

Understanding (Dark) Humour with Internet Meme Analysis

TheWebConf'24 Tutorial

Roy Ka-Wei Lee

Assistant Professor
Singapore University
of Technology and
Design (SUTD)

Mingshan Hee

PhD Student
Singapore University
of Technology and
Design (SUTD)

Rui Cao

PhD Student
Singapore
Management
University (SMU)

Tanmoy Chakraborty

Associate Professor
Indian Institute of
Technology Delhi
(IIT Delhi)



WARNING: The following talk contain act of violence and discrimination that may be disturbing to some participants. Discretion is advised

What is a Meme?

- “An amusing or interesting item (such as a captioned picture or video) or genre of items that is spread widely online, especially through social media.” - Merriam-Webster



The dancing baby, 1995



Do
tutorial
slides early

Do slides
3 hours
before
the tutorial

Most popular meme template
all time (imgflip)

What make a Meme?

- Composed of **funny images and text**, which are often used as a form of **social or political commentary**. A meme will generate hundreds of different interpretations as it **gets passed around**, usually via social media, text and email, as people put their own personal spins on a popular meme **template**.
 - **Funny or humorous** (at least to some audience)
 - **Multimodal**
 - **Convey a certain message or ideology**
 - **It is shared or it get viral**
 - **It is usually template-based**



Meme used in Social Movements

- Memes are very effective tools to spread ideas and beliefs



Images of rice and bunnies were used by women in China to spread word about the #MeToo movement in the country



Bernie Sanders memes being used in many settings. Help in advocating his political beliefs

Meme with a Darker Twist

- **Hateful memes** - target certain communities and or individuals by portraying them in a derogatory manner
- **Misinformation memes** - spread misinformation using humor



Can AI understand Memes?

You

In Canada, They use special stickers to slow the car



In my country, We use the 3D techniques



Explain this meme

ChatGPT

This meme humorously contrasts road safety measures between two countries using two different techniques.

Top Part (Canada):

- **Image:** Shows a man applying a sticker on the road that looks like a pothole.
- **Text:** "In Canada, They use special stickers to slow the car."
- **Explanation:** In Canada, they use optical illusions (stickers) that look like potholes to make drivers slow down, enhancing road safety.

Bottom Part (Other Country):

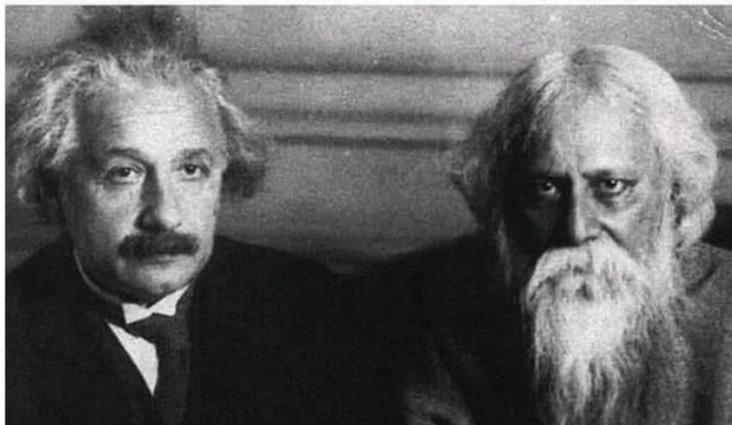
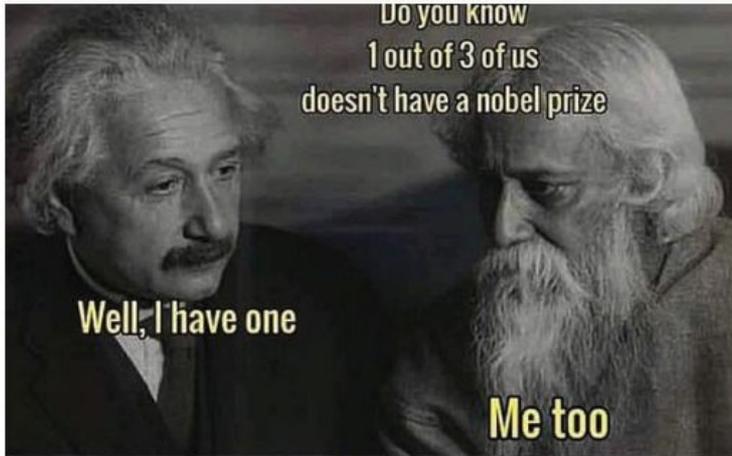
- **Image:** Shows actual potholes on a road with vehicles navigating around them.
- **Text:** "In my country, We use the 3D techniques."
- **Explanation:** In contrast, the meme suggests that in the creator's country, real potholes serve as a natural way to slow down traffic. This is a humorous way to point out the poor road conditions and lack of maintenance, implying that these unintentional "3D techniques" (actual potholes) slow down vehicles just as effectively, if not more so, than the stickers.

The humor lies in the exaggerated comparison of sophisticated safety measures in Canada with the unintended but effective "3D" potholes in the meme creator's country.



Can AI understand Memes?

You



Explain this meme

ChatGPT

This meme humorously depicts a conversation between two famous Nobel Prize winners, Albert Einstein and Rabindranath Tagore, highlighting an unspoken third person.

Top Panel:

- **Image:** Shows Albert Einstein and Rabindranath Tagore in conversation.
- **Text:**
 - Tagore: "Do you know 1 out of 3 of us doesn't have a Nobel Prize."
 - Einstein: "Well, I have one."
 - Tagore: "Me too."

Bottom Panel:

- **Image:** A photograph of Albert Einstein and Rabindranath Tagore sitting together, revealing that there is no third person.

Explanation: The meme humorously suggests that there is a third, unnamed person who doesn't have a Nobel Prize, while both Einstein and Tagore proudly affirm that they do. The punchline is that there is no third person in the photo, making the statement technically true but playfully misleading. The humor lies in the clever wordplay and the setup that implies a third party when there is none, emphasizing the accolades of Einstein and Tagore while poking fun at the concept of statistical statements.



Can AI understand Memes?

You



STRAITSTIMES.COM

\$255 million in Medisave top-ups to be given to Pioneer and Merdeka Generation Seniors in July



Explain this meme



ChatGPT

This meme humorously comments on the anticipation of a government official appearing to announce financial benefits for senior citizens in Singapore.

Top Panel:

- **Image:** Shows a newspaper headline with elderly people sitting on a bench.
- **Text:** "WITH THE GOODIES COMING IN"
- **Headline:** "\$255 million in Medisave top-ups to be given to Pioneer and Merdeka Generation Seniors in July."
- **Explanation:** The "goodies" refer to the financial benefits (Medisave top-ups) that are about to be distributed to senior citizens.

Bottom Panel:

- **Image:** Shows a government official at a podium.
- **Text:** "VERY SOON HE WILL APPEAR LIAO"
- **Explanation:** The phrase "very soon he will appear liao" is Singlish (a colloquial form of English spoken in Singapore), indicating that the official will soon make an appearance to announce or talk about these benefits.

Overall Explanation: The meme pokes fun at the predictability of government announcements and the officials who deliver them. It suggests that whenever there are financial benefits ("goodies") to be distributed, it's expected that a government official will soon make a public appearance to take credit or inform the public. The use of Singlish adds a local flavor and humor to the meme, making it relatable to Singaporeans.



Why is Challenging to Understand Meme?

- **Multimodality** - Recognize the text and objects in the images
- **Complex reasoning** - Interaction between modalities (and even the reader!)
- **Cultural context understanding** - Recognize the cultural background of the objects in the meme and the message convey

Why is Challenging to Understand Meme?

- **Multimodality** - Recognize the text and objects in the images
- **Complex reasoning** - Interaction between modalities (and even the reader!)
- **Cultural context understanding** - Recognize the cultural background of the objects in the meme and the message convey

Research are done
to advance these
aspects!



Tutorial Outline

Part I Introduction

❓ *What do you meme?*

Part II Meme Analysis Methods

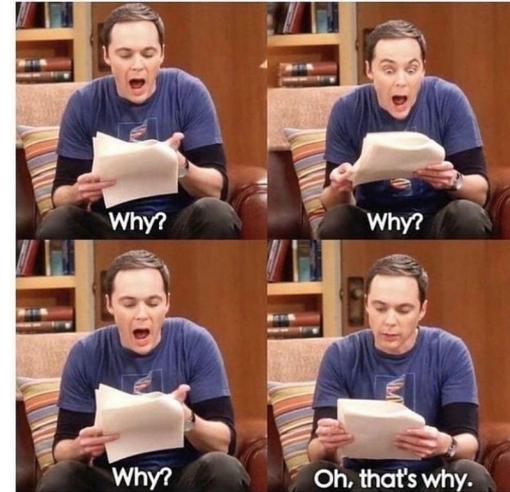
❓ *What methods are available?*

Part III Hands-On with MATK

❓ *How to do it?*

Part IV Gaps & Opportunities

❓ *What's next?*



- Overview of Internet meme culture (including dark scenarios)
- Research challenges in understanding memes
- Existing research efforts and directions

Tutorial Outline

Part I Introduction

① *What do you
meme?*

Part II Meme Analysis Methods

① *What methods
are available?*

Part III Hands-On with MATK

① *How to do it?*

Part IV Gaps & Opportunities

① *What's next?*



- Meme analysis tasks such as meme classification and explanation, etc.
- Meme analysis models leveraging Large Language Models (LLMs) and Large Multimodal Models (LMMs)

Tutorial Outline

Part I Introduction

① *What do you
meme?*

Part II Meme Analysis Methods

① *What methods
are available?*

Part III Hands-On with MATK

① *How to do it?*

Part IV Gaps & Opportunities

① *What's next?*



Listening
to 3
hours lecture

Guided
hands-on
session

- Hands-on session with Meme Analysis Tool Kit (MATK)
 - Data Preprocessing
 - Applying various meme analysis models for various tasks

Tutorial Outline

Part I Introduction

① *What do you meme?*

Part II Meme Analysis Methods

① *What methods are available?*

Part III Hands-On with MATK

① *How to do it?*

Part IV Gaps & Opportunities

① *What's next?*



- Discussion on the gaps in existing meme analysis researches
- Dialogue on emerging trends and research opportunities

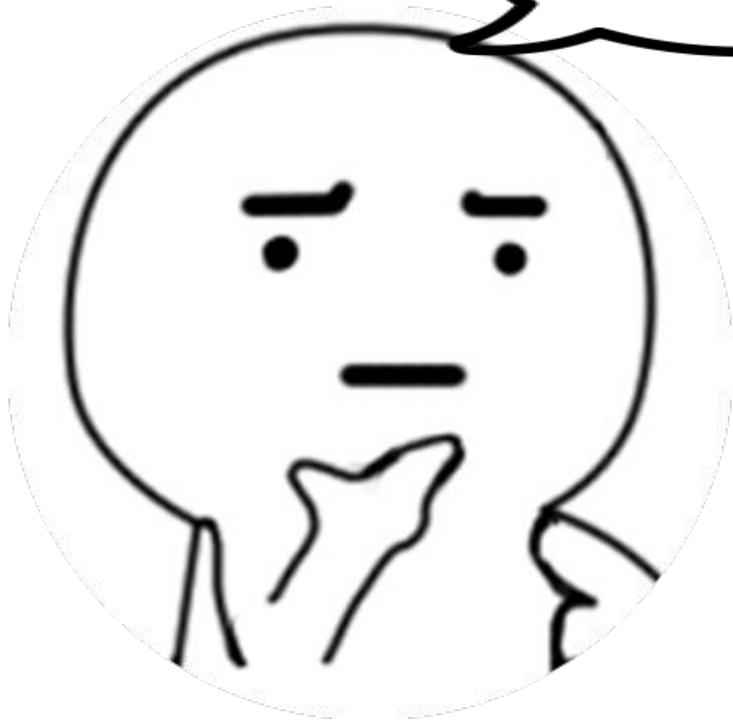
Part II: Meme Analysis Methods

Multimodal Memes: Images with Short Texts

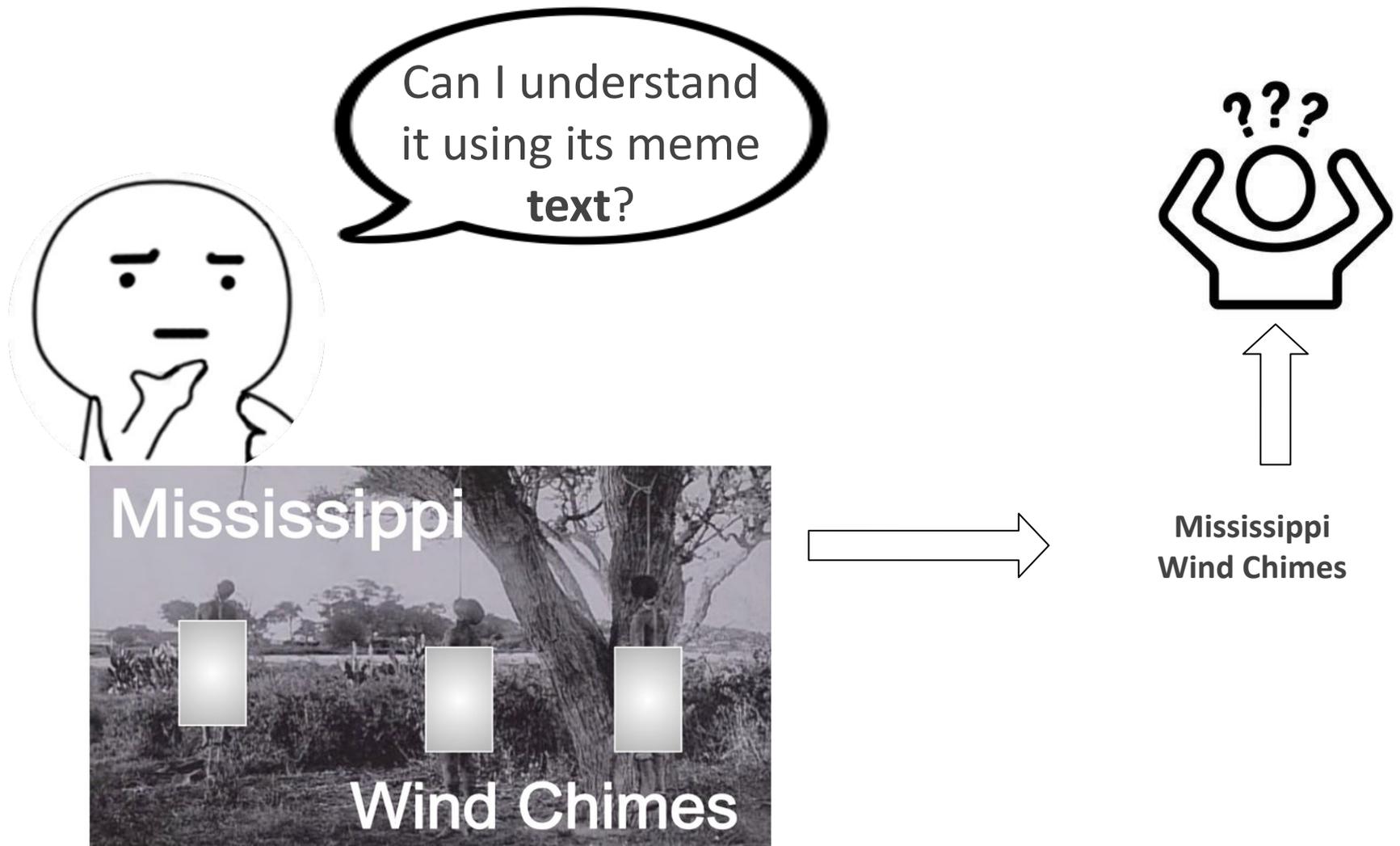


Understanding Multimodal Memes

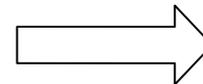
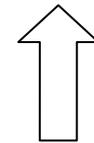
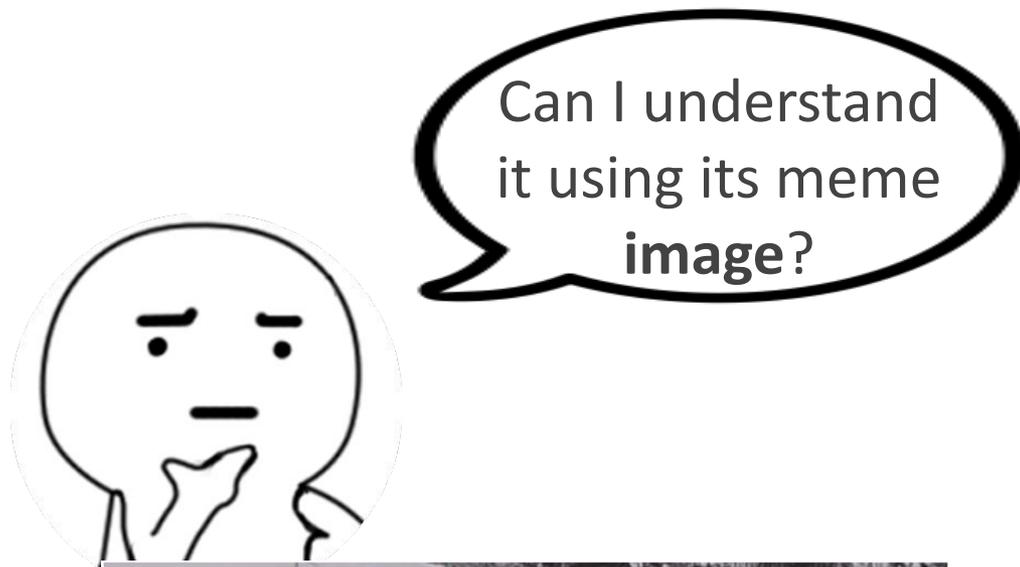
What is the meaning of the meme?



