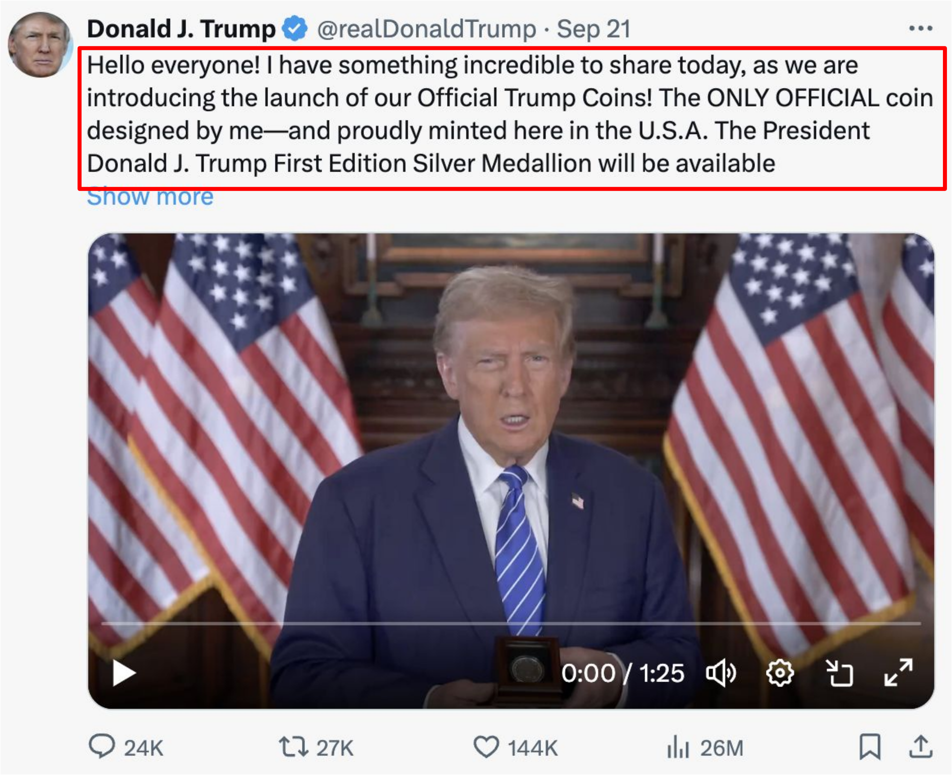
Supervised machine learning is often used to identify a defined concept in large amounts of data

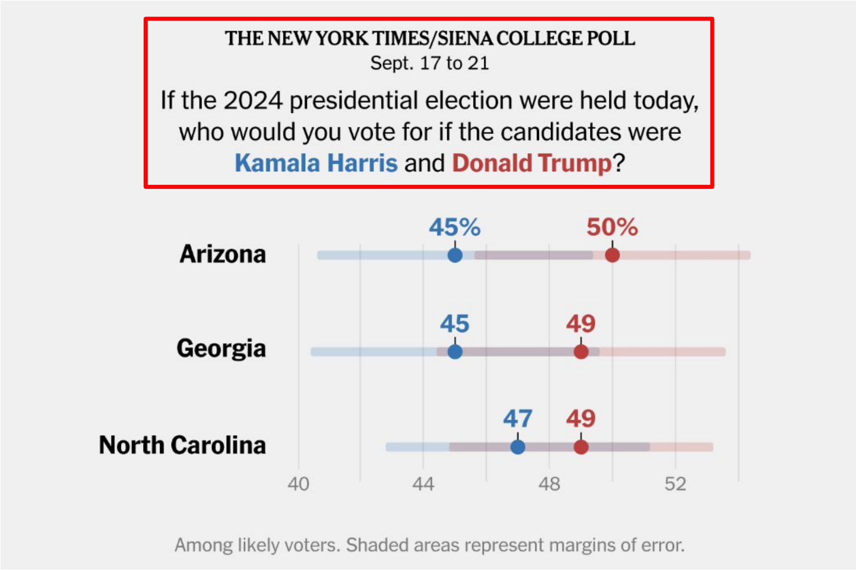


Often this data is **multimodal:** with text, visuals, audio



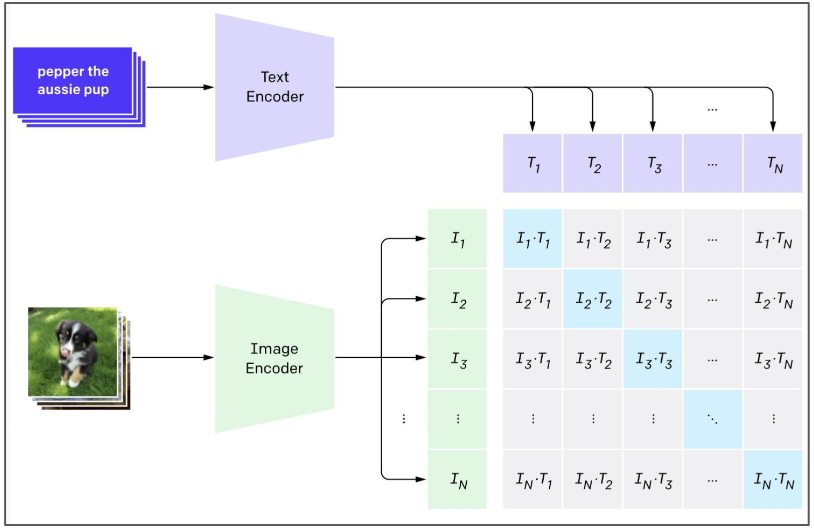
Yet **we mostly only use one data modality** to train supervised ML (e.g. text)