Understanding Multimodal Memes

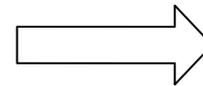
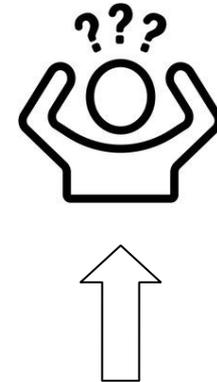
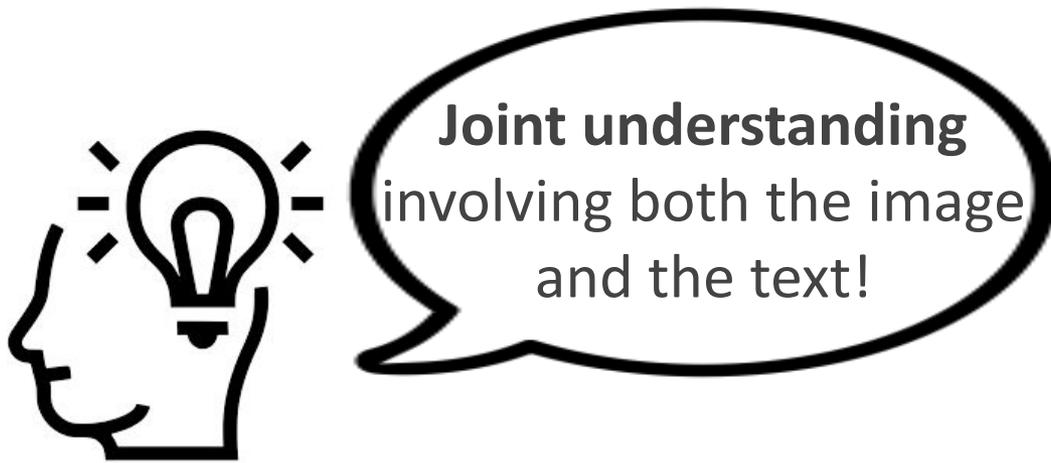


Understanding Multimodal Memes



Three people hanging over trees

Understanding Multimodal Memes

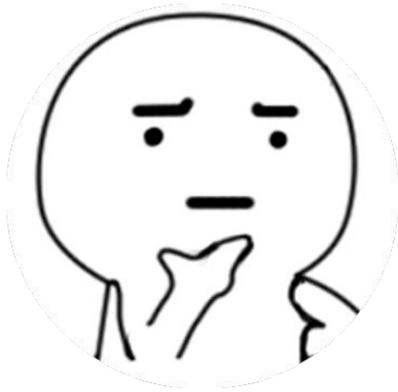


The meme is comparing the hanged people to wind chimes

Understanding Multimodal Memes



Understanding Multimodal Memes



The hanged people in the image were died from the slaughter of black people in Mississippi

The meme is comparing the hanged people to wind chimes



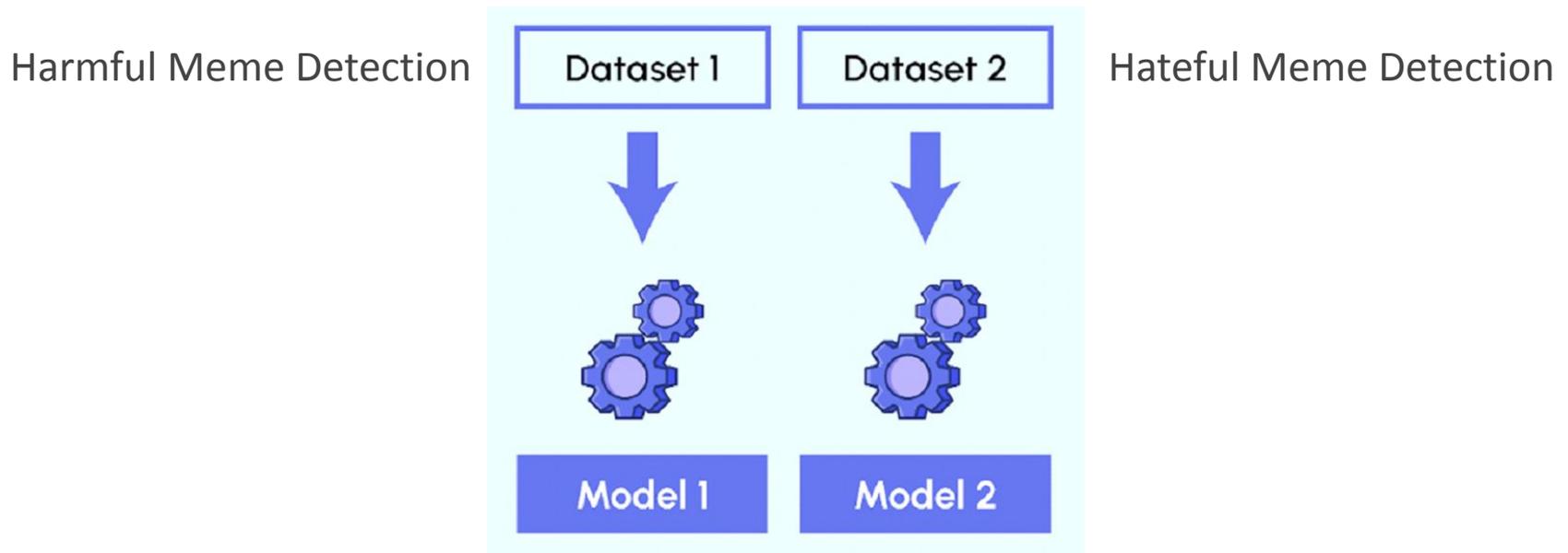
The meme is making joke of the slaughter of black people, by comparing them to wind chimes. **The meme is hateful.**

Major Challenges: Meme Understanding

- Joint Vision-Language Understanding
 - Interactions between the image and the meme text
 - Visual metaphors
 - E.g., referring “*wind chimes*” to “*hanged people*”
- External Knowledge beyond the Meme
 - Cultural background knowledge, commonsense
 - E.g., the image is related to the slaughter of black people
- Complex reasoning
 - Reasoning across modalities; reasoning with knowledge
 - E.g., the meme is making joke of the slaughter of black people; it is hateful

Initial Solutions

- Specially designed architectures for different challenges
 - Fusion mechanism: acquire vision-language understanding
 - Train models from scratch with task-specific data



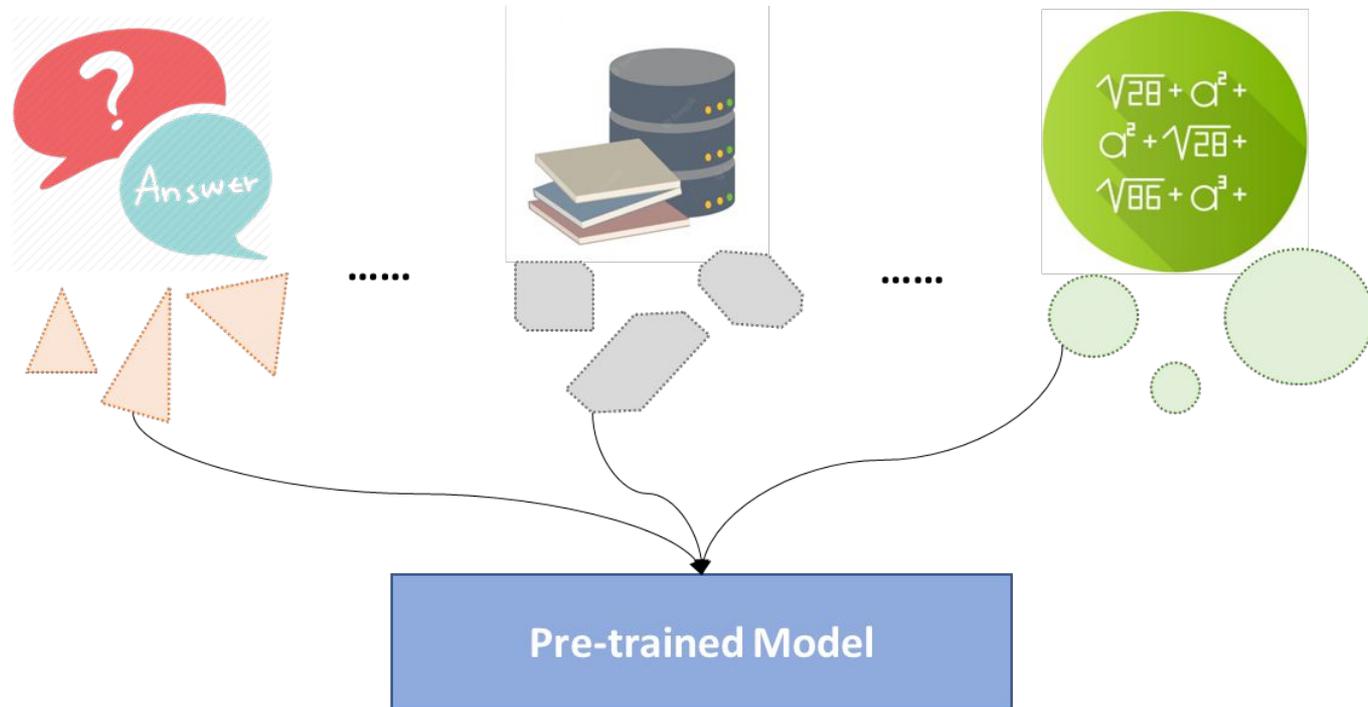
Scarcity of Data

- Three commonly used dataset for hateful meme detection
 - Training data: less than 10k
 - Annotating large-scale datasets: expensive

Datasets	Train		Test	
	#Hate.	#Non-hate.	#Hate.	#Non-hate.
FHM	3,050	5,450	250	250
HarM	1,064	1,949	124	230
MAMI	5,000	5,000	500	500

- Impact
 - Overfitting to training data
 - Incapability of generalization

Pre-trained Models



Pre-trained with **large amount** of data from **different sources (mostly unsupervised)**

Pre-trained models learn universal representations / good at generalization

Outline

- Using PTMs for meme analysis
 - Brief introduction to PTMs
 - Utilization of PTMs: fine-tuned/frozen; a single/a composition
 - Hateful meme detection: testbed
- Beyond classification tasks
 - Interpretation of memes
 - Counter hateful memes

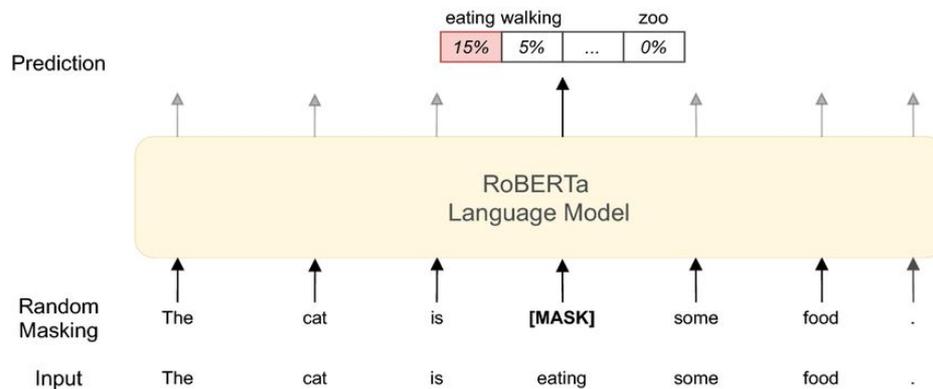
Outline

- Using PTMs for meme analysis
 - Brief introduction to PTMs
 - Utilization of PTMs: fine-tuned/frozen; a single/a composition
 - Hateful meme detection: testbed
- Beyond classification tasks
 - Interpretation of memes
 - Counter hateful memes

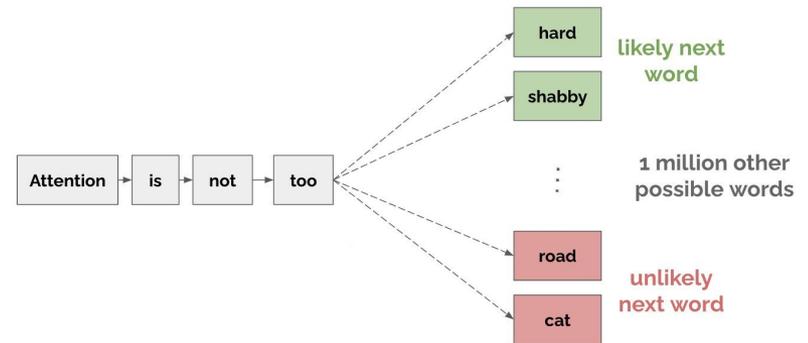
Pre-trained Language Model

- For Natural Language Processing
- Unsupervised pre-training

Masked Language Modeling

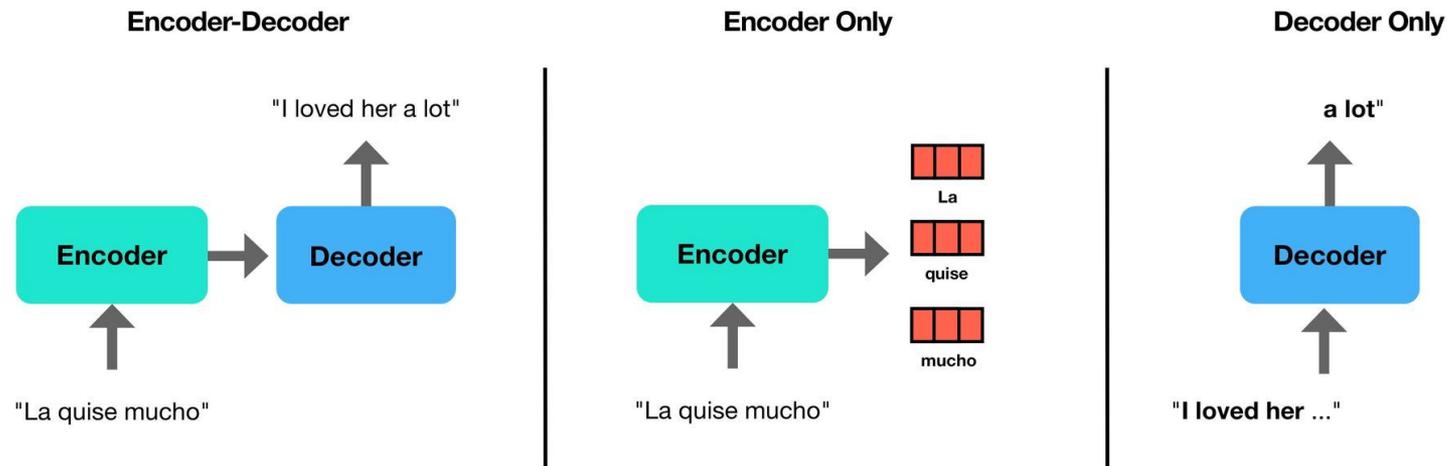


Next Word Prediction



Pre-trained Language Model

- Transformer architectures



Pre-trained Language Models

- Architectures of recent PT-LMs

- Encoder-only

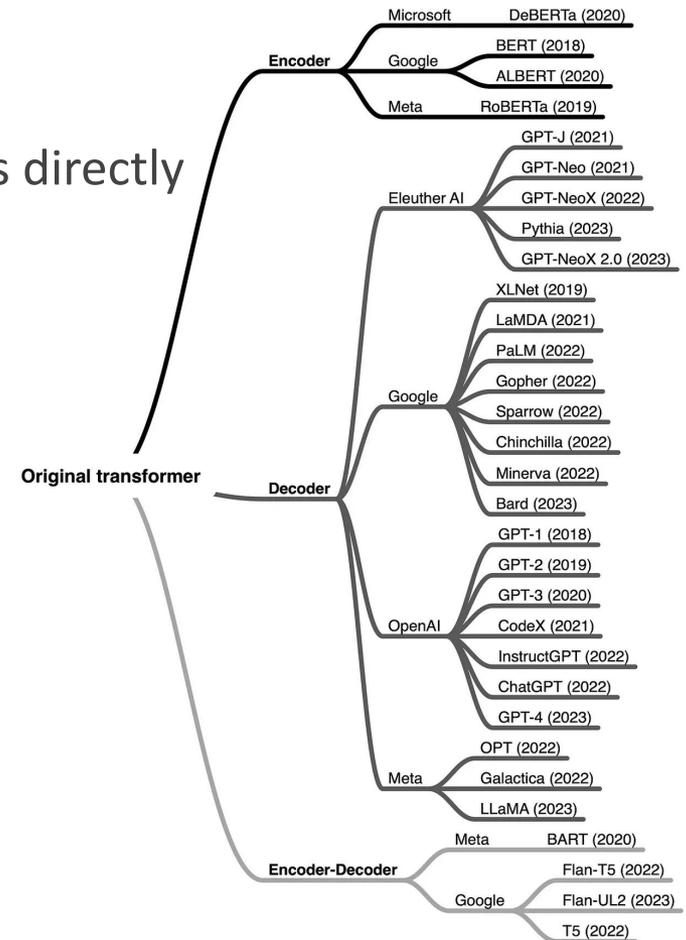
- encode texts; cannot generate texts directly

- Decoder-only

- strong at text generation

- Encoder-Decoder

- powerful at natural language understanding tasks



Pre-trained Language Model

- Instruction tuning
 - finetuning PT-LMs on a collection of datasets described via instructions
 - improve generalization to unseen tasks

Finetune on many tasks (“instruction-tuning”)

Input (Commonsense Reasoning)

Here is a goal: Get a cool sleep on summer days.
How would you accomplish this goal?
OPTIONS:
 -Keep stack of pillow cases in fridge.
 -Keep stack of pillow cases in oven.

Target

keep stack of pillow cases in fridge

Input (Translation)

Translate this sentence to Spanish:
The new office building was built in less than three months.

Target

El nuevo edificio de oficinas se construyó en tres meses.

Sentiment analysis tasks

Coreference resolution tasks

...

Inference on unseen task type

Input (Natural Language Inference)

Premise: At my age you will probably have learnt one lesson.
Hypothesis: It's not certain how many lessons you'll learn by your thirties.
Does the premise entail the hypothesis?
OPTIONS:
 -yes -it is not possible to tell -no

FLAN Response

It is not possible to tell

Pre-trained Vision-Language Model

- For Vision-Language understanding (Image-Text)
- Unsupervised pre-training

Masked vision-language modeling

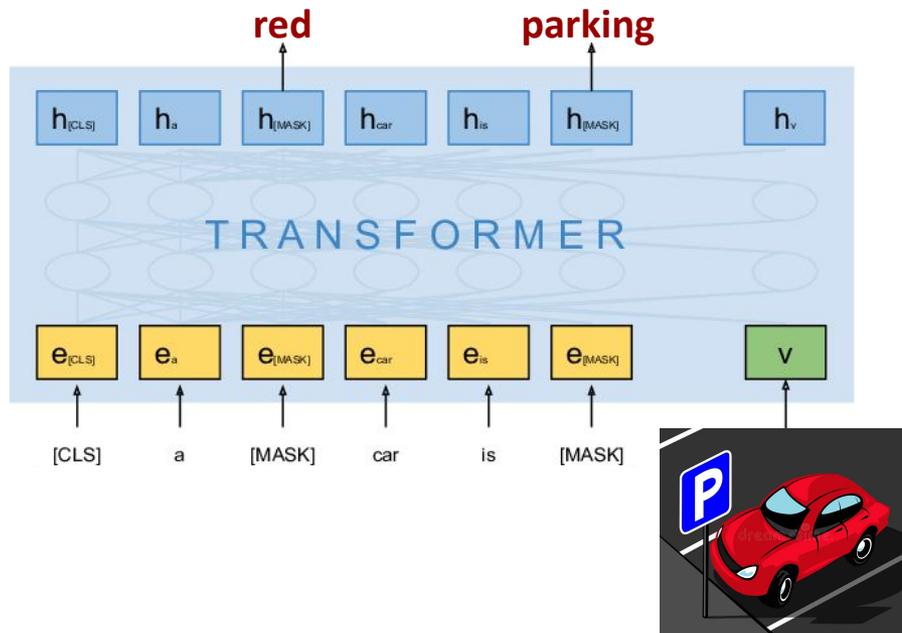
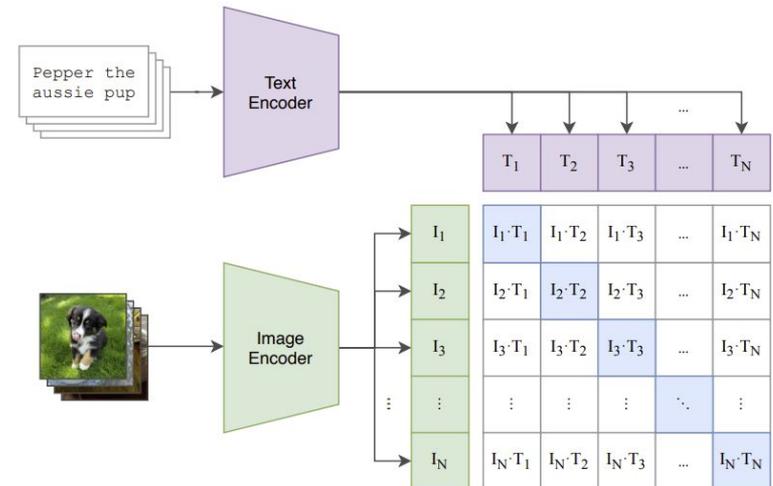


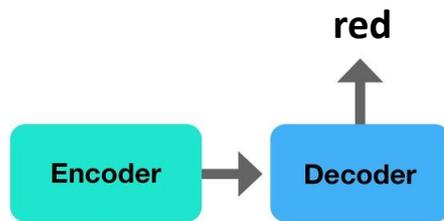
Image-Text Contrastive Learning



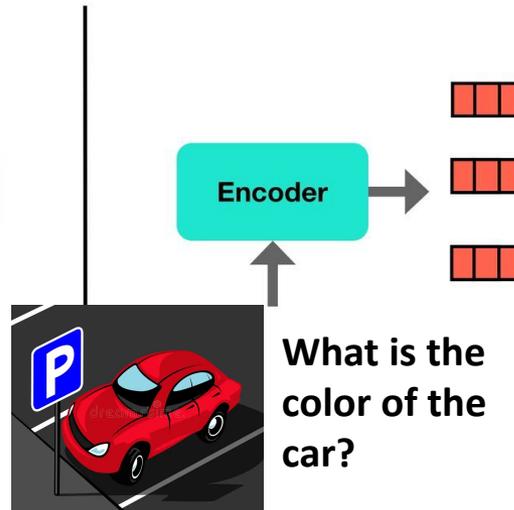
Pre-trained Vision-Language Model

- Transformer architectures

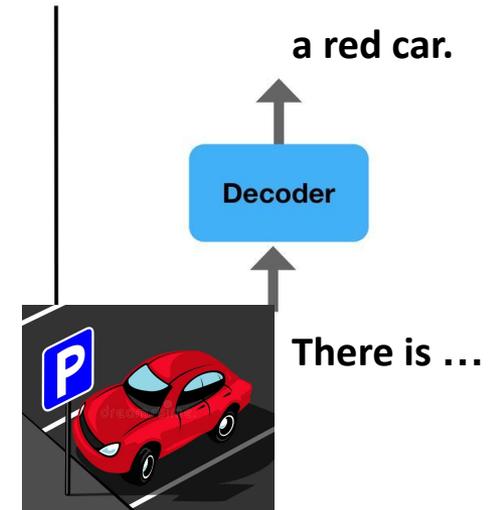
Encoder-Decoder



Encoder Only

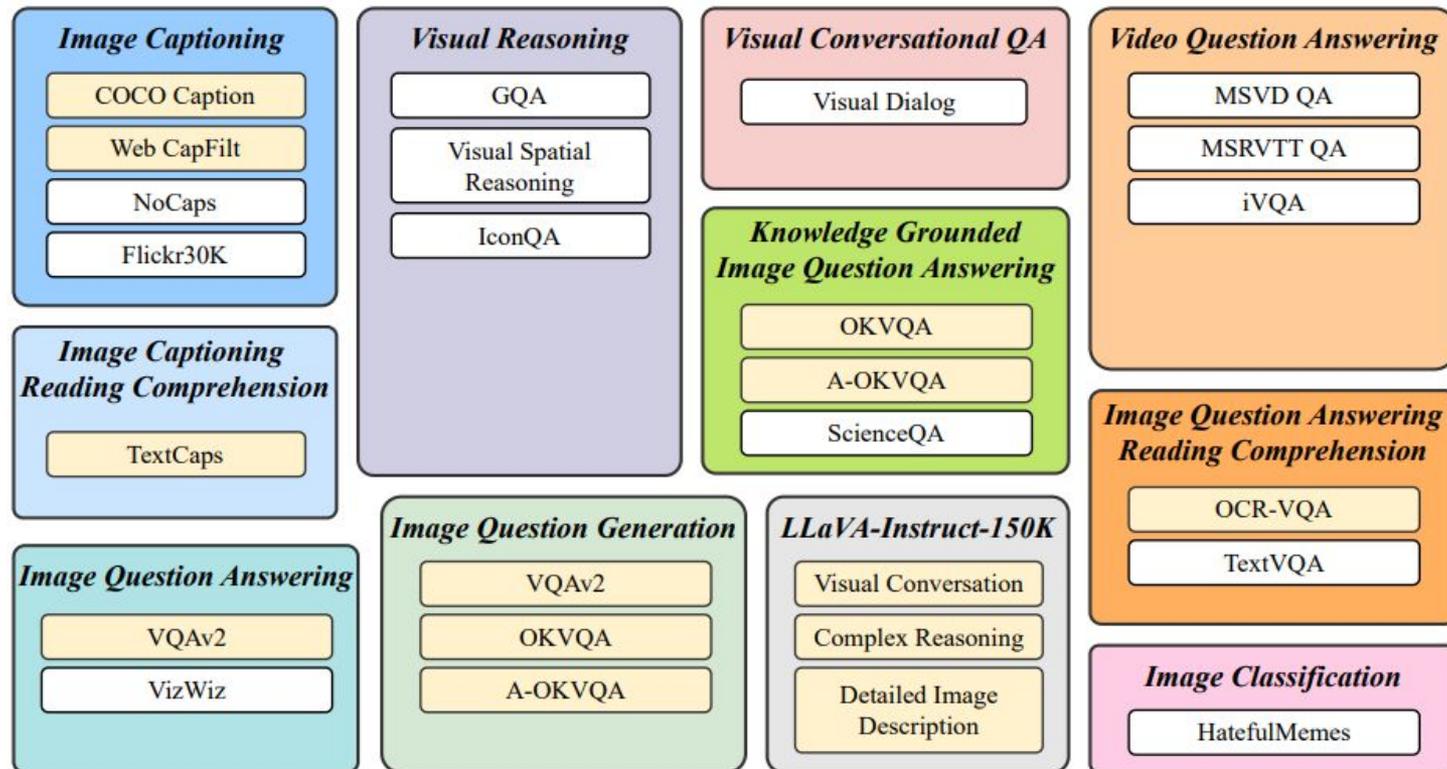


Decoder Only



Pre-trained Vision-Language Model

- Vision-Language instruction tuning
 - Fine-tuned on multiple vision-language tasks via instructions



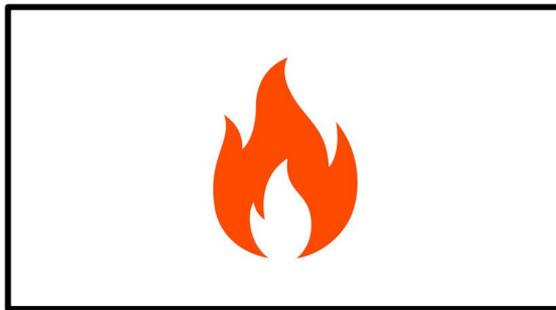
Outline

- Using PTMs for meme analysis
 - Brief introduction to PTMs
 - **Utilization of PTMs: fine-tuned/frozen; a single/a composition**
 - Hateful meme detection: testbed
- Beyond classification tasks
 - Interpretation of memes
 - Counter hateful memes

Power of Pre-trained Models (PT-VLMs)

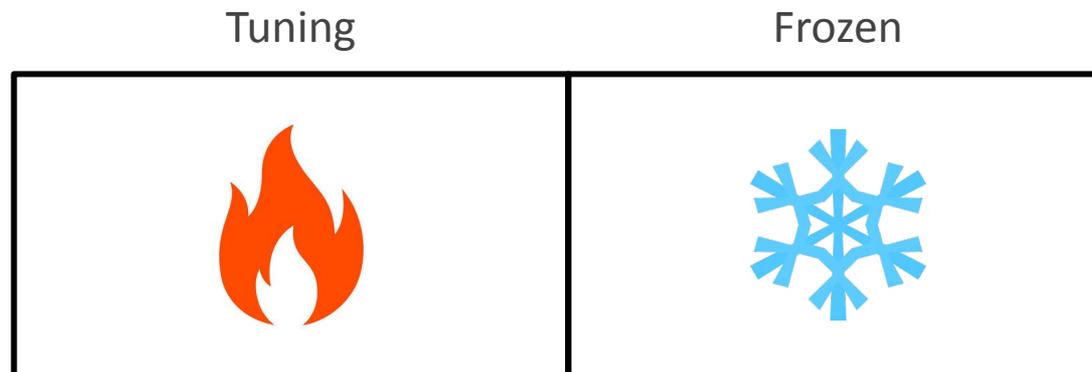
- Good tuning performance with down-stream task data
 - Expressive vision-language representation
 - VisualBERT, ViLBERT, LXMERT, OSCAR, UNITER ...

Tuning



Power of Pre-trained Models (PT-VLMs)

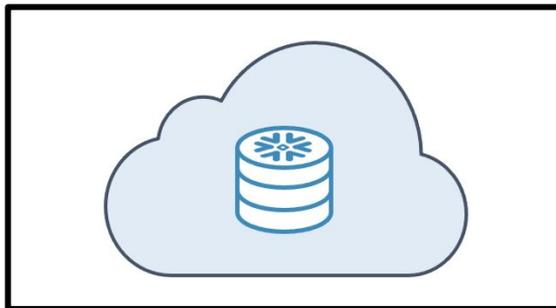
- Good tuning performance with down-stream task data
 - Expressive vision-language representation
 - VisualBERT, ViLBERT, LXMERT, OSCAR, UNITER ...
- Strong zero-shot performance with frozen pre-trained models
 - Without requiring any adaptation
 - BLIP-2, LLaVA, mPLUG-OWL, Instruct-BLIP ...



Power of Pre-trained Models

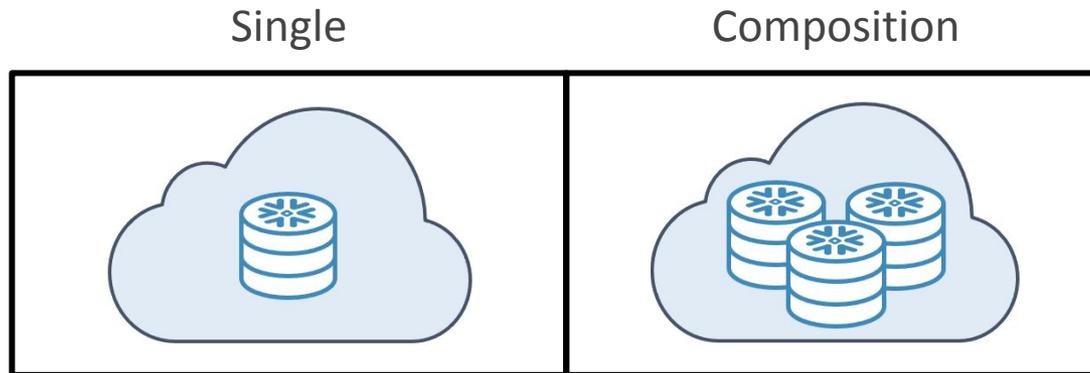
- Using a single model

Single



Power of Pre-trained Models

- Using a single model
- Using a composition of models
 - complex tasks requiring several skills
 - Hateful meme detection: visual metaphor understanding, comprehension of the definition of “*hateful*”, decoding the underlying meaning of hateful memes ...



Utilization of Pre-trained Models

Tuning 

Frozen 

Single



Composition



Meme Related Tasks

- Classification
 - offensive/harmful/hateful meme detection
 - protected category classification
 - vulnerable targets of hateful memes
- Generation
 - meme comprehension
 - sarcastic/hateful meme interpretation
 - decoding the underlying meaning of sarcastic/hateful memes
 - why sarcastic/hateful
 - Counter hateful memes

Outline

- Using PTMs for meme analysis
 - Brief introduction to PTMs
 - Utilization of PTMs: fine-tuned/frozen; a single/a composition
 - **Hateful meme detection: testbed**
- Beyond classification tasks
 - Interpretation of memes
 - Counter hateful memes

Hateful Meme Detection

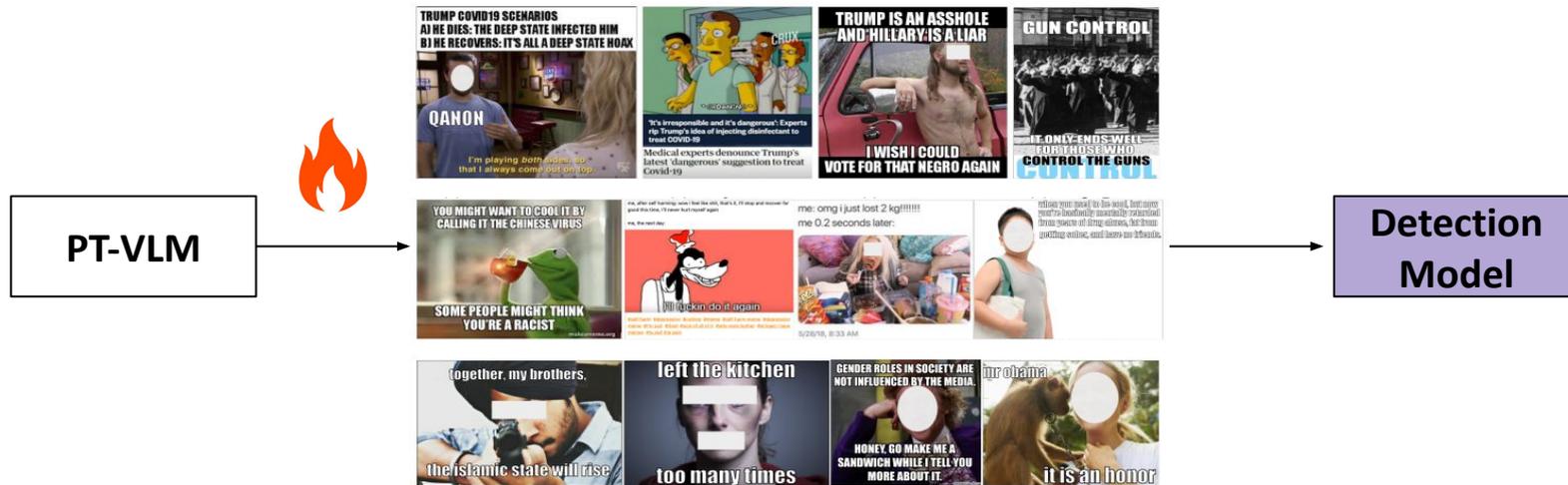
- Given a meme, predict whether it is hateful/non-hateful
 - hateful: attacks or uses discriminatory content targeting a person or a group based on their race, religion, gender, etc.
 - binary classification
 - vision-language classification task



Prevalent Approach: Finetuning PT-VLMs

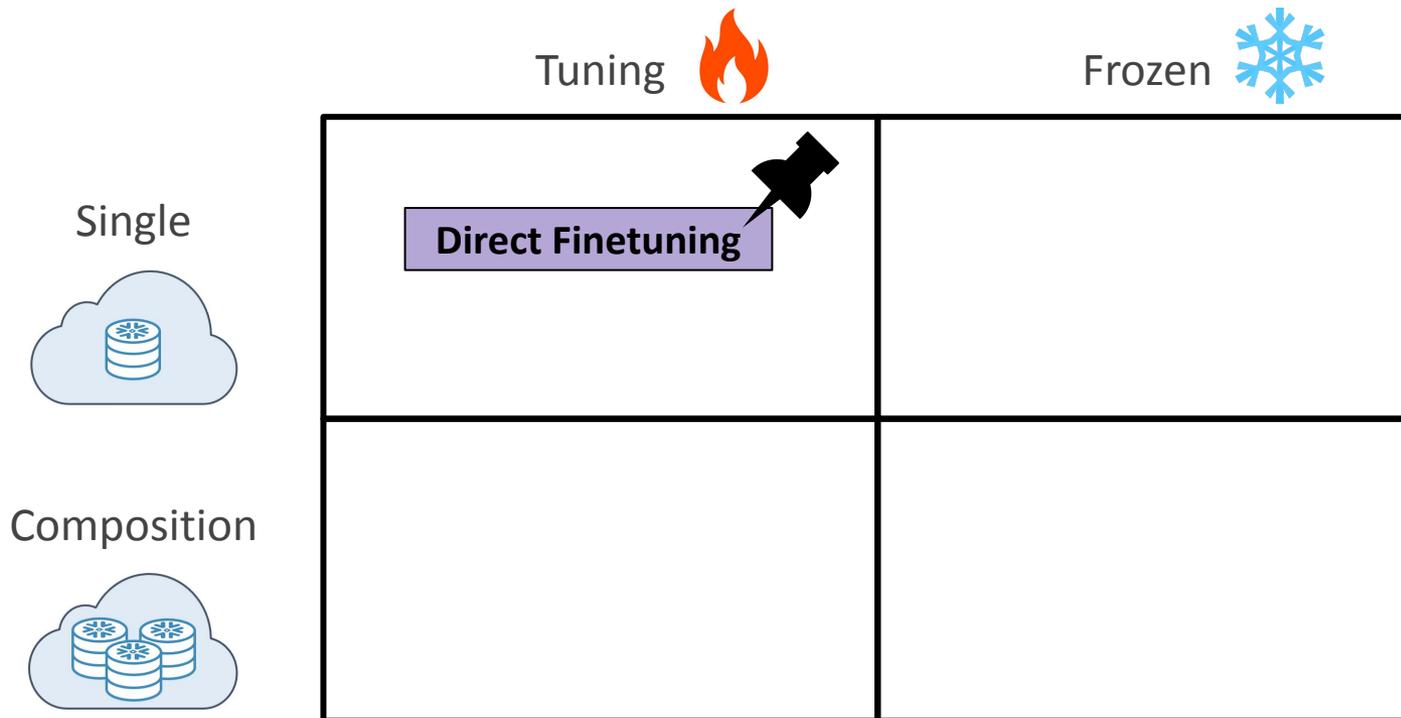
- Bridging the vision-language gap
 - Finetuning PT-VLMs with hateful meme detection data

Hateful Meme Detection Data



Prevalent Approach: Finetuning PT-VLMs

- Bridging the vision-language gap
 - Finetuning PT-VLMs with hateful meme detection data



Limitation of Direct Fine-tuning

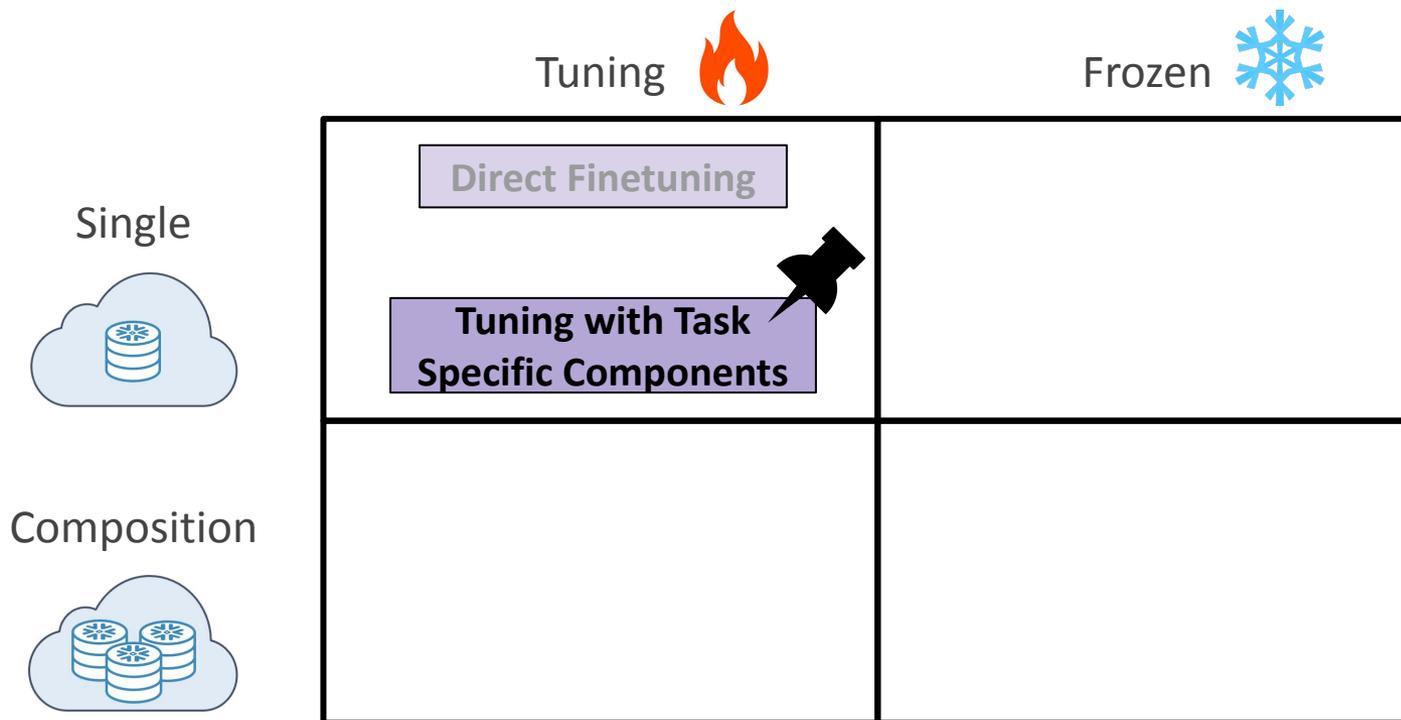
- Regard hateful meme detection as a general vision-language classification task
- Ignore task-specific characteristics
 - understanding targets/victims: important
 - meme texts and meme image: weakly aligned
- Incorporate task-specific components
 - task-specific components + PTMs



No black people in the text

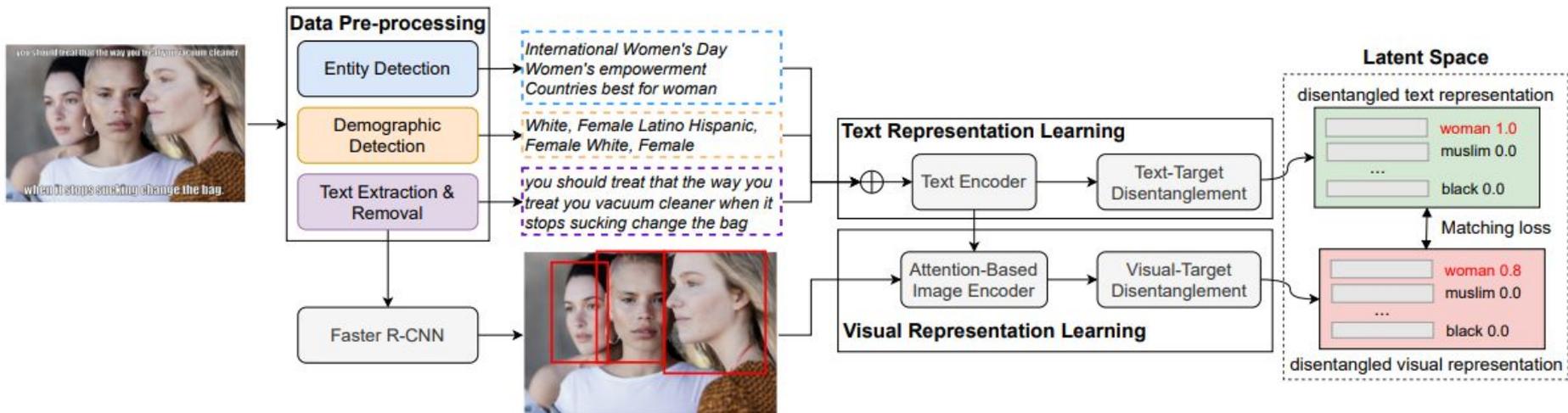
PTMs with Task-specific Components

- Tuning PTMs with task-specific components end-to-end



Tuning with Task-Specific Components

- DisMultiHate
 - disentangle the representations of hate speech related target entities



Lee, Roy Ka-Wei, et al. "Disentangling hate in online memes." *Proceedings of the 29th ACM international conference on multimedia*. 2021.