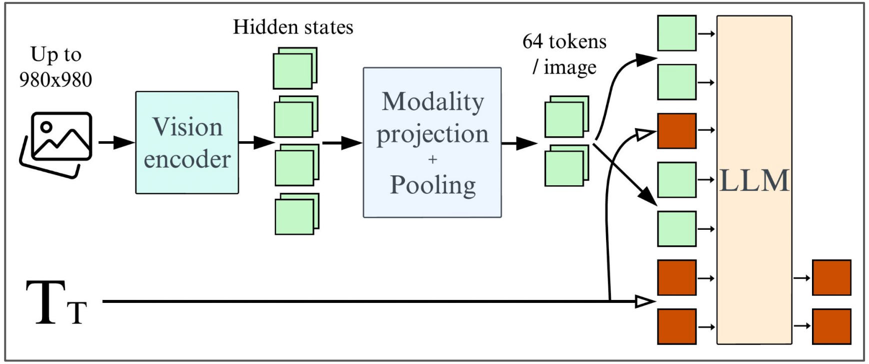




Recent computational advances make **multimodal modeling easier**

CLIP

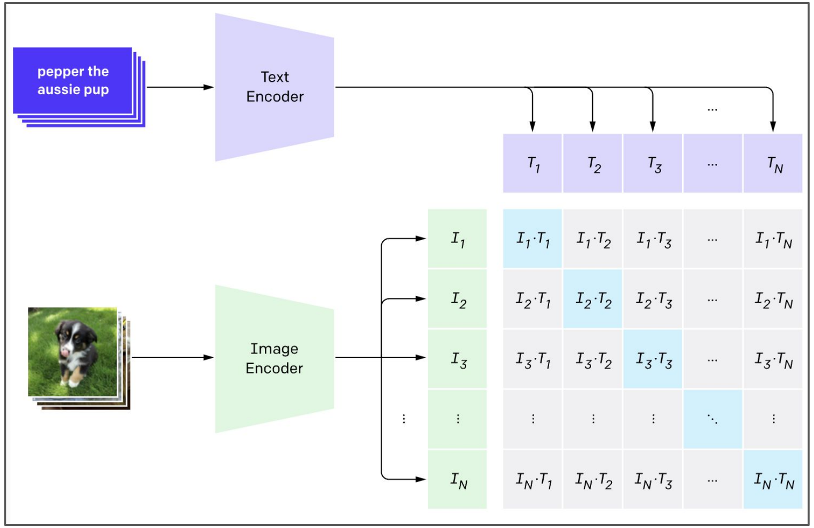


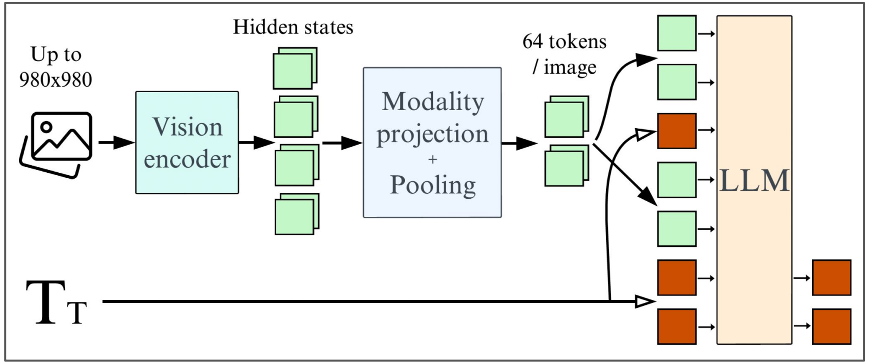


Idefics2

But we don't know much about **whether** nor **the conditions under which multimodality can help improve the performance of supervised ML**

CLIP





Idefics2

Two **original annotated dataset** (10 tasks)

(1) **YouTube Videos** from channels posting on US politics (N = ~4,000) → 6 tasks

(2) **Twitter Posts** from interest groups from US, ES, DK, GE (N = ~4,000) → 4 tasks

Seven **models**:

(1) **Text only**: SVM, BERT, Llama2, and Llama3

(2) **Image only**: CNN

(3) **Text + Image**: CLIP, Idefics2

This draft: 1 dataset and **4 tasks**

/!\ Work in Progress

(1) **YouTube Videos** from channels that post about US politics (N = ~4,000)

- (a) *US politics*: is the video about US politics?
- (b) *Hateful*: does the video have hateful content?
- (c) *Typology*: what type of video? (e.g. opinion, high-quality news, etc.)
- (d) *Ideology:* ideology of the video (liberal, moderate, conservative, neutral)

Seven **models**:

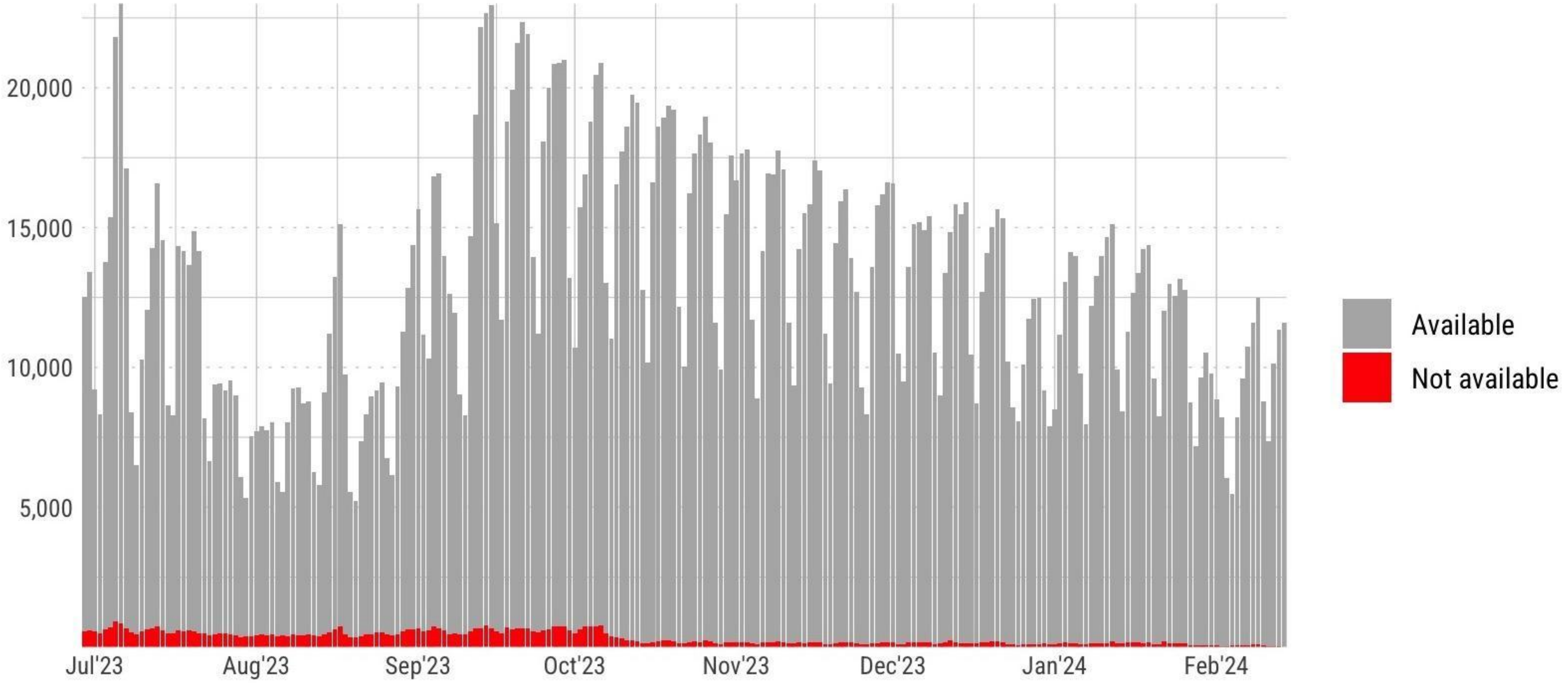
(1) **Text only**: SVM, BERT, Llama2, and Llama3

(2) **Image only**: CNN

(3) **Text + Image**: CLIP, Idefics2

Data

- about **6 million videos**
- from about **12k channels**
- **3% channels** not available
- **2% videos** not available



Data

Video data

Metadata

- Title, Description, Topic labels, Duration, Tstamp

Transcript

OpenAI/Whisper

- *Tiny* version
- Multilingual
- Quite fast (~1.5min. for 10min. video)