Tuning with Task-Specific Components

- Image target entity augmentation
 - Explicit mention of targets in texts: more aligned
- Task-specific component: target latent space
 - Force the generated joint representation (for meme image and text) can disentangle the target entities
 - Target aware representation



Entities

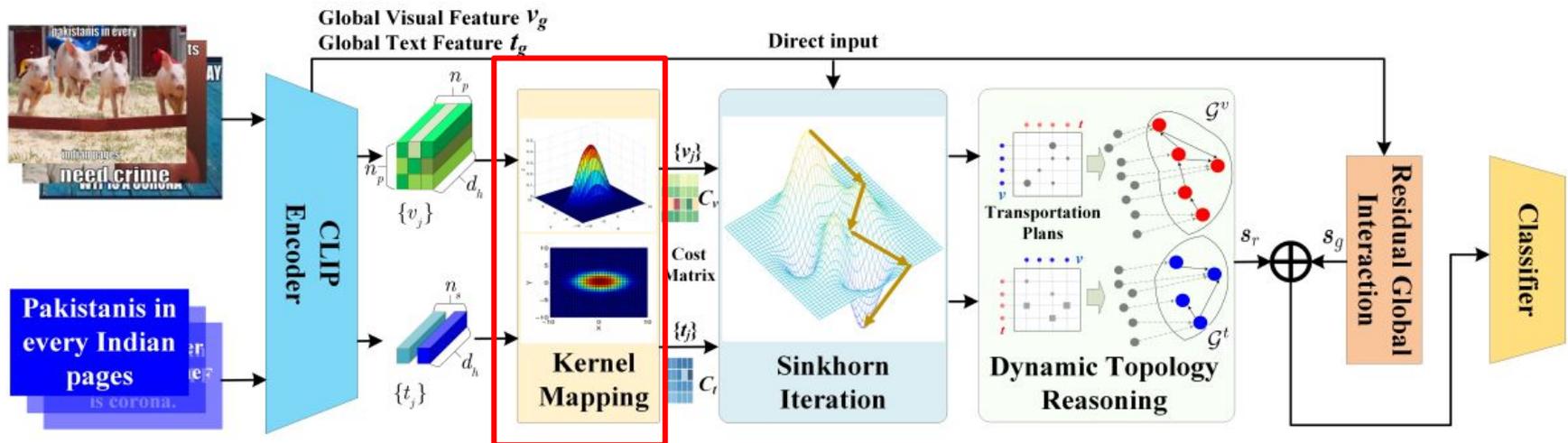
Black people slaughter in Mississippi
Premonitions
Black people lynched
....

10% absolute improvement over direct fine-tuning VisualBERT

Tuning with Task-Specific Components

- TOT: Topology-Aware Optimal Transport
 - formulate the cross-modal alignment problem as solutions for optimal transportation plans

cost matrix: the pairwise costs to align each elements in the image and the text

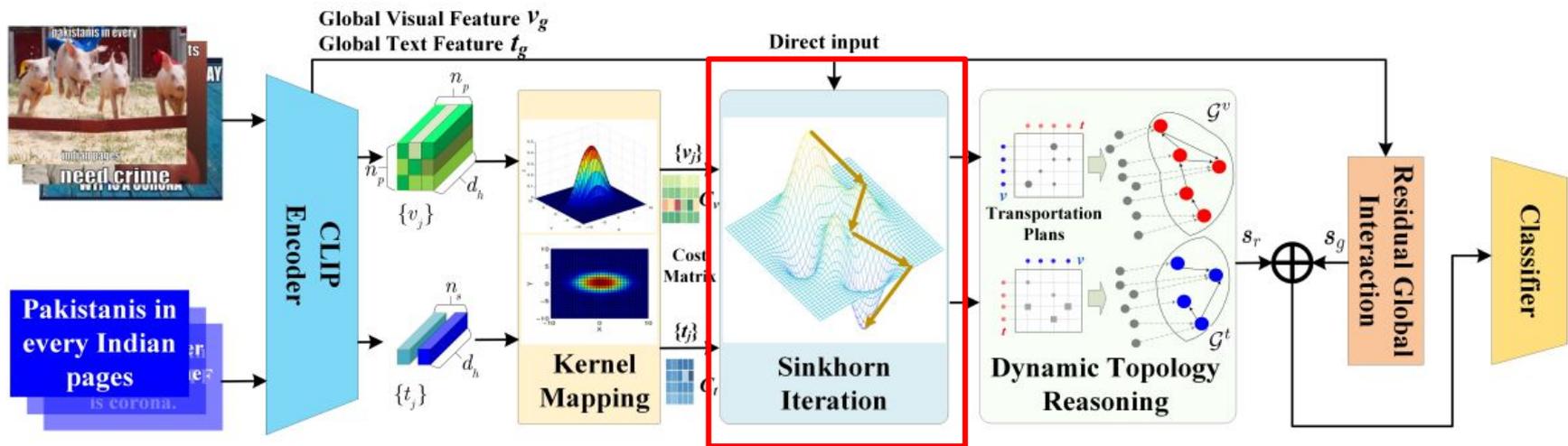


Zhang, Linhao, et al. "TOT: topology-aware optimal transport for multimodal hate detection." *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 37. No. 4. 2023.

Tuning with Task-Specific Components

- TOT: Topology-Aware Optimal Transport

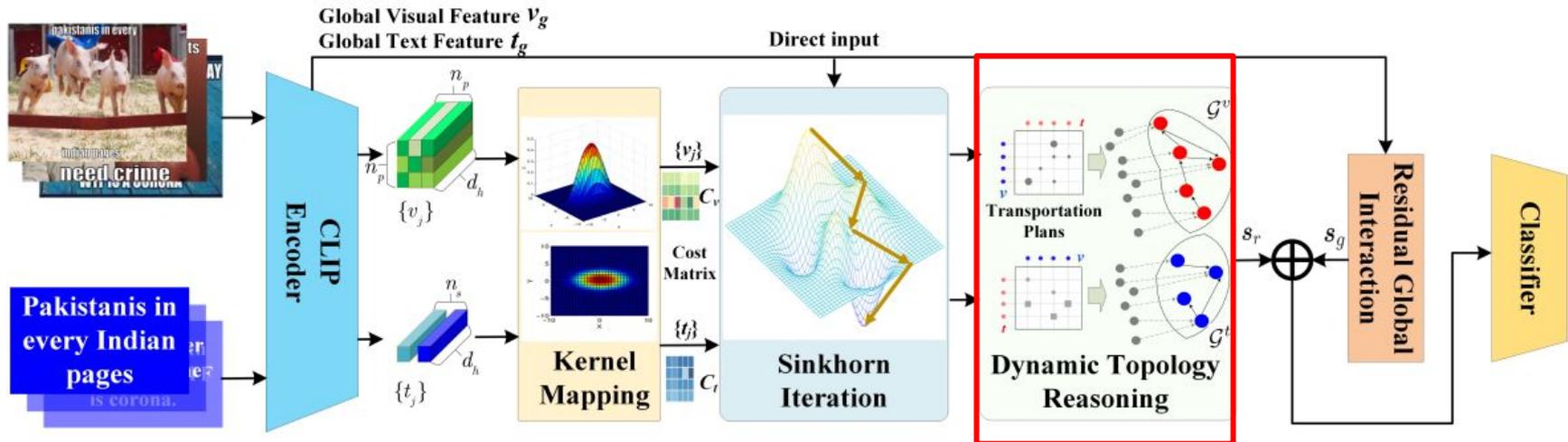
Transportation plan: minimize the cost



Tuning with Task-Specific Components

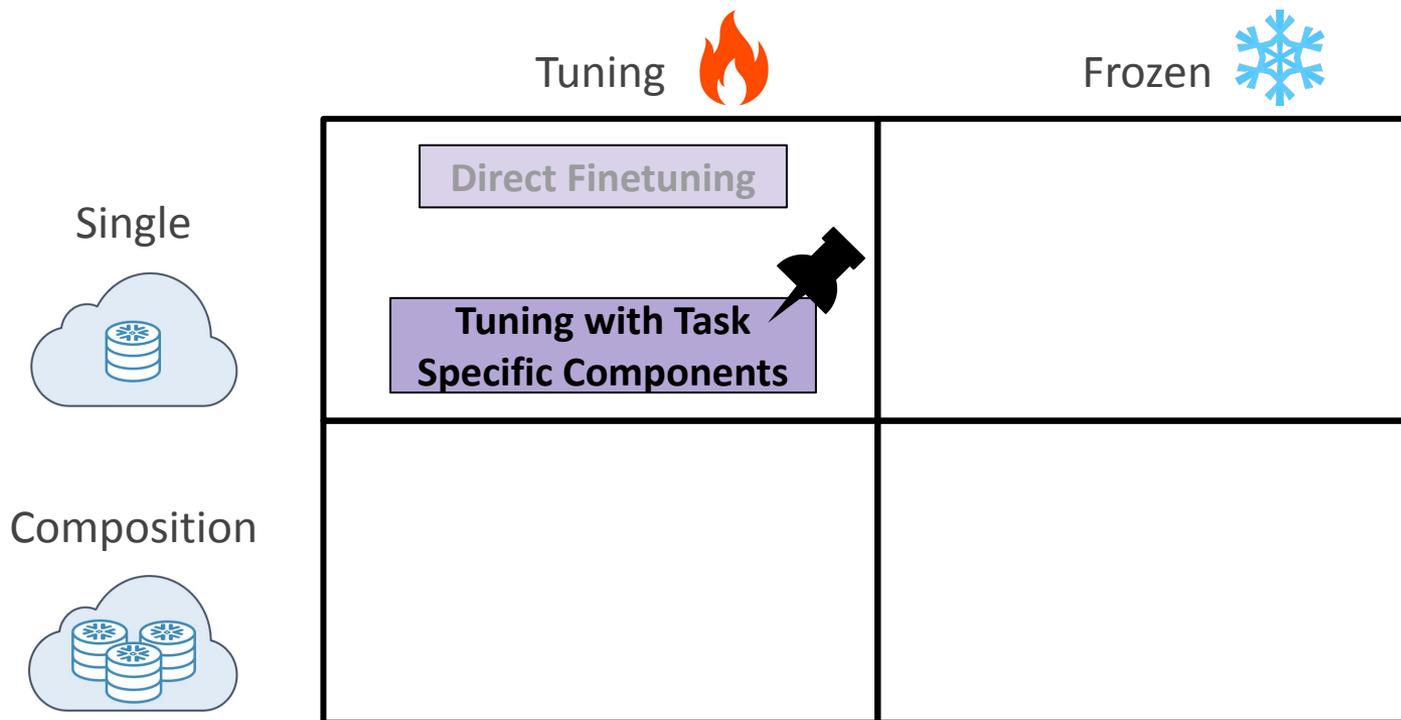
- TOT: Topology-Aware Optimal Transport

Aligned representation generation



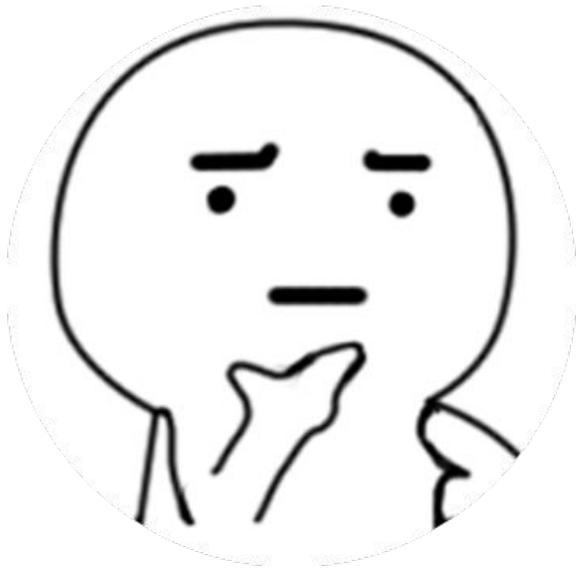
PTMs with Task-specific Components

- Tuning PTMs with task-specific components end-to-end



Need of Background Knowledge

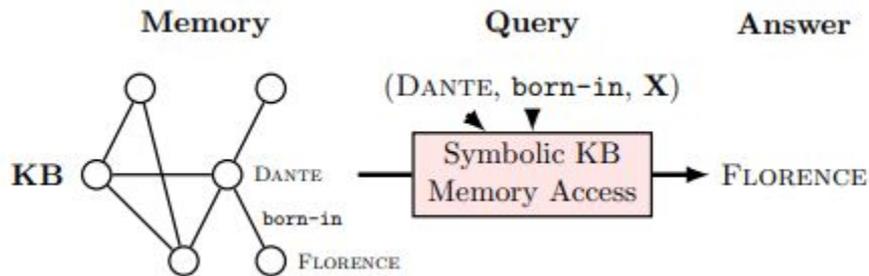
What is the smell
of roses?



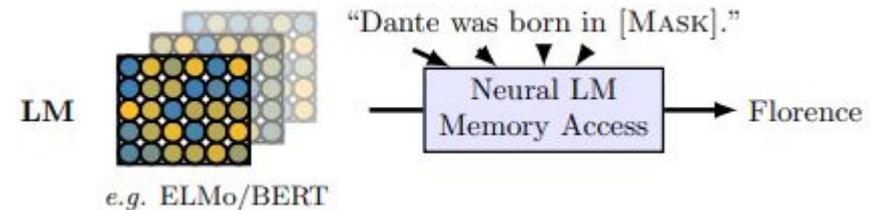
PT-LM: Knowledge Base

- Pre-trained on large corpus of data
 - implicit knowledge embedded in PT-LMs

Traditional Knowledge Base

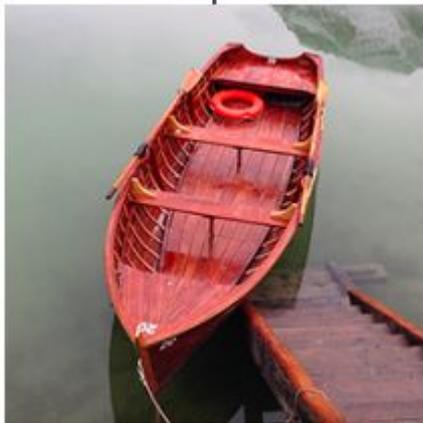


PT-LMs as Knowledge Base



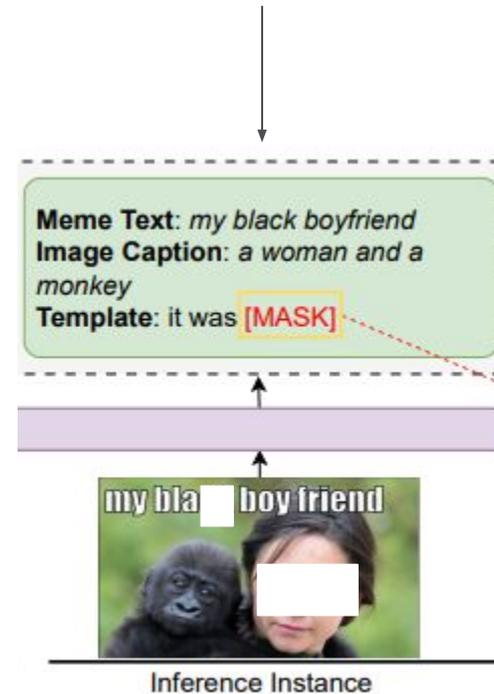
Prompting PT-LMs for Knowledge!

- Meme images are not comprehensible to PT-LMs
 - use a frozen PT-VLMs for converting images to textual representations



Prompting PT-LMs for Knowledge!

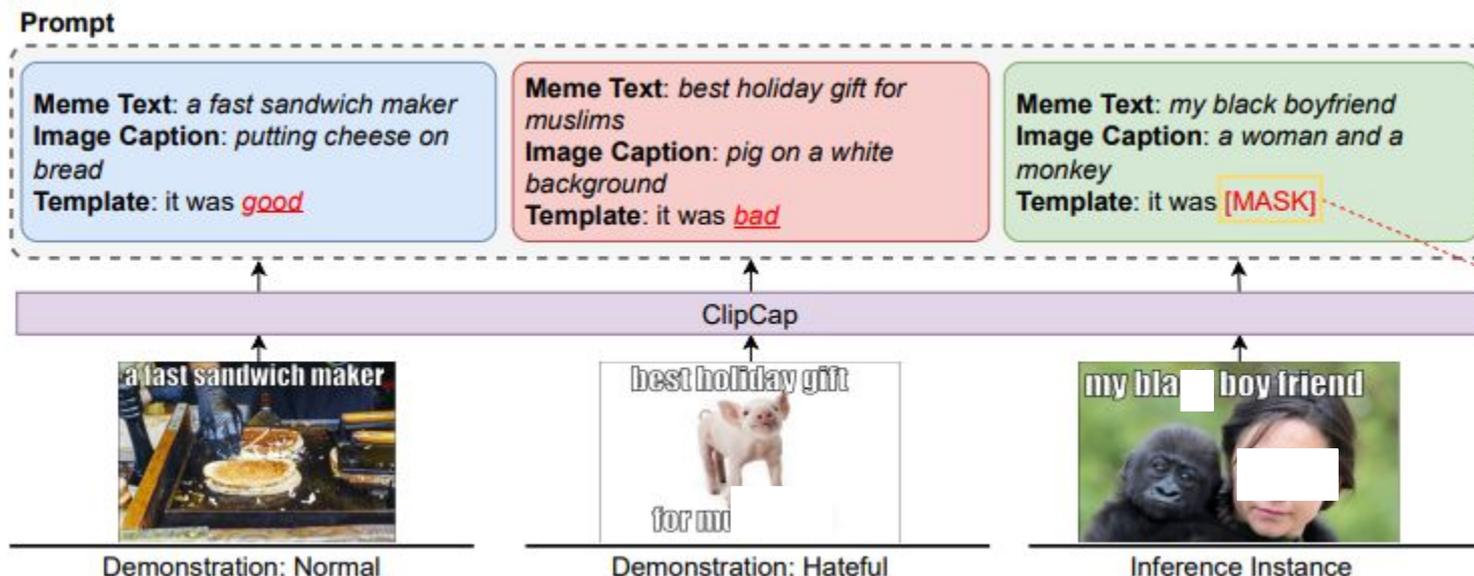
- PromptHate:
 - Hateful meme classification → Masked Language Modeling



Cao, Rui, et al. "Prompting for multimodal hateful meme classification." *arXiv preprint arXiv:2302.04156* (2023).

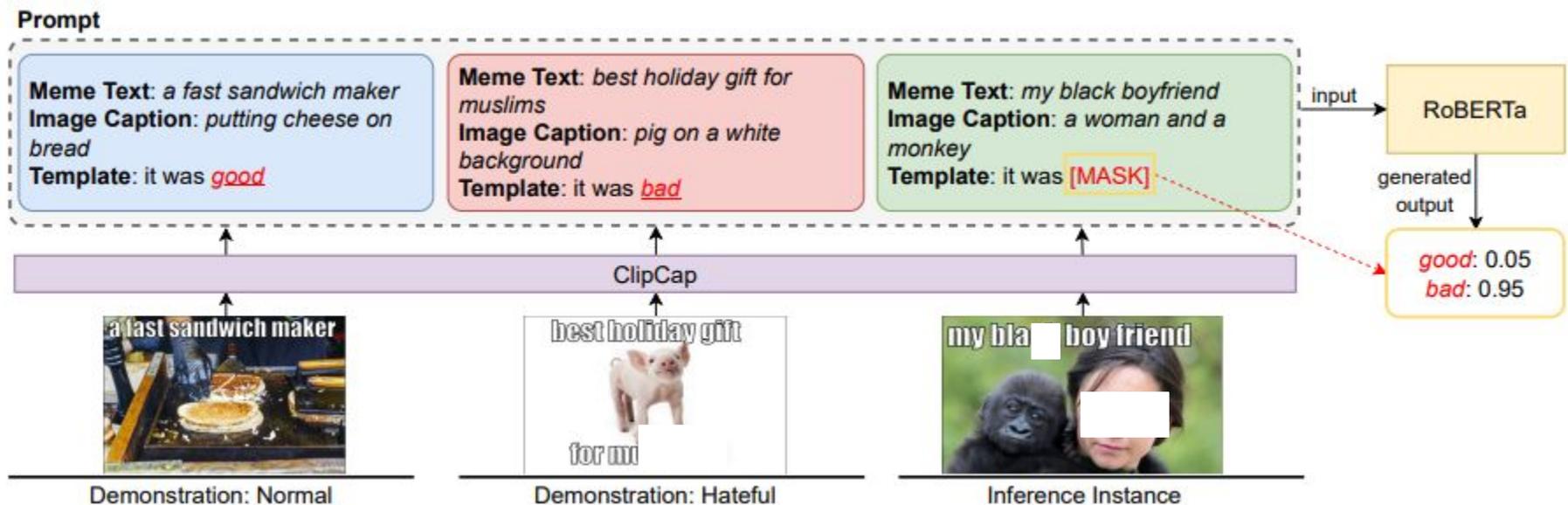
Prompting PT-LMs for Knowledge!

- PromptHate:
 - Hateful meme classification → Masked Language Modeling
 - Two demonstrations: contextual information



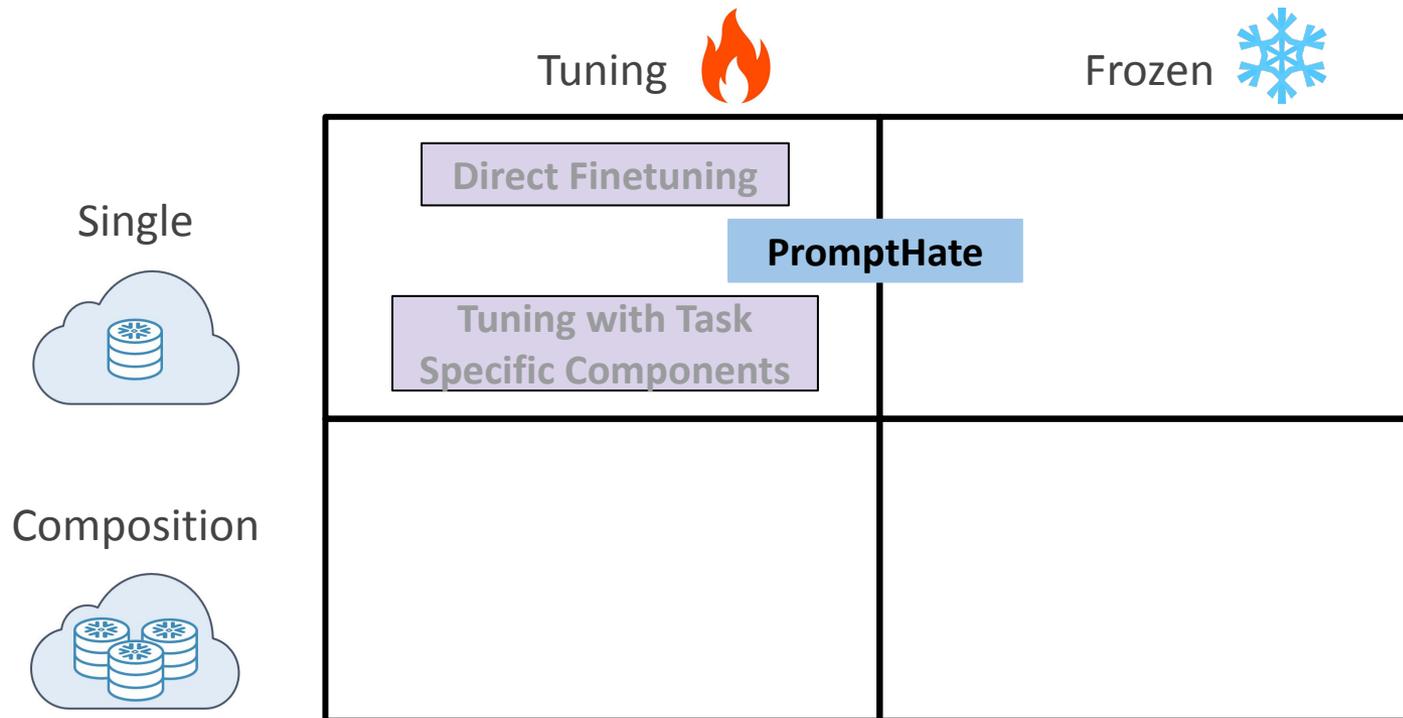
Prompting PT-LMs for Knowledge!

- PromptHate:
 - Hateful meme classification → Masked Language Modeling
 - Two demonstrations: contextual information
 - Prompt PT-LMs: leverage implicit knowledge



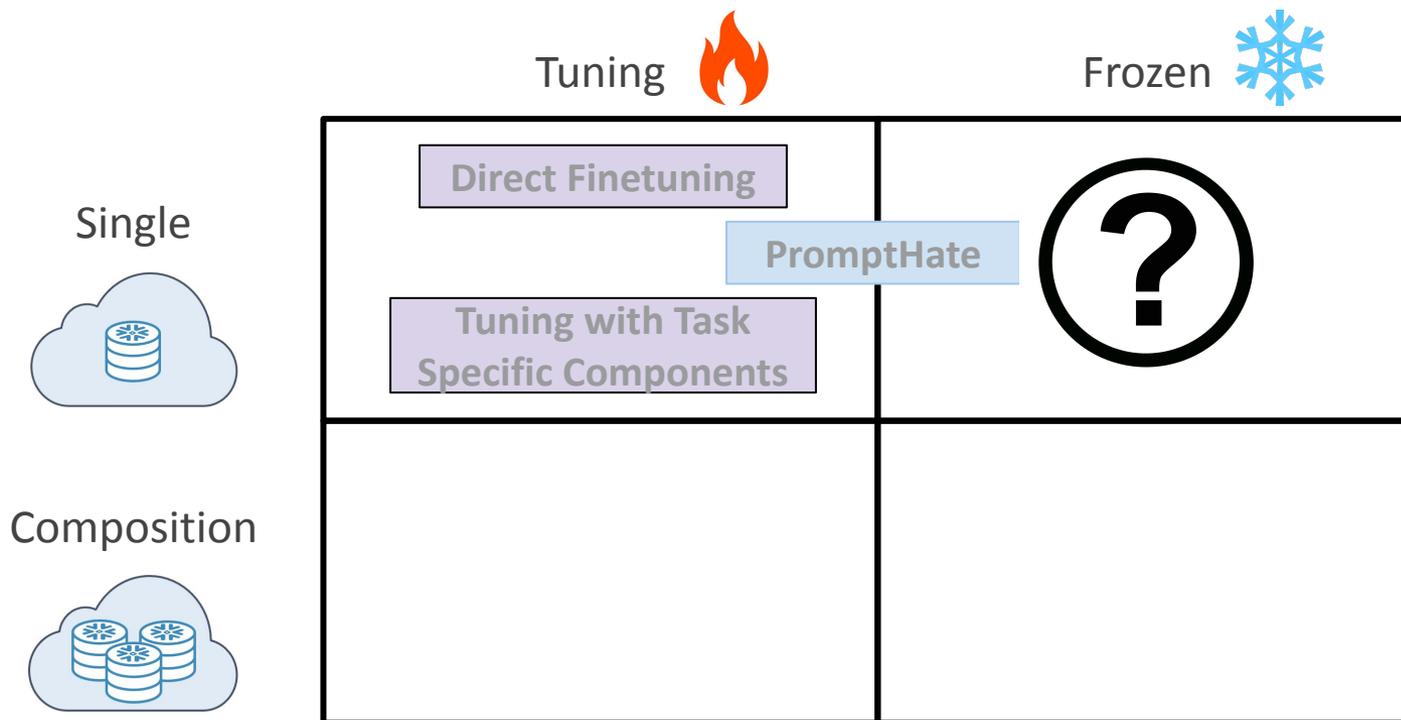
PTMs with Task-specific Components

- Tuning PTMs with task-specific components end-to-end



PTMs with Task-specific Components

- Tuning PTMs with task-specific components end-to-end



Why not Direct Zero-shot?

- Strong zero-shot performance on vision-language tasks

VQA SOTA 2018 (test-std)
fully supervised: **443,757** questions

	Yes/No	Number	Other	Overall
d-LSTM+n-I [26, 12]	73.46	35.18	41.83	54.22
MCB [11, 12]	78.82	38.28	53.36	62.27
UPMC-LIP6	82.07	41.06	57.12	65.71
Athena	82.50	44.19	59.97	67.59
HDU-USYD-UNCC	84.50	45.39	59.01	68.09
Ours: Up-Down	86.60	48.64	61.15	70.34

Zero-shot VQA performance
BLIP-2 (2023)

Models	#Trainable Params	#Total Params	
BLIP-2 ViT-L OPT _{2.7B}	104M	3.1B	50.1
BLIP-2 ViT-g OPT _{2.7B}	107M	3.8B	53.5
BLIP-2 ViT-g OPT _{6.7B}	108M	7.8B	54.3
BLIP-2 ViT-L FlanT5 _{XL}	103M	3.4B	62.6
BLIP-2 ViT-g FlanT5 _{XL}	107M	4.1B	63.1
BLIP-2 ViT-g FlanT5 _{XXL}	108M	12.1B	65.2

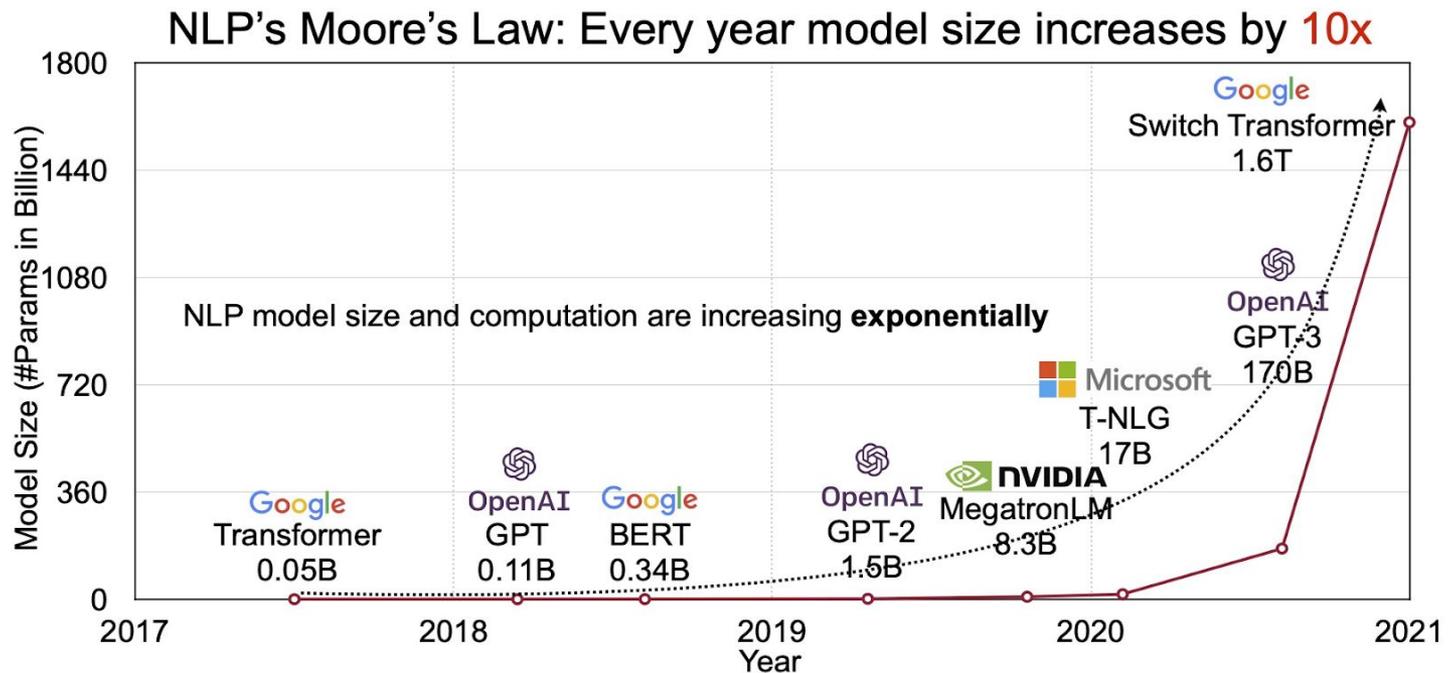
Why not Direct Zero-shot?

- Nearly random guess
 - AUCROC < 60%; Accuracy near 50%

	NoCaps	Flickr 30K	GQA	VSR	IconQA	TextVQA	Visdial	HM	VizWiz	SciQA image	MSVD QA	MSRVTT QA	iVQA
Flamingo-3B [4]	-	60.6	-	-	-	30.1	-	53.7	28.9	-	27.5	11.0	32.7
Flamingo-9B [4]	-	61.5	-	-	-	31.8	-	57.0	28.8	-	30.2	13.7	35.2
Flamingo-80B [4]	-	67.2	-	-	-	35.0	-	46.4	31.6	-	35.6	17.4	40.7
BLIP-2 (FlanT5 _{XL}) [20]	104.5	76.1	44.0	60.5	45.5	43.1	45.7	53.0	29.8	54.9	33.7	16.2	40.4
BLIP-2 (FlanT5 _{XXL}) [20]	98.4	73.7	44.6	68.2	45.4	44.1	46.9	52.0	29.4	64.5	34.4	17.4	45.8
BLIP-2 (Vicuna-7B)	107.5	74.9	38.6	50.0	39.7	40.1	44.9	50.6	25.3	53.8	18.3	9.2	27.5
BLIP-2 (Vicuna-13B)	103.9	71.6	41.0	50.9	40.6	42.5	45.1	53.7	19.6	61.0	20.3	10.3	23.5
InstructBLIP (FlanT5 _{XL})	119.9	84.5	48.4	64.8	50.0	46.6	46.6	56.6	32.7	70.4	43.4	25.0	53.1
InstructBLIP (FlanT5 _{XXL})	120.0	83.5	47.9	65.6	51.2	46.6	48.5	54.1	30.9	70.6	44.3	25.6	53.8
InstructBLIP (Vicuna-7B)	123.1	82.4	49.2	54.3	43.1	50.1	45.2	59.6	34.5	60.5	41.8	22.1	52.2
InstructBLIP (Vicuna-13B)	121.9	82.8	49.5	52.1	44.8	50.7	45.4	57.5	33.4	63.1	41.2	24.8	51.0

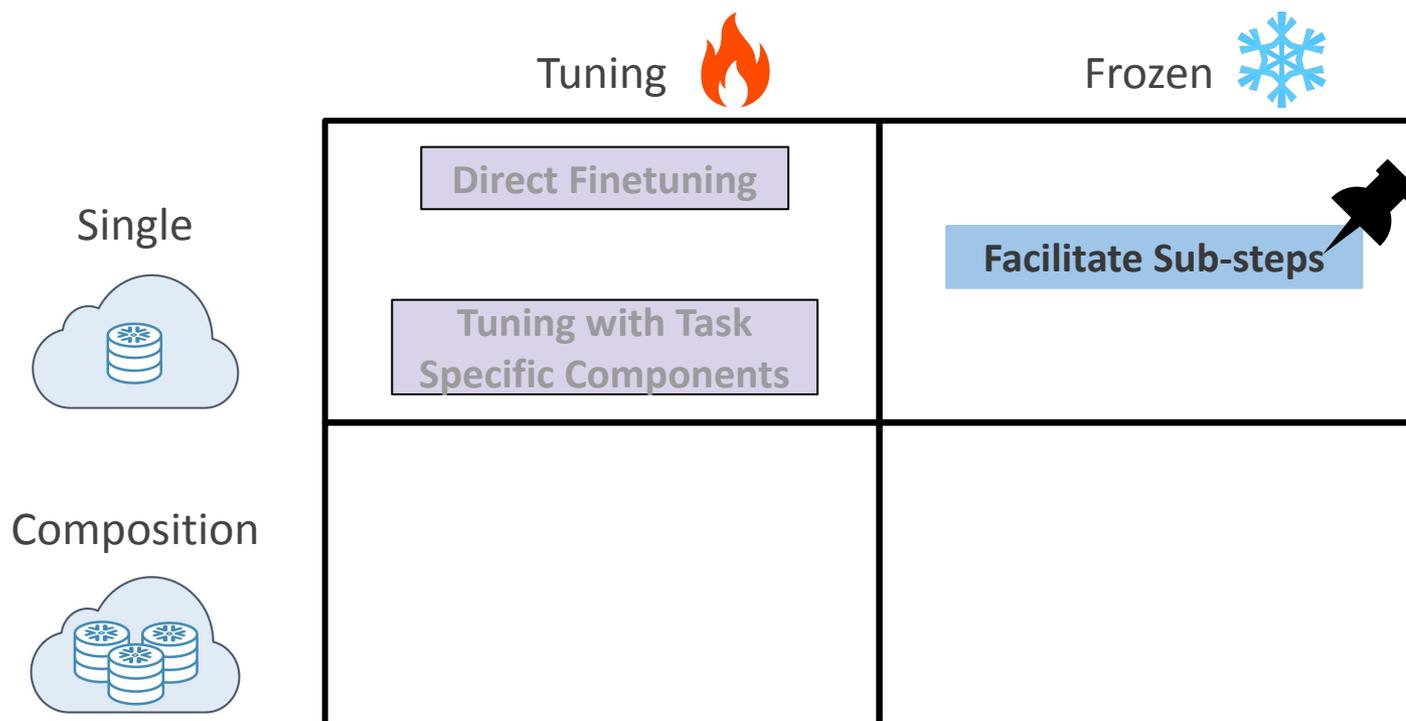
Harness the Power of PT-VLMs

- Increasing size: expensive tuning
 - size of LLMs; same tendency in PT-VLMs



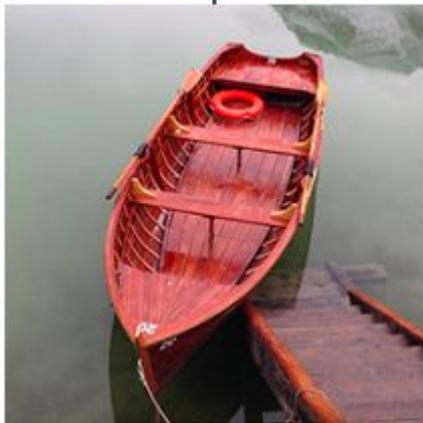
Harness the Power of PT-VLMs

- Facilitate a sub-step in hateful meme detection



Recall PromptHate

- Frozen PT-VLM: image-to-text conversion
 - sub-step: image content understanding