Certainly, a fairly big deal he has said that then X Vice President Biden in May 2017 and TECS seemed to show that he met with then X Vice President Biden in early 2017 along with it seems like Hunter Biden and James Biden. And Bobulinski says that at sort of a high level the business efforts that were ongoing were discussed and that he thinks it was pretty clear that Joe Biden was at least aware of the business efforts that were going on with Hunter. And if you don't believe Bobulinski then listen to the President himself from a phone message found on the laptop. He tells us that it's E-16 and once you know you can get a chance to be called not the business effort. If you want to talk to us like the article, at least just then online, it's going to be planned to run time and spirit and you can clear. Anyway, if you get a chance to be called. You can hear Joe Biden leaving the voice mail for Hunter and he's talking about a new talk time story that's just come out and it was investigating CFC, Chairman Yee basically all the partners of Hunter Biden and Jim Biden that were giving the millions of dollars. And the Biden family was panic that the whole story was going to block as it turned out when the story was published. There was hardly any mention of the Biden's, the Biden's were in the clear. In addition to evidence on the laptop, Chairman Comer says he's received new leads from whistleblowers. We've had whistleblowers from a financial institution. We've had whistleblowers who were former business associate. We've also had a whistleblower who the lead is to a bit on a call with the President and Hunter Biden in one of these shady business deals. So these are the people that we want to hear from under oath. One allegation that especially worries Comer is that the Biden family tried to sell China access to US oil and gas. According to Josh Boswell of the Daily Mail, Joe Biden was named in an October 2017 email found on Hunter's abandoned laptop discussing a multi-million dollar natural gas deal with China. In the email, Louisiana Attorney Robert Finett told Hunter Biden and Jim Biden he had secured a contract for them with a Houston-based energy company to be available for a call from Joe Biden and Hunter Biden on Monday morning to discuss the purchase of 5 million tons of gas. We're Hunter Biden was trying to not only sell them, liquefied natural gas from the United States, but also, according to a whistleblower, helping them try to get their foot in the door to where they can start purchasing interest in the natural gas drillers. Can you imagine China own in the natural gas drillers in the United States? That's a huge national security risk. That's why we're investigating Joe Biden. I mean, this is very serious stuff. Another tip led to a new house Republican inquiry into Hunter Biden's former Chinese secretary, Jackie Baal. She goes and works for the Chinese Communist Party. And then somehow she ends up in the United States working for a hundred Biden. We are concerned that she may be a spy. Baal, however, has not been implicated in any professional or criminal misconduct in connection with her work for Hunter Biden who has congressional investigators curious about his banking records. Comers says they contain an alarming number of suspicious activity reports. With suspicious activity reports, this is the bank notified on the federal government that we're pretty certain our client has committed a crime. Now think about that. I can tell you with confidence that 99.99% of Americans never even come close to getting a suspicious activity report. Hunter Biden has had 150. That's not really implicated anybody. But what they do is they lay down the financial trail. It shows you where the money was coming from and where it was going to. It was coming from China. It was coming from Burisma. It was coming from Ukraine. Coming from Russia. Wherever it was coming from, we now see the money trail through the bank accounts and into Biden's family coffers or the bank accounts of their business partners. Hunter Biden has not been charged with any criminal misconduct. In connection with his banking representative Comer, however, believes Joe Biden tried to hide those reports from Congress and he now wants to know why. But here's the interesting thing about the suspicious activity reports prior to Joe Biden coming president under George Bush, Barack Obama and Donald Trump. Any committee in Congress could request those suspicious activity reports from Treasury. When Joe Biden came president, one of the first rule changes he made was he changed the rule to where Congress couldn't have access to suspicious activity reports. Now why would he do that? At the time, nobody knew. And in fact, Maxine Waters, when she heard about it, she put language and legislation that would give Congress the authority to get those sars again. Because she didn't know that Hunter Biden had 150. None of us knew that. But then when it came out, the Hunter Biden had 150 suspicious activity reports. And I tried to access those. They said, now that was a rule change that Joe Biden made. Well, that's a perfect example of Joe Biden being compromised because of his son, shady business dealings. These were created to help everyone communicate to try to ward off illegal foreign transaction coming in for crime for terrorism. And he changed rules because his son had 150 thumbs. That's very serious. And that's something that we hope that the Treasury department will give those to us. And new emails further suggest Joe Biden was indeed aware of Hunter's overseas business deals. According to the Daily Mail, emails dating back to 2014 show the VEP met in the White House with two Chinese energy executives from a company called Wen Shen, which partnered with Hunter Biden's company, Sen. A global partners for a $1.25 billion natural gas plant in China. According to the report and the White House visitor logs, Biden met with Wen Shen America, President Pinny on July 25, 2014. We have lots of evidence that Vice President Biden met in the White House, the Obama White House, with associates of Hunter Biden, people who are business

Frames

- Download video (temporarily)
- Extract frames (e.g. 1 frame/sec.)
- Frame embeddings (MoblieNetv3 small)
- Cosine similarity all pairs of frames
- identify "duplicate" frames (e.g. > 90%)
- Only keep "unique" frames

Data

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%

Models

Text only:

(1) SVM

(2) BERT

(3) Llama2

(4) Llama3

Image only

(5) CNN

Text + Image

(6) CLIP

(7) Idefics2

Models

Text only:

(1) SVM

(2) BERT

(3) Llama2

(4) Llama3

Image only

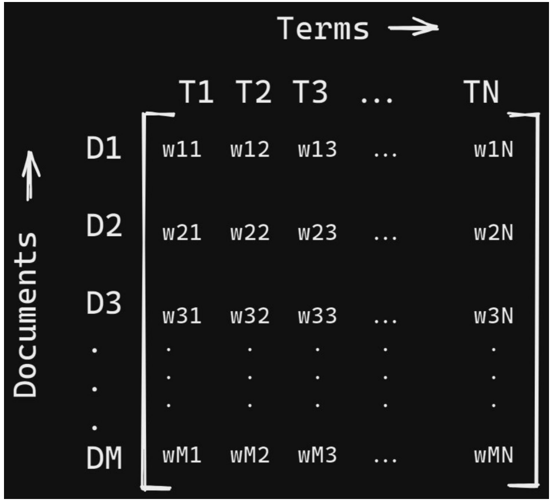
(5) CNN

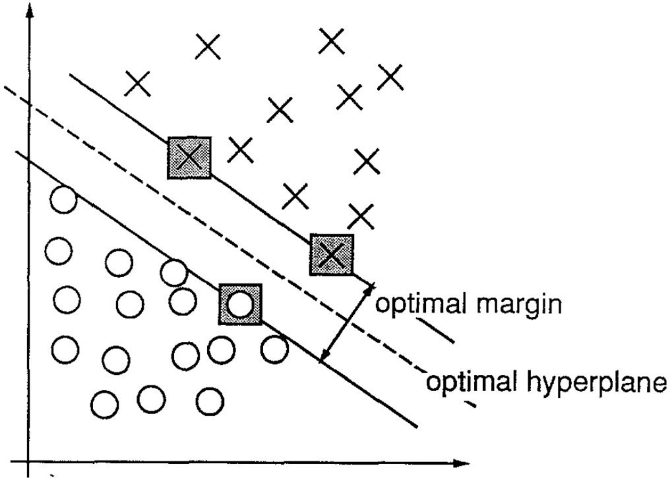
Text + Image

(6) CLIP

(7) Idefics2

- Ngram-based, **no prior language knowledge**, fully task dependent.





Cortes & Vapnik (1995)

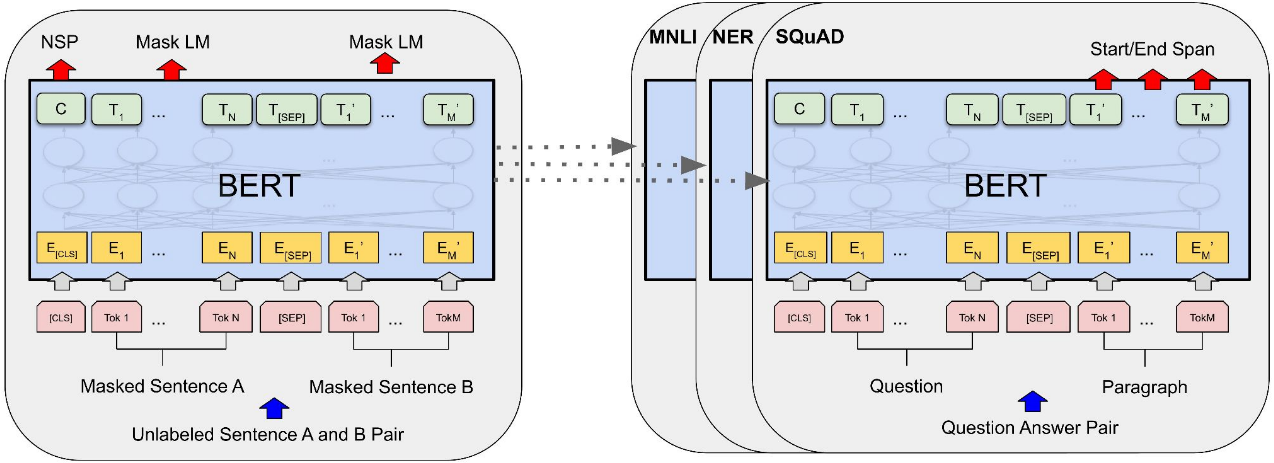
Models

Text only:
(1) SVM
(2) BERT
(3) Llama2
(4) Llama3

Image only
(5) CNN

Text + Image
(6) CLIP
(7) Idefics2

- Transformed-based, self-supervised, language model (prior knowledge), **fine-tuned on next sentence**, can be fine-tuned to do new tasks.
- Trained on: 11k books (800mil tokens) + English wikipedia (**2.5 bil** tokens)
- **bert-base-uncased**: **110 mil** parameters
- Context length: **512** tokens