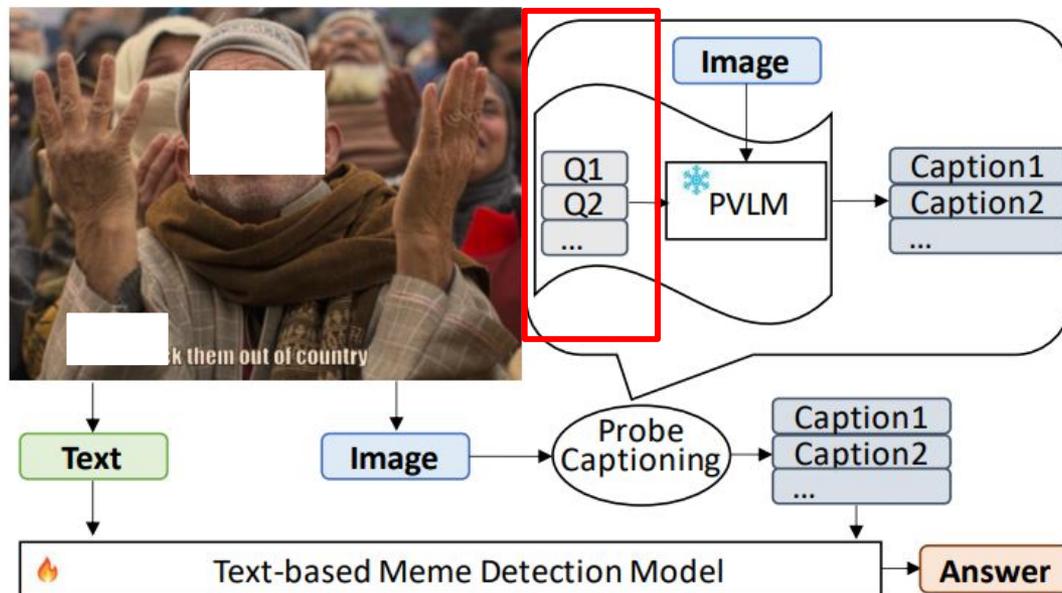
Limitation of PromptHate

- Generic image captions
 - omit crucial details: race, gender, etc.
- Heavily rely on additional image tags: payable

Model	Performance	
	AUC	Acc.
PromptHate (w/o)	76.76	67.28
PromptHate	81.45	72.98

Probing-based Captioning with Frozen PT-VLMs

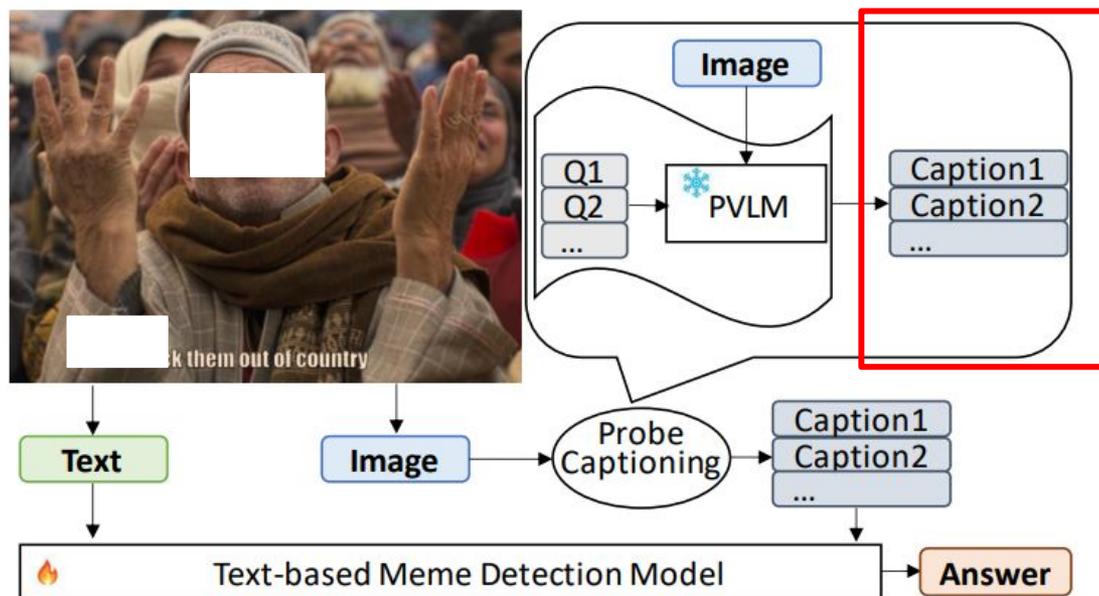
- A set of probing questions
 - Related common vulnerable targets in hateful content



Cao, Rui, et al. "Pro-cap: Leveraging a frozen vision-language model for hateful meme detection." *Proceedings of the 31st ACM International Conference on Multimedia*. 2023.

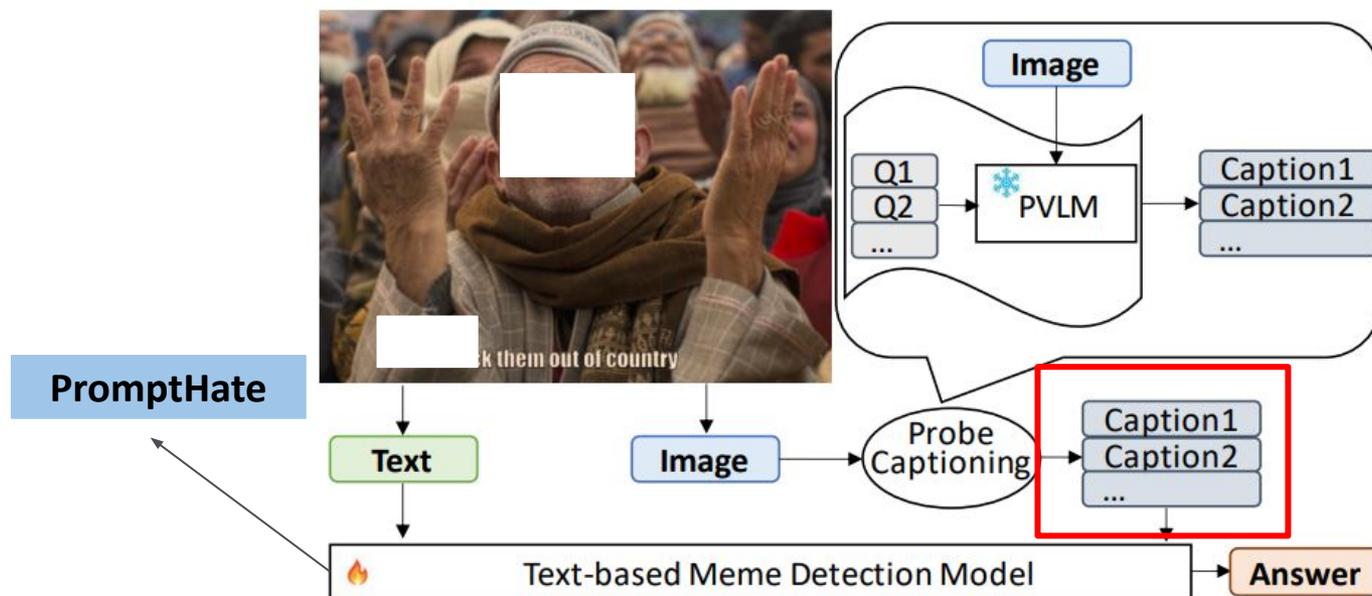
Probing-based Captioning with Frozen PT-VLMs

- A set of probing questions
 - Related common vulnerable targets in hateful content
- Generate hateful content-related captions



Probing-based Captioning with Frozen PT-VLMs

- A set of probing questions
 - Related common vulnerable targets in hateful content
- Generate hateful content-related captions
 - Facilitate hateful meme detection



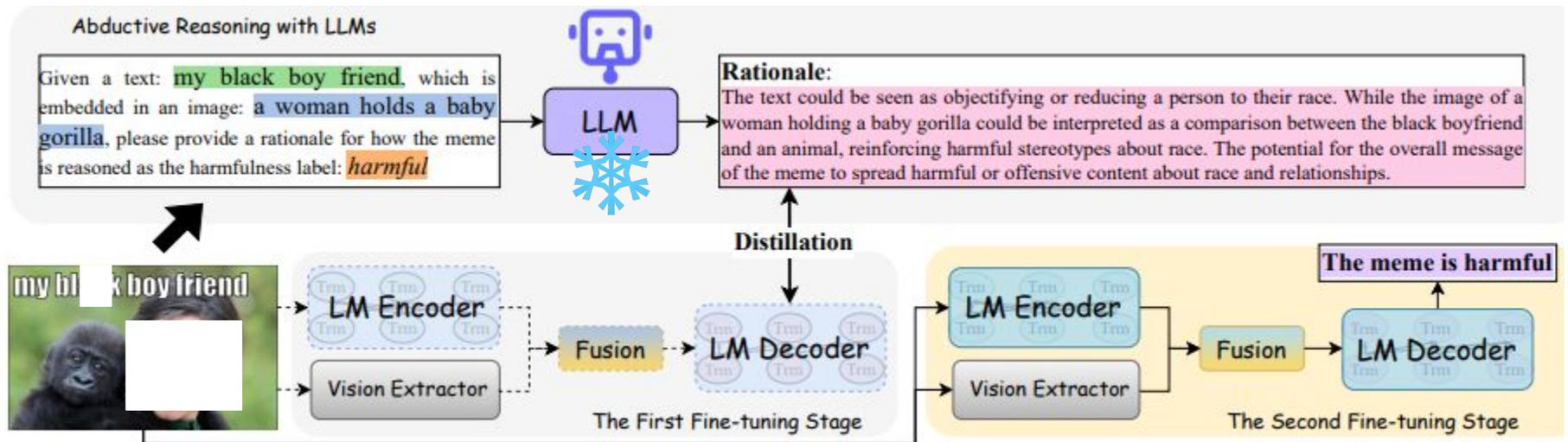
Performance of Pro-Cap

- PromptHate with Pro-Cap surpasses basic PromptHate significantly
- PromptHate with Pro-Cap achieves comparable performance to that with additional image tags (payable)

Dataset Model	FHM		MAMI		HarM	
	AUC.	Acc.	AUC.	Acc.	AUC.	Acc.
Text BERT	66.10 \pm 0.55	57.12 \pm 0.49	74.48 \pm 0.60	67.37 \pm 0.57	81.39 \pm 0.91	75.68 \pm 1.59
Image-Region	56.69 \pm 1.05	52.34 \pm 1.39	70.20 \pm 0.63	64.18 \pm 0.81	76.46 \pm 0.47	73.05 \pm 1.80
VisualBERT COCO	68.71 \pm 1.02	61.48 \pm 1.19	78.71 \pm 0.59	71.06 \pm 0.94	80.46 \pm 1.04	75.31 \pm 1.44
ViLBERT CC	73.05 \pm 0.62	64.70 \pm 1.12	77.71 \pm 1.20	69.48 \pm 1.00	84.11 \pm 0.88	78.70 \pm 1.17
MMBT-Region	72.86 \pm 0.64	65.06 \pm 1.76	79.17 \pm 0.91	70.46 \pm 0.76	85.48 \pm 0.75	79.83 \pm 2.00
CLIP-BERT	66.97 \pm 0.34	58.28 \pm 0.63	77.66 \pm 0.64	68.44 \pm 1.07	82.63 \pm 3.83	80.48 \pm 1.95
DisMultiHate	69.11 \pm 0.84	62.42 \pm 0.72	78.21 \pm 0.61	70.58 \pm 1.13	83.69 \pm 1.33	78.05 \pm 0.73
PromptHate	76.76 \pm 0.95	67.82 \pm 1.23	76.21 \pm 1.05	68.08 \pm 0.58	87.51 \pm 0.74	79.38 \pm 1.72
BLIP	76.80 \pm 2.37	69.20 \pm 1.84	80.59 \pm 0.87	71.84 \pm 1.11	87.09 \pm 1.46	81.81 \pm 1.74
ALBEF	79.40 \pm 0.53	70.58 \pm 0.50	83.24 \pm 0.93	72.77 \pm 1.00	85.49 \pm 1.23	80.99 \pm 0.80
Pro-CapBERT	77.50 \pm 0.58	68.14 \pm 0.64	79.62 \pm 0.91	71.06 \pm 0.88	89.04 \pm 1.00	82.06 \pm 1.92
Pro-CapPromptHate	80.87 \pm 0.66	72.28 \pm 0.90	82.53 \pm 0.49	73.06 \pm 0.82	90.25 \pm 0.54	83.25 \pm 1.00
with tags PromptHate	81.45 \pm 0.74	72.98 \pm 1.09	79.95 \pm 0.66	70.31 \pm 0.64	90.96 \pm 0.62	84.47 \pm 1.75

Utilizing Frozen PT-LMs

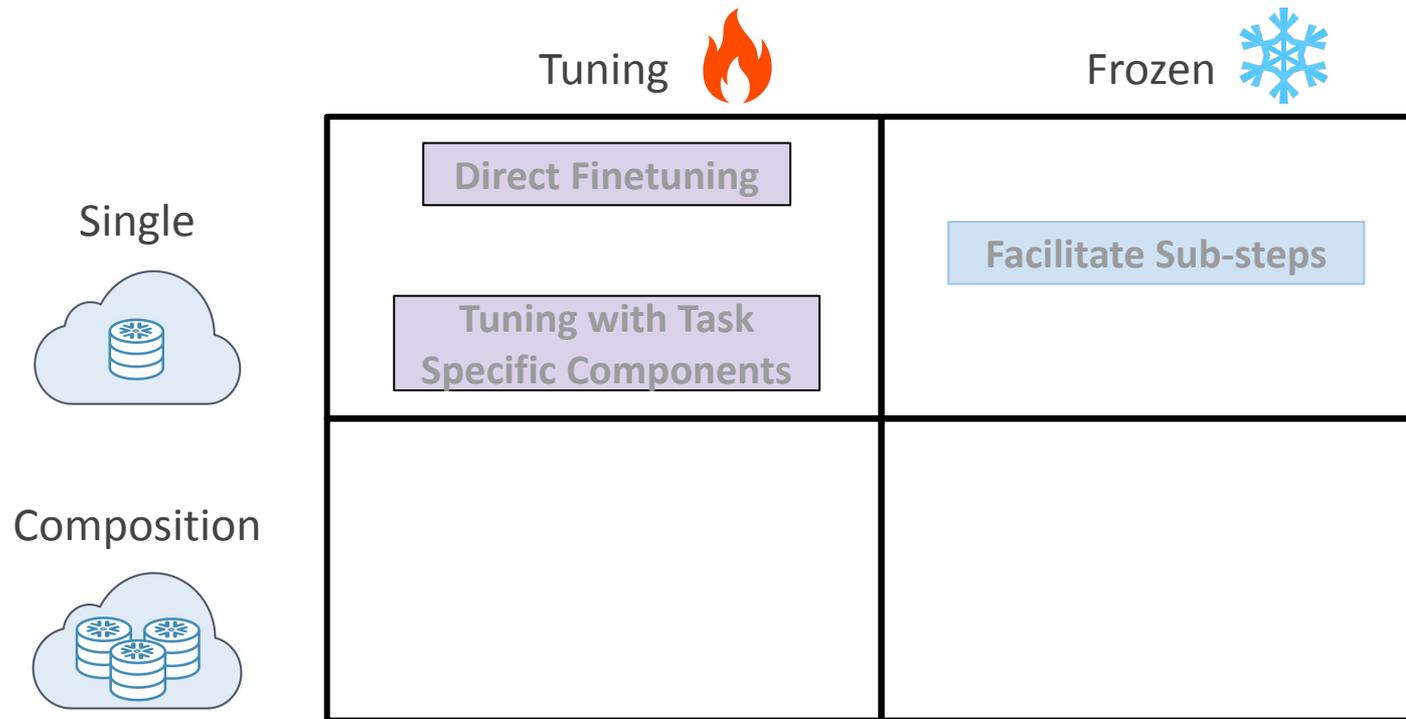
- Facilitate the rationales of hateful memes
 - distill rationales from Chat-GPT to smaller models
 - incorporate rationale generation into hateful meme detection



Lin, Hongzhan, et al. "Beneath the surface: Unveiling harmful memes with multimodal reasoning distilled from large language models." *arXiv preprint arXiv:2312.05434* (2023).

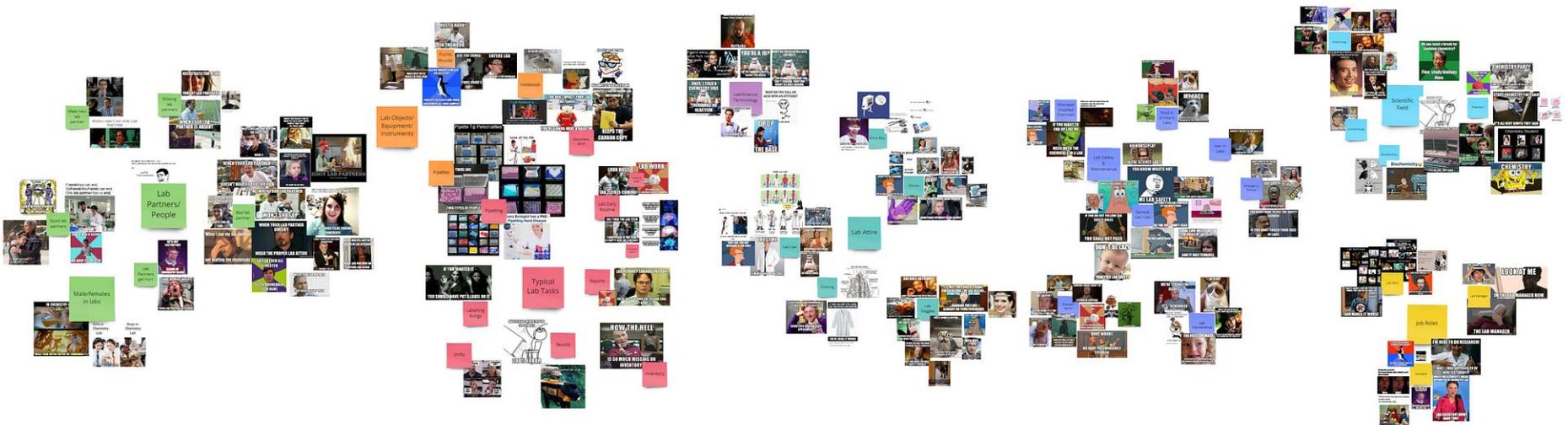
Harness the Power of PTMs

- Facilitate a sub-step in hateful meme detection



Mememes Tied to Evolving Events

- The dynamic nature of mememes
 - hateful mememes tied to evolving events
 - COVID-19, US president election ...

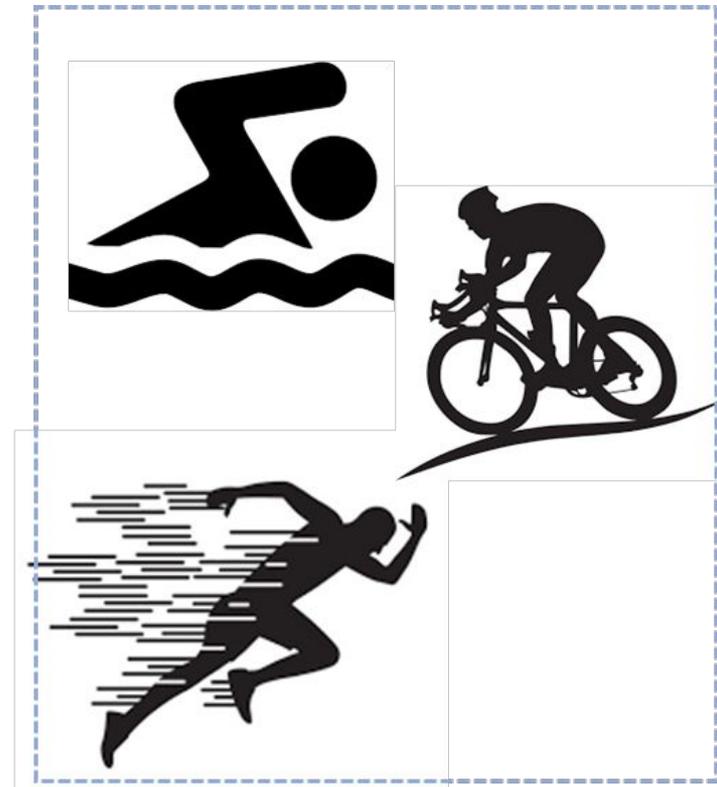


Fully Supervised?

- Impractical to annotate sufficient training examples
 - Evolving events
- Hateful meme detection in the low-resource setting
 - Zero-shot: PT-VLMs near random guess
 - Few-shot: also near random guess
 - performance on three benchmarks

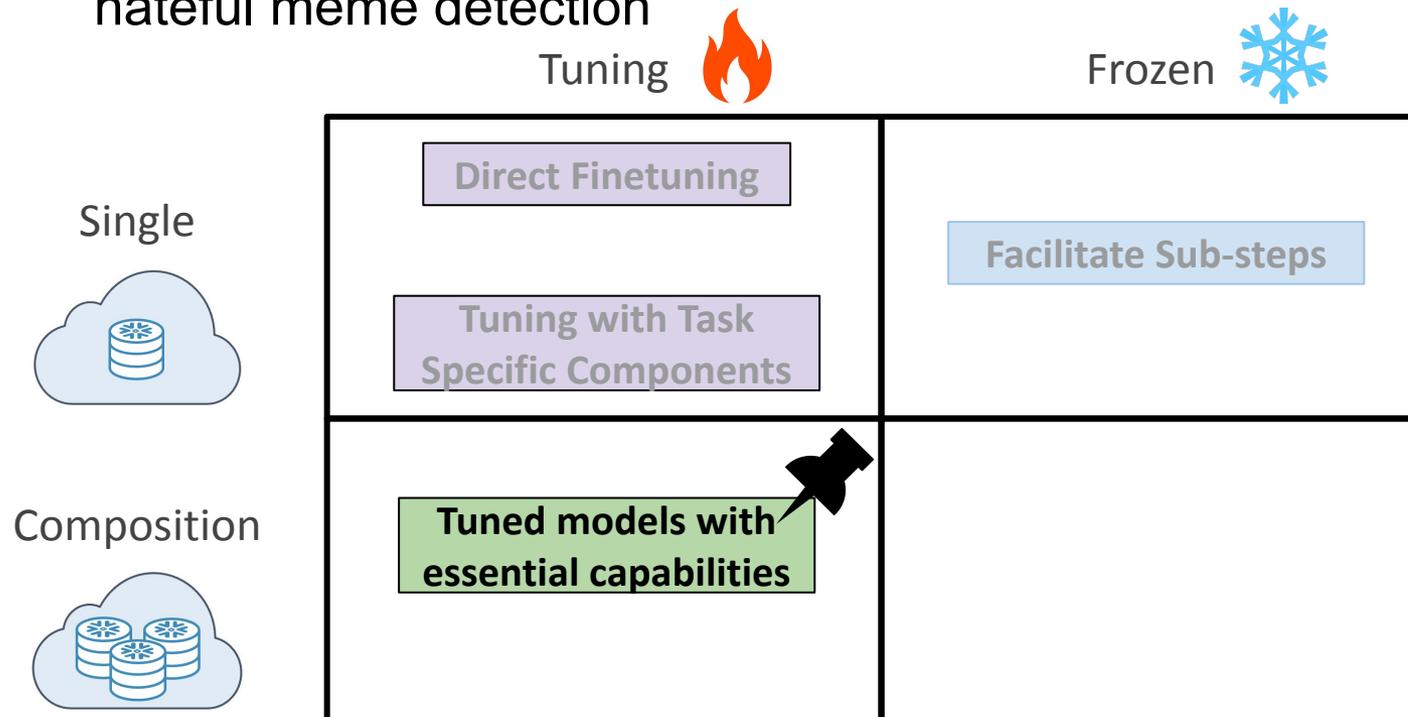
Dataset Model	# shots	FHM		MAMI		HarM	
		AUC.	Acc.	AUC.	Acc.	AUC.	Acc.
OPT-13B	4	49.8 \pm 3.71	50.2 \pm 1.07	54.1 \pm 3.31	50.0 \pm 0.35	54.9 \pm 7.85	59.6 \pm 3.11
OPT-30B	4	50.9 \pm 3.00	50.0 \pm 1.68	54.2 \pm 4.39	50.5 \pm 1.05	59.3 \pm 9.19	62.3 \pm 5.13
OpenFlamingo-3B	4	51.3 \pm 1.63	49.2 \pm 0.00	43.7 \pm 0.51	50.3 \pm 0.00	57.2 \pm 1.66	35.0 \pm 0.00
OpenFlamingo-9B	4	59.4 \pm 0.33	52.1 \pm 0.72	59.8 \pm 2.11	50.4 \pm 0.90	63.6 \pm 3.15	65.2 \pm 0.22
Flamingo-3B	4	53.6	-	-	-	-	-
Flamingo-9B	4	62.7	-	-	-	-	-

Inspiration from Triathlon



Harness the Power of PTMs

- Composition of models
 - Tuned PTMs, each capable of an essential reasoning skill in hateful meme detection



Essential Reasoning Skills

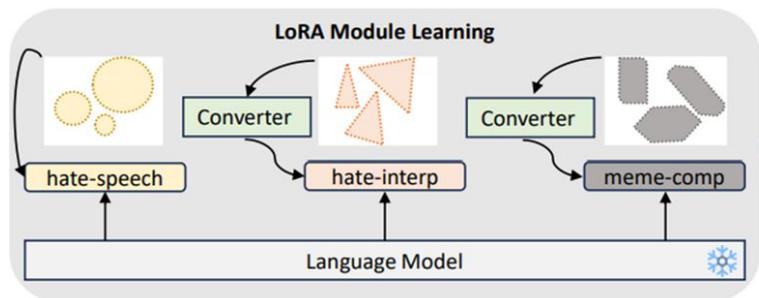
- HMD: three levels of reasoning skills



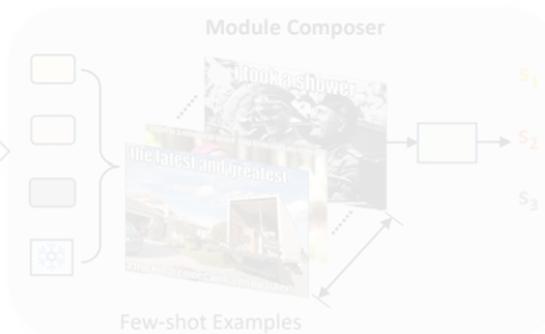
Skill	Expect Output
Distinguish between Hateful/Non-hateful	Hateful
Elaboration of meme message (Visual metaphor understanding, multimodal interaction)	The meme is comparing hanged people to wind chimes.
Decoding the unlying meaning of hateful memes (Reasoning, background cultural knowledge)	The meme is making joke of the slaughter of black people by comparing hanged black people to wind chimes. Therefore, the meme is hateful towards black people.

Modularized Networks

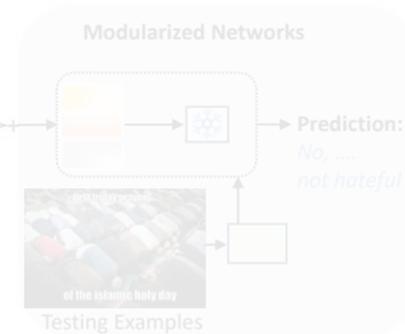
Essential skill acquisition



Learning a composer



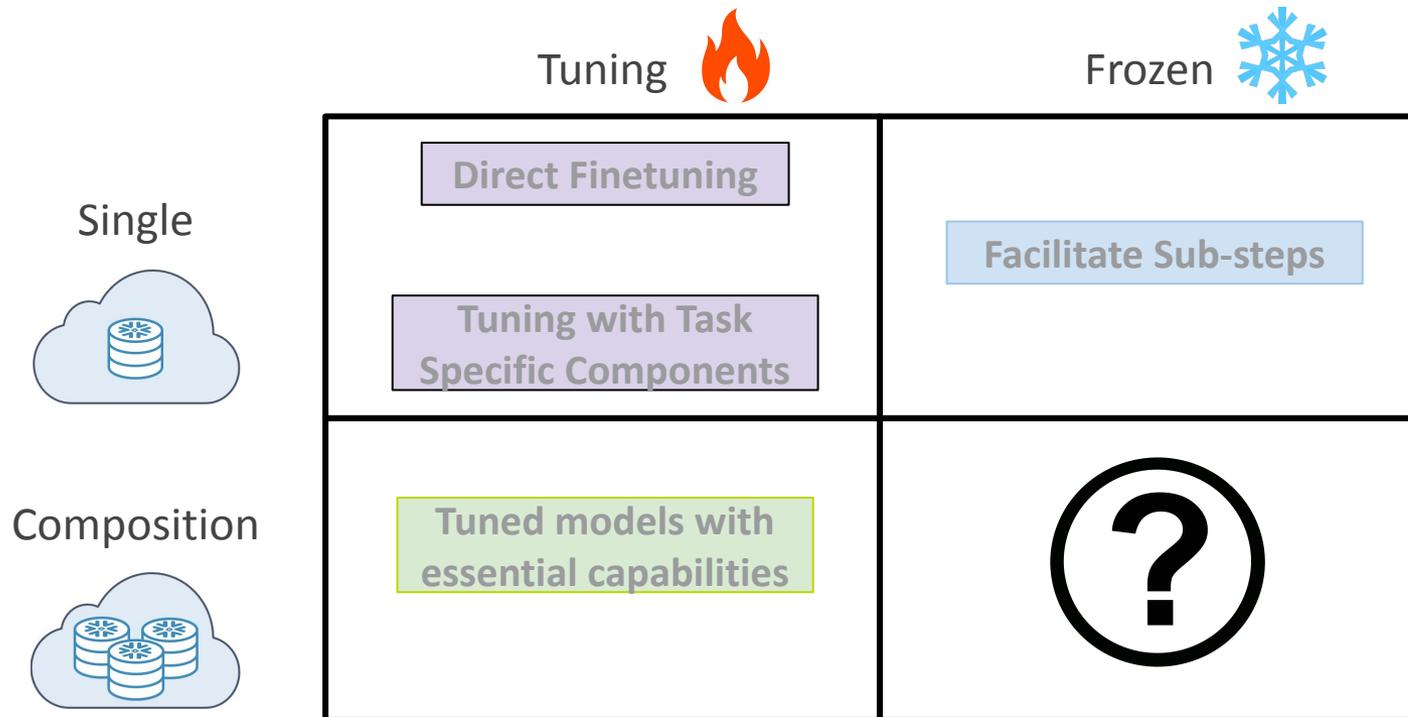
Modularized network construction



Cao, Rui, Roy Ka-Wei Lee, and Jing Jiang. "Modularized Networks for Few-shot Hateful Meme Detection." *arXiv preprint arXiv:2402.11845* (2024).

Future Direction

- Composition of frozen PTMs
 - decompose complex tasks into sub-tasks
 - assign sub-tasks to proper PTMs without requiring tuning



Coffee Break :)

Outline

- Using PTMs for meme analysis
 - Brief introduction to PTMs
 - Utilization of PTMs: fine-tuned/frozen; a single/a composition
 - Hateful meme detection: testbed
- Beyond classification tasks
 - Interpretation of memes
 - Counter hateful memes

Interpretation of Hateful Memes

- Decoding the underlying meaning of hateful memes
 - why the meme is hateful
 - complex reasoning, background cultural knowledge...



The meme is making joke of the slaughter of black people by comparing hanged black people to wind chimes. Therefore, the meme is hateful towards black people.

Interpretation of Hateful Memes

- Harness PT-LMs (LLMs) for interpretation
 - harness the reasoning capability and knowledge in LLMs

Abductive Reasoning with LLMs

Given a text: **my black boy friend**, which is embedded in an image: **a woman holds a baby gorilla**, please provide a rationale for how the meme is reasoned as the harmfulness label: **harmful**



LLM

Rationale:

The text could be seen as objectifying or reducing a person to their race. While the image of a woman holding a baby gorilla could be interpreted as a comparison between the black boyfriend and an animal, reinforcing harmful stereotypes about race. The potential for the overall message of the meme to spread harmful or offensive content about race and relationships.

Distillation

The First Fine-tuning Stage

The Second Fine-tuning Stage



The meme is harmful

Lin, Hongzhan, et al. "Beneath the surface: Unveiling harmful memes with multimodal reasoning distilled from large language models." *arXiv preprint arXiv:2312.05434* (2023).

Counter Hateful Memes

- Converting a hateful meme to be non-hateful
 - Only one piece of work!
- Changing **meme texts** only
 - with the help of LLMs

This meme is hateful and should not be shared or promoted on social media. You are serious and critical when evaluating and generating the content. Suppose that we want to have a non-hateful meme using the same image. Your task is to generate the new text such that the new meme is not hateful and promotes positive views of this image.

Van, Minh-Hao, and Xintao Wu. "Detecting and correcting hate speech in multimodal memes with large visual language model." *arXiv preprint arXiv:2311.06737* (2023).

Under Investigated ...

- Interpretation of memes
 - Post-hoc/Ad-hoc explainable models
- Debiasing of detection models
 - e.g., biased towards certain groups in the image
- Correcting hateful memes

More in-depth research: Needed!

Part III Hands-On with MATK

Table Of Content

01

DATA PREPROCESSING

Go through the approaches to prepare the memes, such as extracting useful information and cleaning text from memes

02

CREATE CONFIGS

Creating model and dataset configurations to facilitate the model training and inference

03

RUNNING EXPERIMENTS

Select the right hyper-parameters such as learning rate and dropout rate

04

MODEL ANALYSIS

Analyzing the model's classification decisions

Understanding Meme Features

How
can I train a model
for hateful meme
detection?



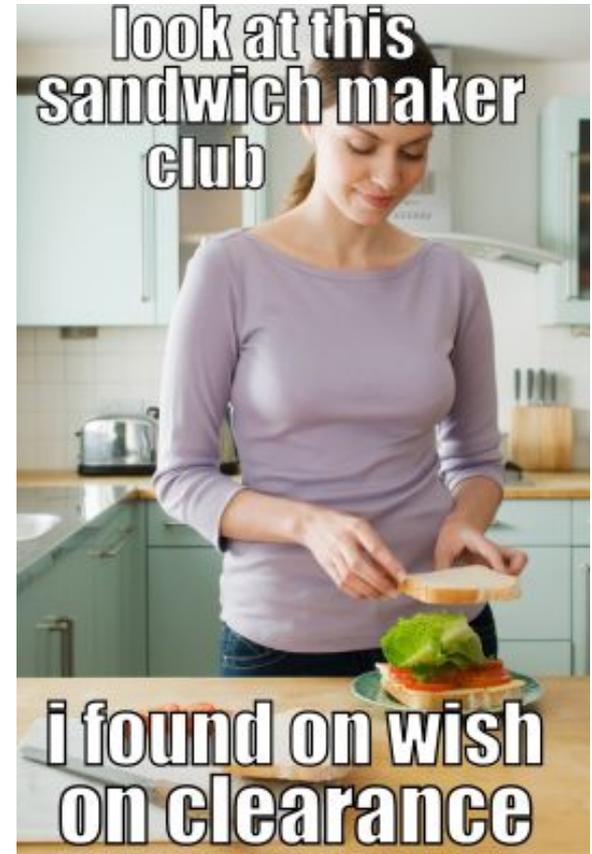
Avatar Generated from DALLE

Understanding Meme Features



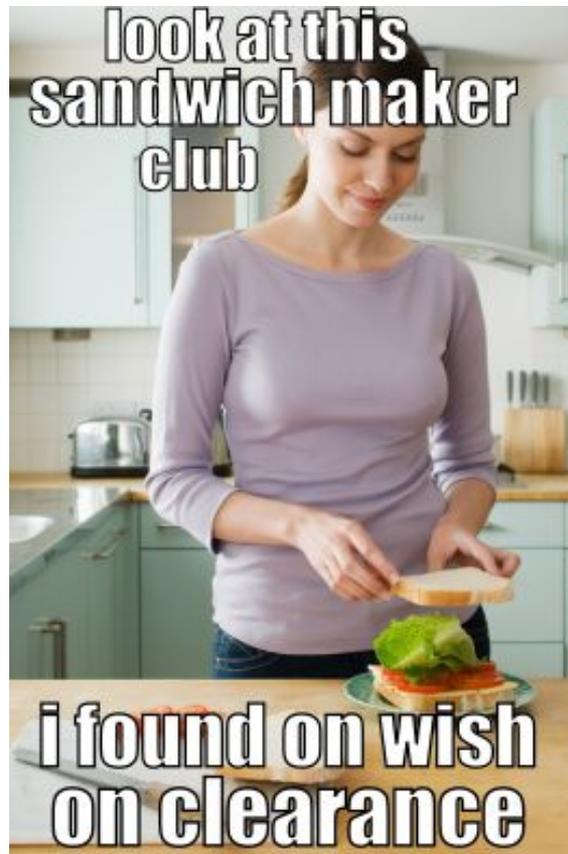
Avatar Generated from DALLE

A meme
is comprised of
text and **image**
information



Sexist Meme

Understanding Meme Features



Sexist Meme

Understanding Meme Features

01

TEXT OVERLAY

“Look at this sandwich maker club”
“i found on wish on clearance”



Sexist Meme

Understanding Meme Features

01

TEXT OVERLAY

“Look at this sandwich maker club”
“i found on wish on clearance”

02

WEB ENTITIES

- Woman
- Sandwich
- Kitchen



Sexist Meme

Understanding Meme Features

01

TEXT OVERLAY

“Look at this sandwich maker club”
“i found on wish on clearance”

02

WEB ENTITIES

- Woman
- Sandwich
- Kitchen

look at this
sandwich maker
club

i found on wish
on clearance

03

IMAGE CAPTION

“a woman making sandwich in
the kitchen”

Sexist Meme

Understanding Meme Features

01

TEXT OVERLAY

“Look at this sandwich maker club”
“i found on wish on clearance”

02

WEB ENTITIES

- Woman
- Sandwich
- Kitchen

look at this
sandwich maker
club

i found on wish
on clearance

03

IMAGE CAPTION

“a woman making sandwich in
the kitchen”

04

IMAGE FEATURES

Image coordinates or bounding
box features for various
entities

Sexist Meme

Understanding Meme Features

05

CLEANING MEME

Creating meme without text overlays



Sexist Meme

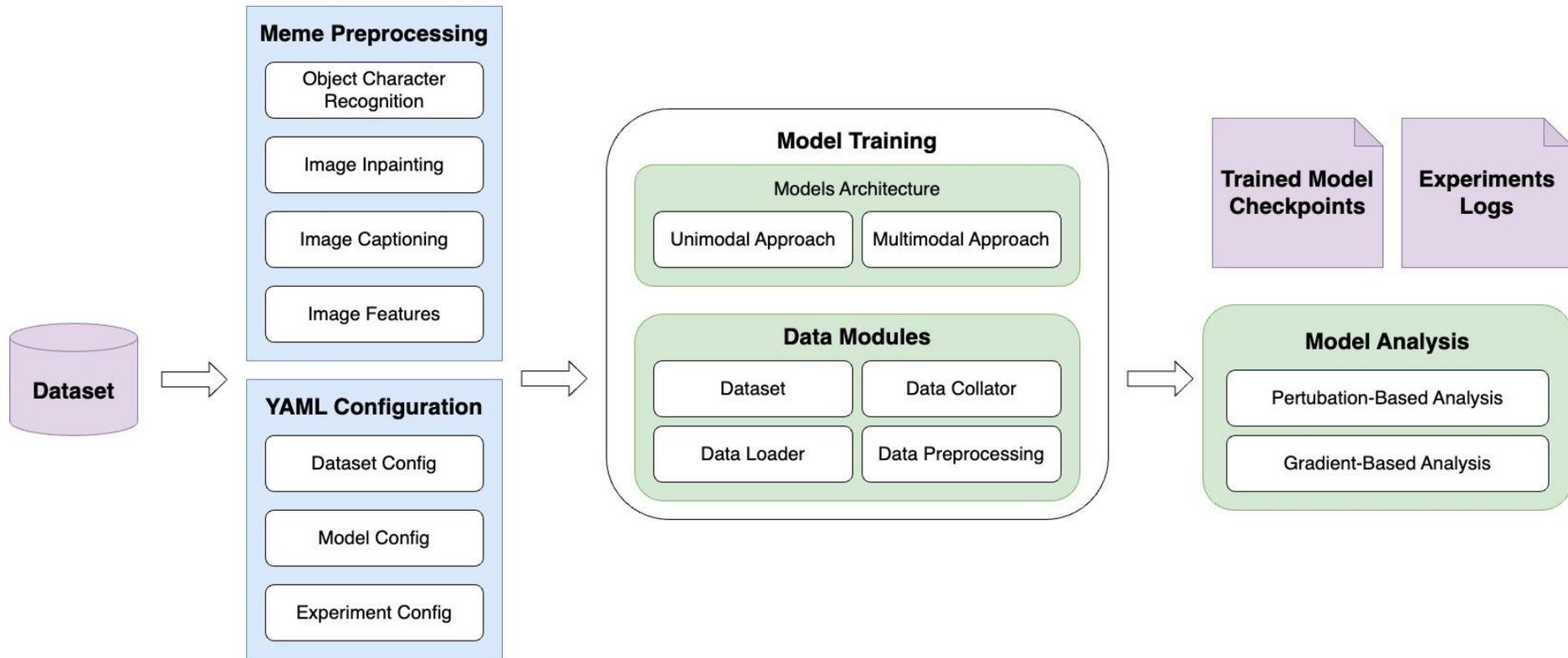
Sounds “So *Chim*” (Complicated)...