Pre-training Fine-Tuning

Models

Text only:

(1) SVM

(2) BERT

(3) Llama2

(4) Llama3

Image only

(5) CNN

Text + Image

(6) CLIP

(7) Idefics2

- Transformed-based, self-supervised, language model (prior knowledge), **fine-tuned on instruction task**, can be fine-tuned on new instructions
- Trained on: **2 tril** tokens, publicly available sources (although *unknown*)
- **7Bil/13bil/70bil** parameters → **Llama-2-7b-chat**
- Context length**: 4,096 tokens**

Models

Text only:

(1) SVM

(2) BERT

(3) Llama2

(4) Llama3

- Transformed-based, self-supervised, language model (prior knowledge), **fine-tuned on instruction task**, can be fine-tuned on new instructions
- Trained on: **15 tril** tokens, publicly available sources (although *unknown*)
- **8Bil/70/405 bil** parameters → **Llama-3-8b-instruct**
- Context length**: 8,000**

Image only

(5) CNN

Text + Image

(6) CLIP

(7) Idefics2

Models

Text only:

(1) SVM

(2) BERT

(3) Llama2

(4) Llama3

Image only

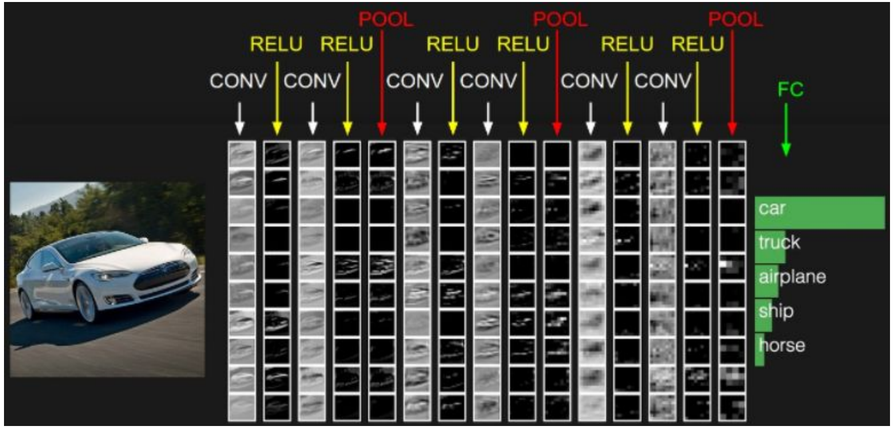
(5) CNN

Text + Image

(6) CLIP

(7) Idefics2

- Pre-trained for object recognition: 1,000 ImageNet object classes
- Trained on: **1.28 mil** images
- **25.6 mil** parameters → **ResNet50**
- Input size: 224 x 224 x 3



Models

Text only:

(1) SVM

(2) BERT

(3) Llama2

(4) Llama3

Image only

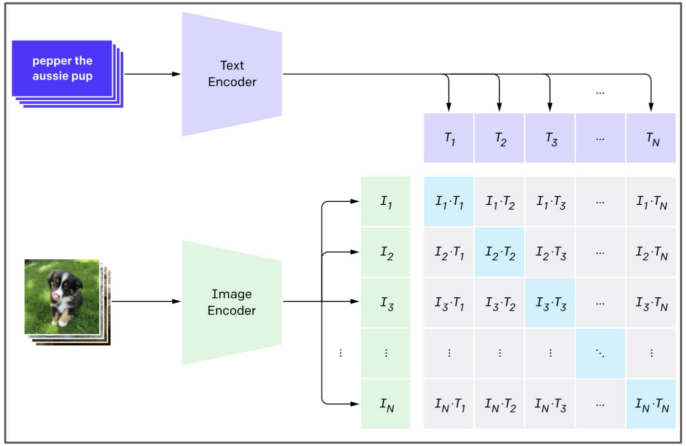
(5) CNN

Text + Image

(6) CLIP

(7) Idefics2

- Text and image encoder
- Trained on **400 mil** text-image pairs: e.g. image and its caption
- **Self-trained**: predicting correct text-image pair
- Image input size: **224 x 224 x 3**
- **150**/400 **mil** parameters → **ViT-B/32**
- Context length: **77 tokens**



Models

Text only:

(1) SVM

(2) BERT

(3) Llama2

(4) Llama3

Image only

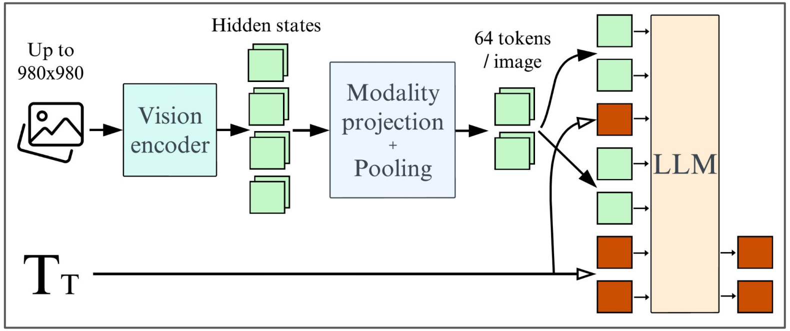
(5) CNN

Text + Image

(6) CLIP

(7) Idefics2

- Text and image **transformer** encoder
- Trained on:
  - **interleaved image-text document: 350 mil images** and **115 bil text tokens.**
  - **Image-text pairs**
  - **PDF OCR extraction: 40 mil**
  - **instruction/chat: 50 open-source datasets**
  - Total: **1.5 bil** images and **225 bil** text tokens
- **8 bil** parameters → **idefics2-8b**
- Image input size: **native resolution (up to 980 x 980)**
- Context length: **?**



Set Up

Text only:

- **(1) SVM**
- **(2) BERT**
- **(3) Llama2**
- **(4) Llama3**

Image only

- (5) CNN

Text + Image

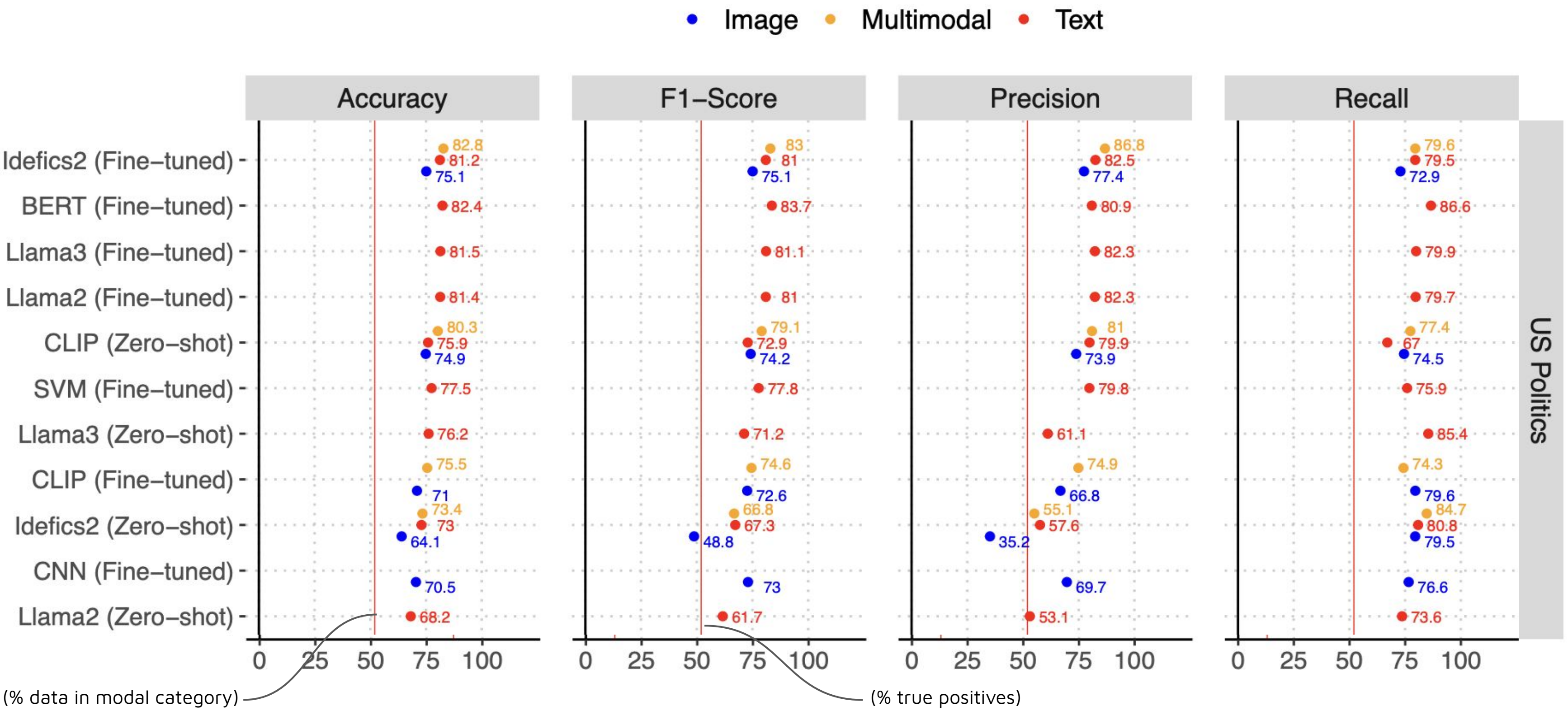
- **(6) CLIP**
- **(7) Idefics2**

- Same **train** (80%) - **test** (20%) sets across models
- Further split train set 80/20 into train/**validation** per **fold**
- **3 folds** (today results only for 1 fold) and **10 epochs**/fold
- Image processing/input:
  - Resize + center_crop all images: 224 x 224 x3
- Text processing/input:
  - 20,000 token vocabulary; no stopwords; linear kernel
  - 512 tokens/video
  - 800 tokens/video
  - 77 tokens/video
- Fine-tuning:
  - $\text{Transcript}_v + \text{Frame}_{v,f} \rightarrow \text{Label}_v$
  - BERT, CNN: new prediction head for each of our tasks
  - Llama2, Llama3, Idefics: same prompts across models
- Evaluation of image and text + image models:
  - Binary: Pred = 1 if $\text{sum}(\text{Frame}_{v,f})$ > threshold (=0 otherwise)
  - Multiclass: Pred = $\text{mode}(\text{Frame}_{v,f})$

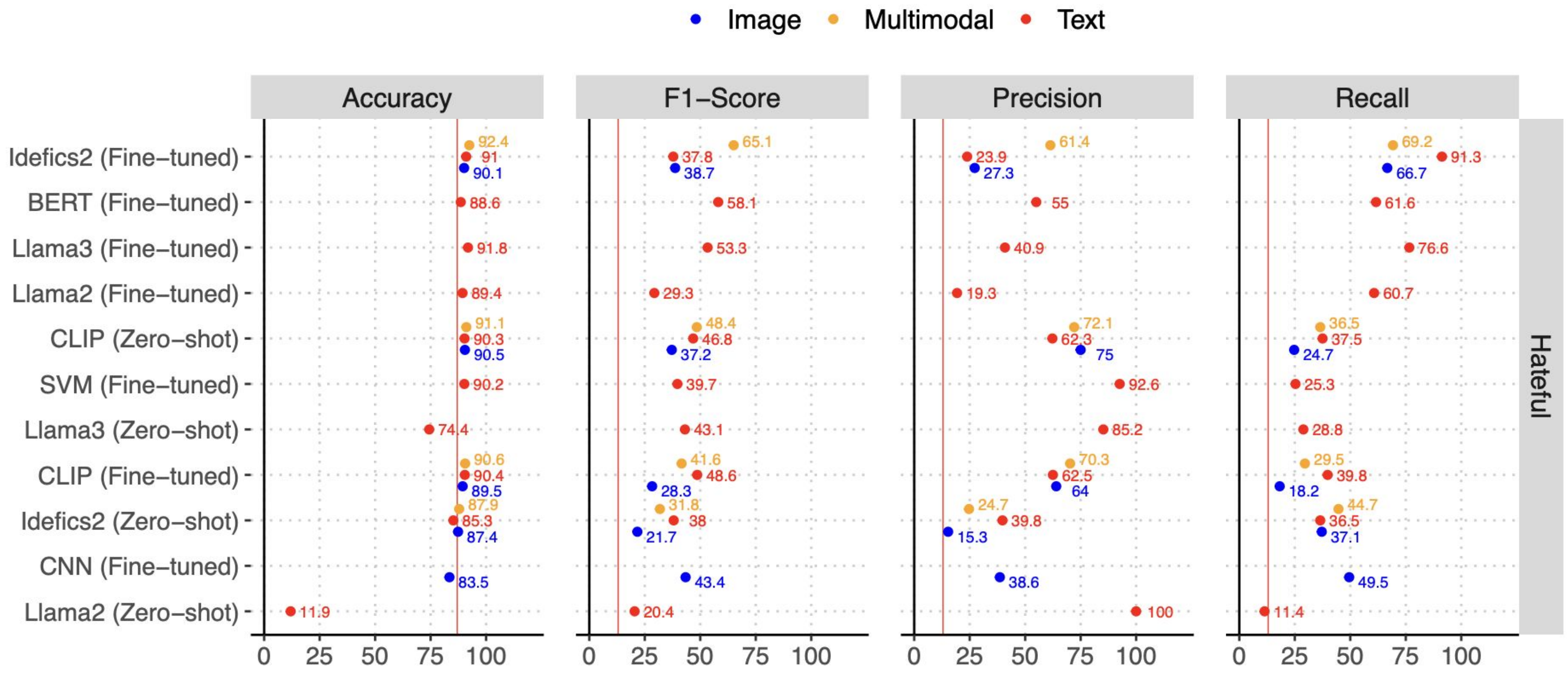
A quick look at how to fine tune a VLM (Idefics2)

```python
messages = [
    {
        "role": "user",
        "content": [
            {"type": "image"}, # frame_v1_f1
            {"type": "image"}, # frame_v1_f2
            {"type": "image"}, # frame_v1_f3
            {"type": "text", "text": transcript_v1,
            {"type": "text", "text": "Is the previous text and images about or relevant to US politics? Answer YES or NO."}
        ]
    },
    {
        "role": "assistant",
        "content": [
            {"type": "text", "text":"YES"}
        ]
    }
]
```

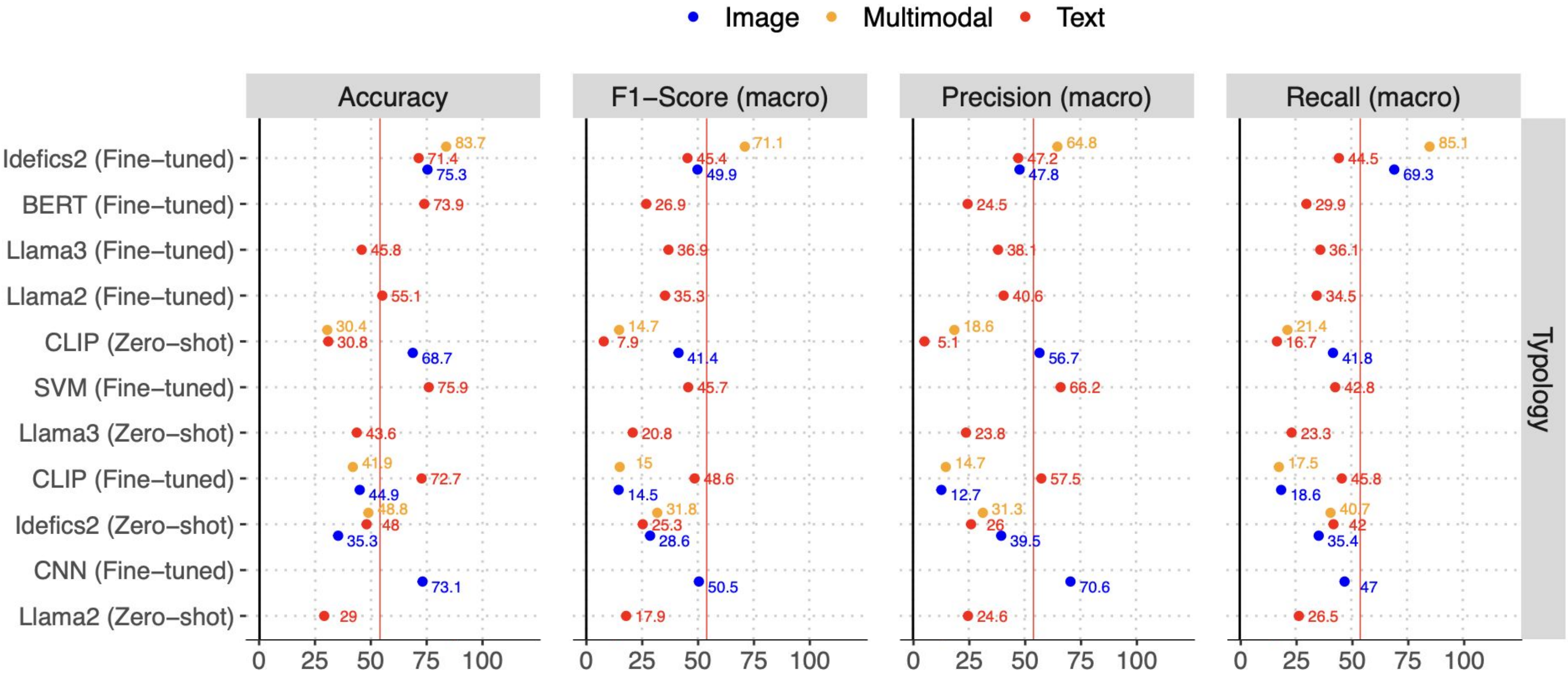
Results: overview



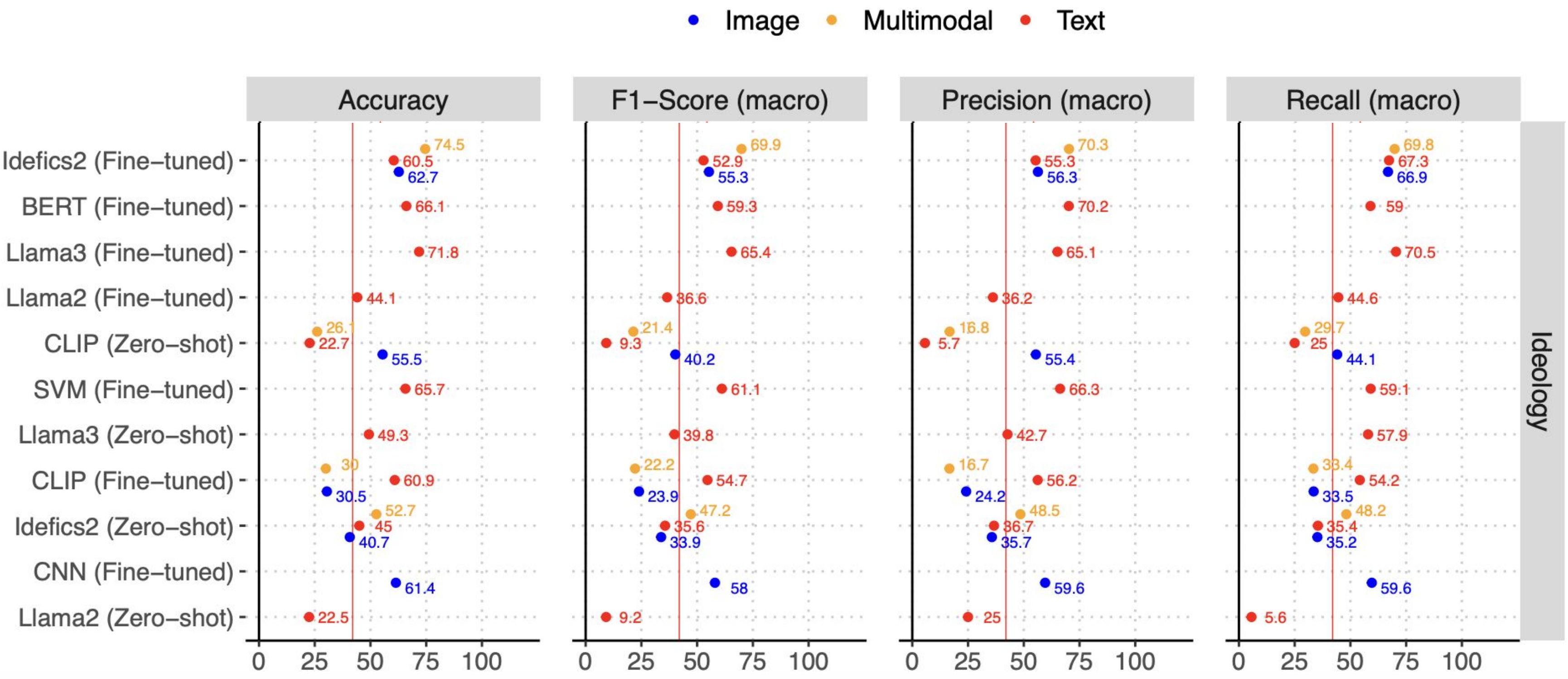
Results: overview



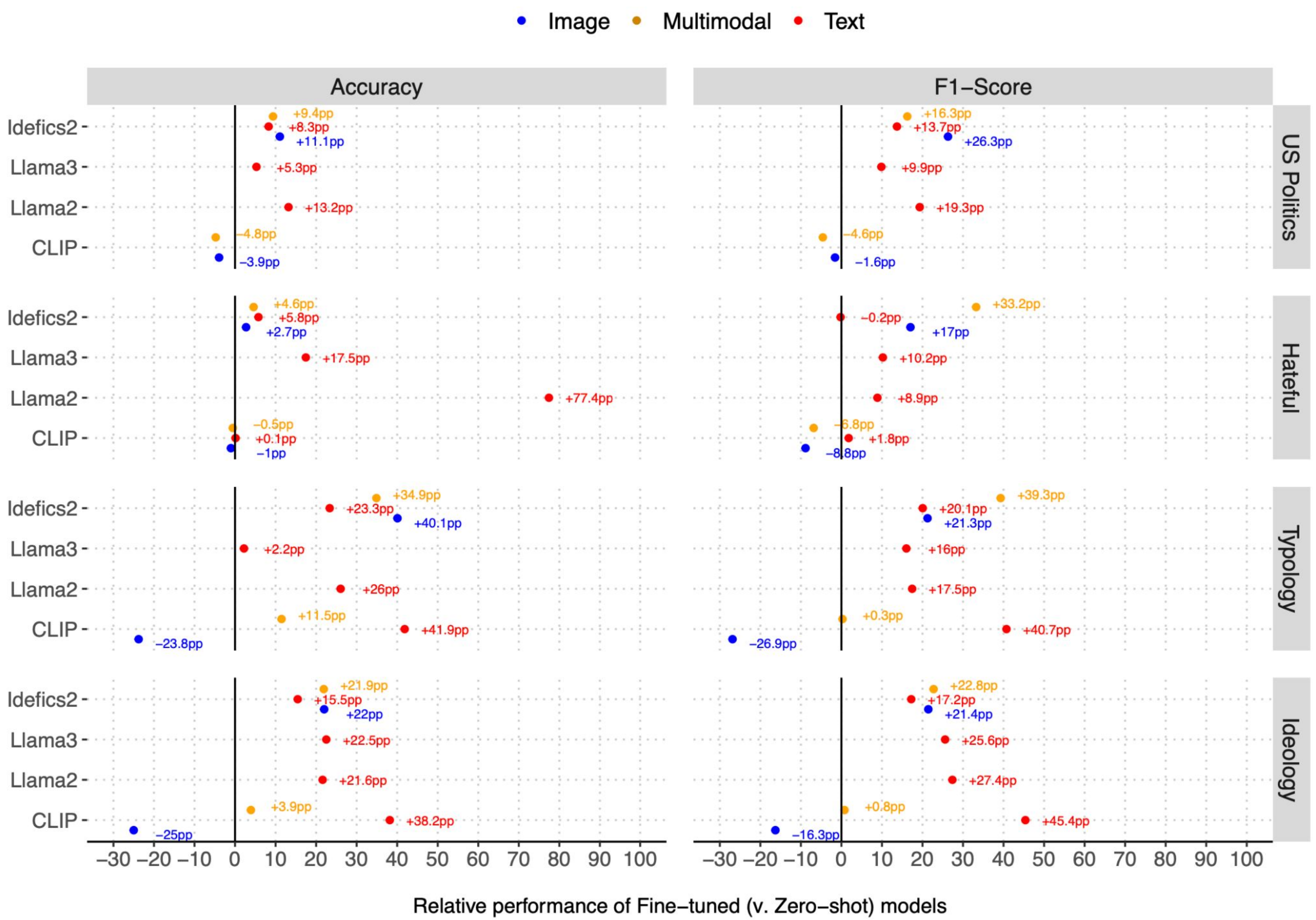
Results: overview



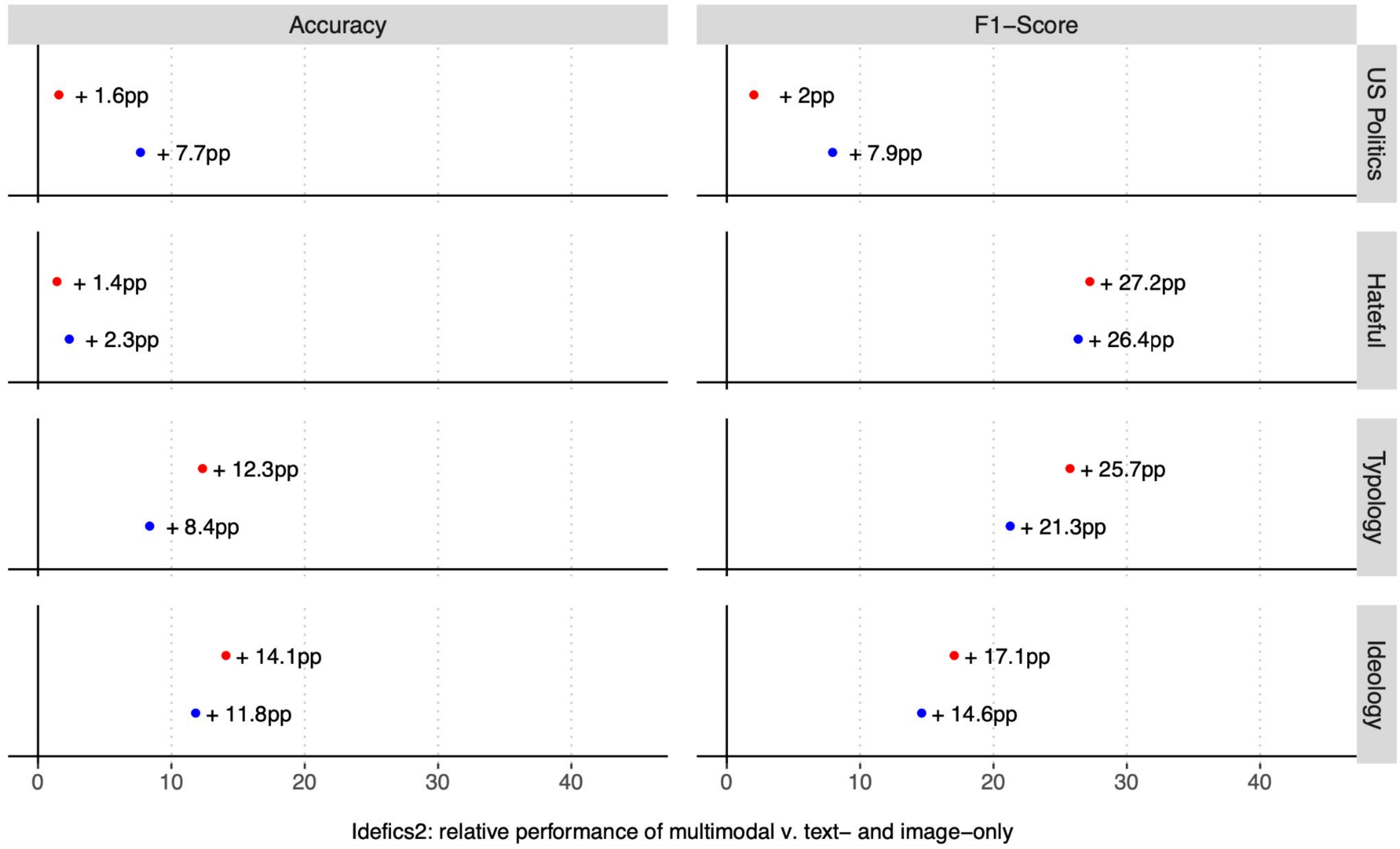
Results: overview



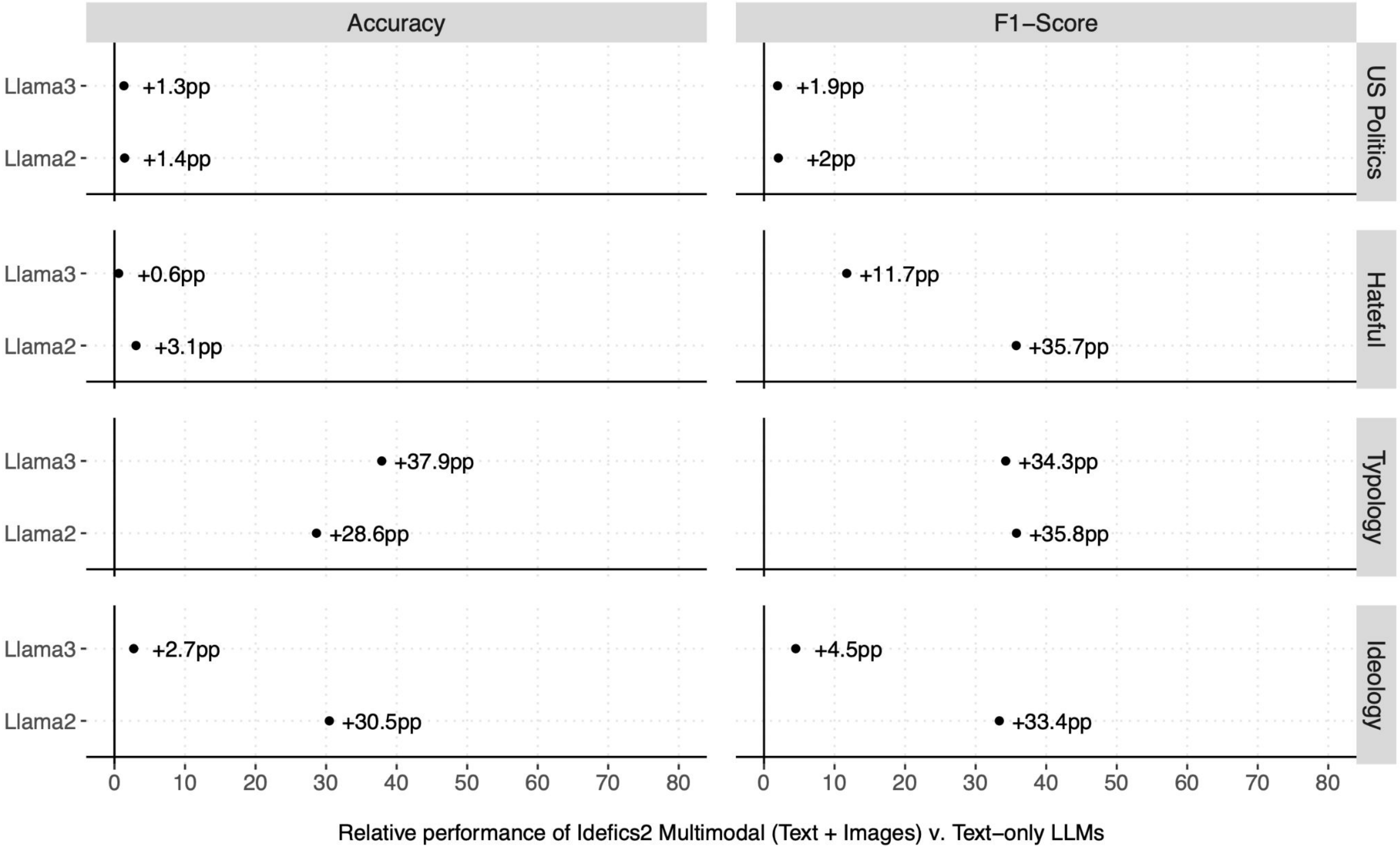
Results: fine-tuning makes a big difference



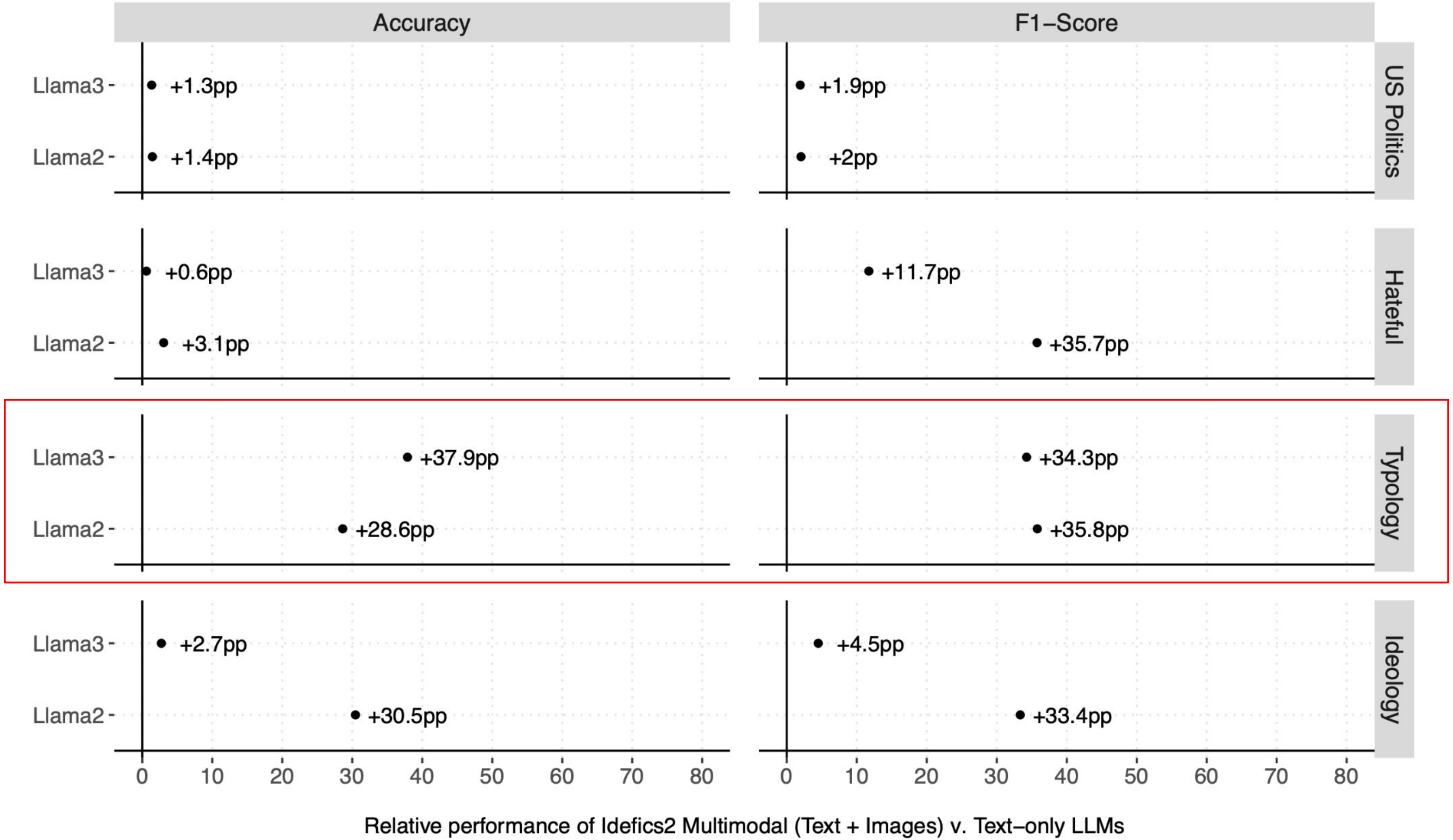
Results: VLM performs better in multimodal settings



Results: VLM outperforms SOTA open-source LLMs



Results: … particularly on more visual-dependent task



Limitations

- Current results based on only 1 fold
- Smaller Llama2 and Llama3 models (7B v. 40-400B parameters)
- Only max. of 800 text tokens per video
- Probably some more prompt engineering is needed
- NEW promising open-source VLM: Idefics3-8B-llama3

Conclusions and next steps

- Multimodality helps to improve performance (v. text-only LLMs)
- Particularly on more visual-dependent tasks
- Next steps:
    - address limitations discussed in previous slide (folds, prompts, add larger/new models)
    - run the same computational experiments on the missing target variables for the YouTube dataset; and also on the Interest Group Twitter data
    - potentially adding one more dataset (TikTok) with higher visual dependence