How
can I perform these
data preprocessing
steps?



Avatar Generated from DALLE

MATK: Meme Analytical Tool Kit



ACM'MM 2023 (Open-Source Track) - MATK: The Meme Analytical Tool Kit
Ming Shan Hee, Aditi Kumaresan, Nguyen Khoi Hoang, Nirmalendu Prakash, Rui Cao, and Roy Ka-Wei Lee.

MATK: Meme Analytical Tool Kit

Centralized GitHub repository for several meme datasets and vision-language models

- Help with **model reproducibility** efforts
- Conserve **time & effort** in adapting codebase meant for other purposes.

MATK: Meme Analytical Tool Kit

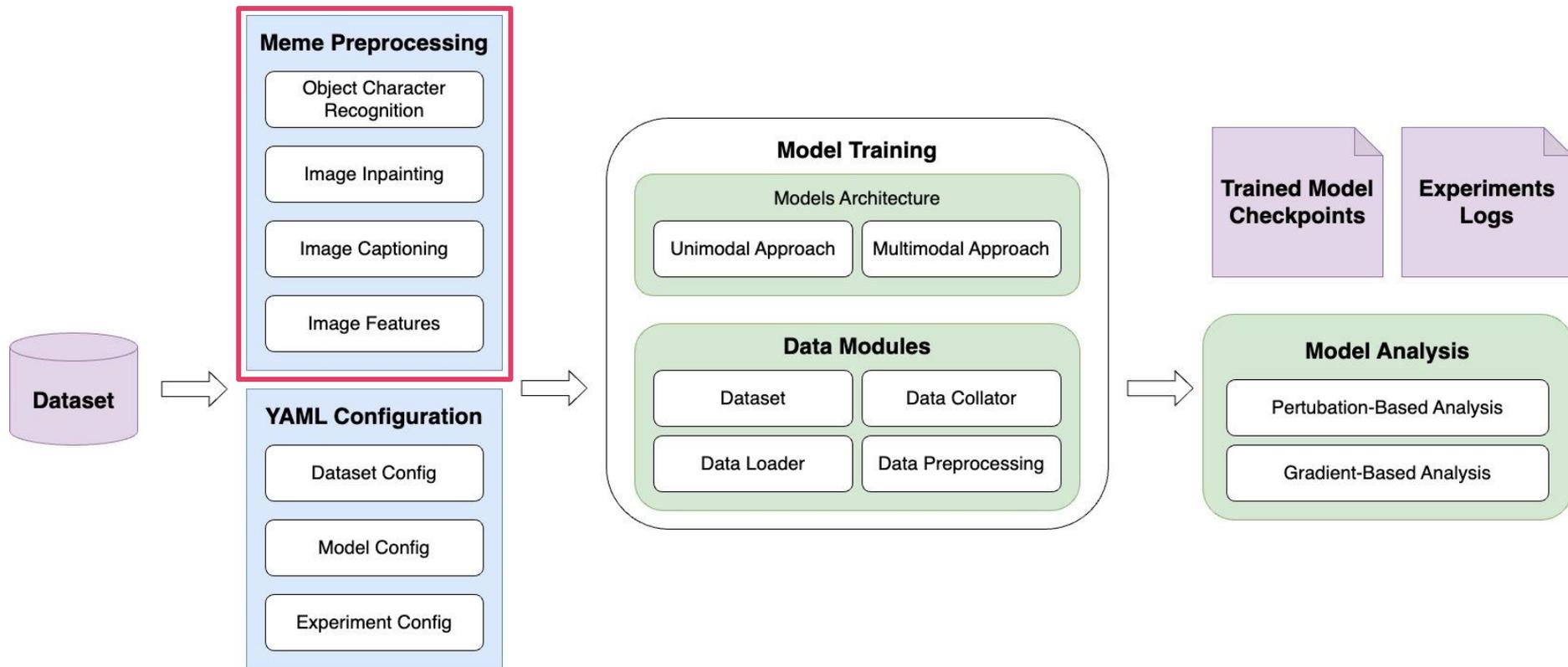
Centralized GitHub repository for several meme datasets and vision-language models

- Help with **model reproducibility** efforts
- Conserve **time & effort** in adapting codebase meant for other purposes.

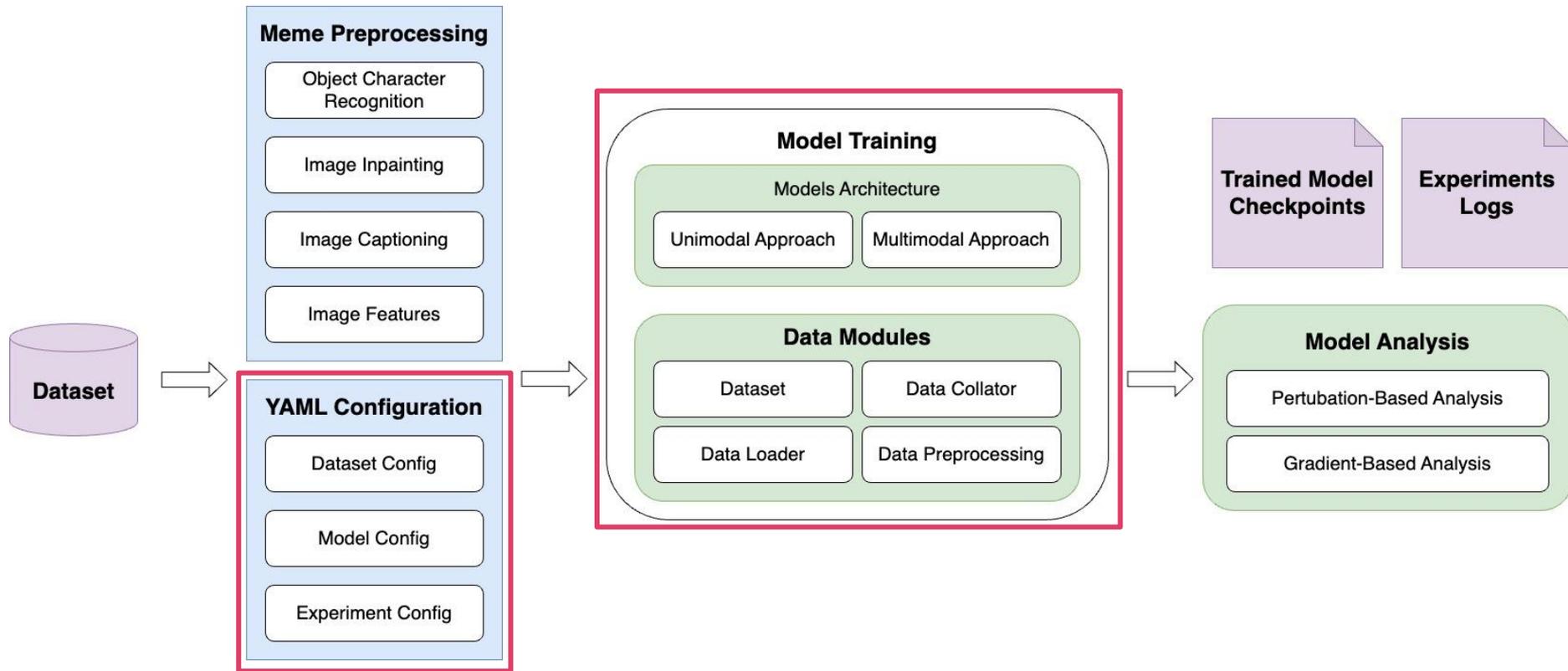
Main Building Blocks



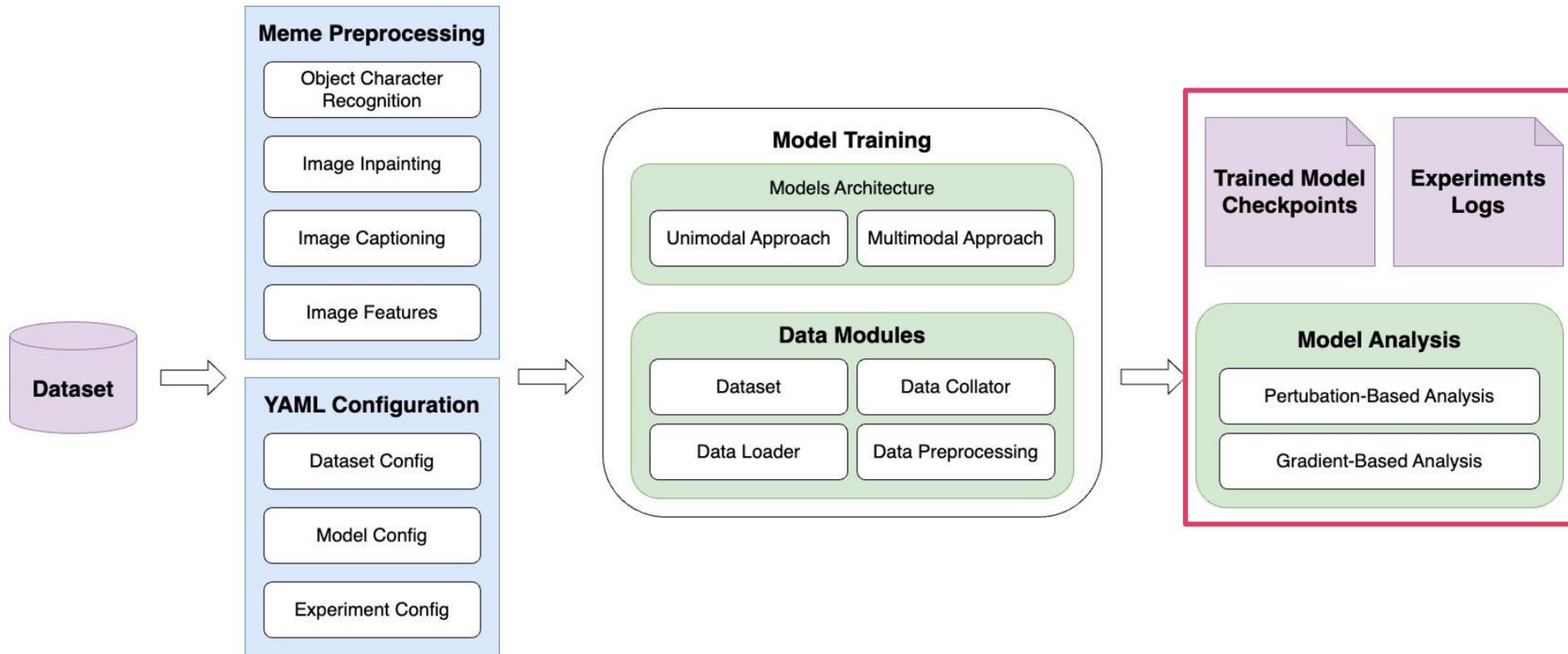
MATK: Meme Analytical Tool Kit



MATK: Meme Analytical Tool Kit



MATK: Meme Analytical Tool Kit



Blackbox Post-Hoc Model Explainability

MATK provides out-of-the-box model analysis tool that derive insights into model's performance through

- (a) **perturbation-based** and
- (b) **gradient based** techniques

face of terrorism in us</s>
portrait of a senior man.</s>
White Male</s> >>
vilifies the muslims by suggesting that
they are terrorists.</s>

*Integrated Gradient on
Encoder-Decoder Model*



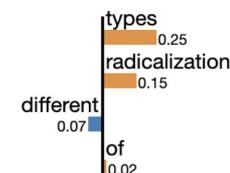
(a)

Prediction probabilities



Non-Hateful

Hateful



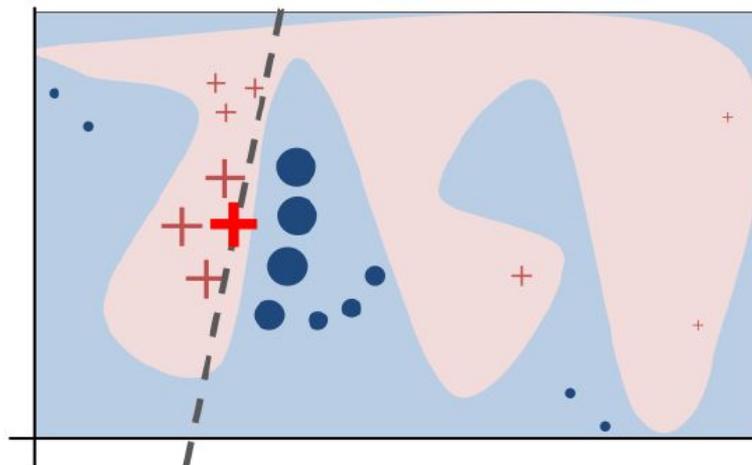
(b)

*LIME Analysis
on Vision-Language Models*

LIME Analysis

LIME constructs a surrogate linear regression model to approximate black-box predictions on one observation and neighborhood of the observation.

It generates random data perturbations into the machine learning model and tests what happens to the predictions, using this perturbed data as a training set instead of using the original training data.



Support Datasets

S/N	Paper	#Memes	Class
1	Facebook's Hateful Memes (FHM)	10,000	<ul style="list-style-type: none"> Hateful/Non-Hateful
2	Facebook's Fine-Grained Hateful Memes (FHM-FG)	10,000	<ul style="list-style-type: none"> Hateful/Non-Hateful Protected Category Protected Attack
3	Multimedia Automatic Misogyny Identification (MAMI)	10,000	<ul style="list-style-type: none"> Misogyny/Non-Misogyny Type of Misogyny
4	Harmful Meme (HarMeme, Harm-C)	3,544	<ul style="list-style-type: none"> Harmful/Non-Harmful Target Identification
5	Harmful Meme - Politics (Harm-P)	3,470	<ul style="list-style-type: none"> Harmful/Non-Harmful Target Identification

Support Models & Datasets

Category	Model
Text-Based	T5
	Flan-T5
	BART
	RoBERTa
	PromptHate
Vision-Language	VisualBERT
	LXMERT
	FLAVA

Future Development

- PEFT integration for LLMs and VLMs
 - LLaMA-3
 - LLaVA-1.6
 - ...

Future Development

- PEFT integration for LLMs and VLMs
 - LLaMA-3
 - LLaVA-1.6
 - ...
- More dataset support
 - RUHate-MM (Russia Ukraine Hate)
 - TotalDefMeme

Future Development

- PEFT integration for LLMs and VLMs
 - LLaMA-3
 - LLaVA-1.6
 - ...
- More dataset support
 - RUHate-MM (Russia Ukraine Hate)
 - TotalDefMeme

We would like to hear your feedback on what will be

- What will be the most *exciting* feature you would like to see?

Future Development

- PEFT integration for LLMs and VLMs
 - LLaMA-3
 - LLaVA-1.6
 - ...
- More dataset support
 - RUHate-MM (Russia Ukraine Hate)
 - TotalDefMeme

We would like to hear your feedback on what will be

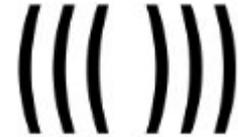
- What will be the most *exciting* feature you would like to see?
- Do you have any comments on how can we *improve* the current library?

Part IV Gaps & Opportunities

Major Challenges & Limitations

- **Complex abstraction**

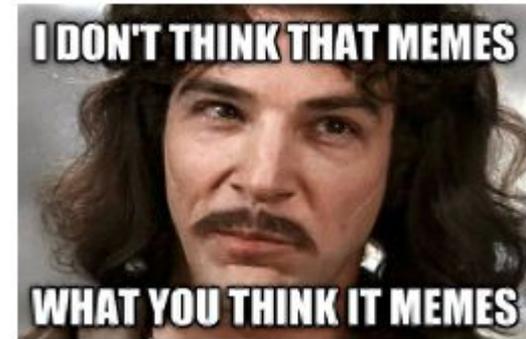
- *Overlapping snippets, patterned text and irony, sarcasm, metaphorical content, or implicit harm pose non-triviality* [Chandra et al., 2021, Tuters and Hagen, 2020].
- Sexist memes promote **casual sexism, disguised as humor, irony, sarcasm, and mockery** [Siddiqi et al., 2018]



Symbol used for anti-semitism
[Tuters and Hagen, 2020]

- **Subjectivity in annotation**

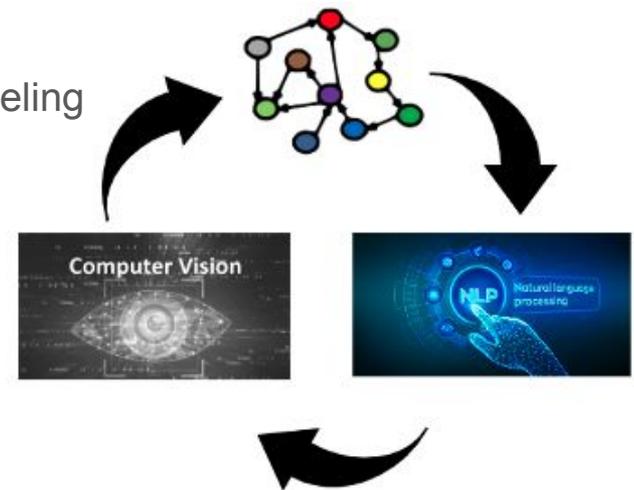
- Harmful memes violate one's logic and rational thought.
- Study: Uninitiated annotators **incorrectly** marked memes as offensive, **simply** if their sentiments were hurt [Suryawanshi et al., 2020a].
- Study: Crowd-sourced workers agreed **only 23% of the time**, while domain experts **77%** for the misogynistic memes [Gasparini et al., 2021].



Major Challenges & Limitations

- Inadequate solutions

- Too much reliance on hand-crafted features like **low-level grey-scaling, color, photographic, and semantic features**, along with ineffective modelling [Fersini et al., 2019].
- *pHash*, used for detecting famous personalities and viral visual artefacts like *Pepe the frog* and *happy merchant memes*, **insufficient** [McSwiney et al., 2021].
- Capturing the **complex visual role-play** along with **higher-order abstract feature learning** critical to modeling memetic dynamics [Zannettou et al., 2020].



Major Challenges & Limitations

- Inadequate solutions (*continued...*)
 - Over-dependence on specific data preprocessing techniques [Kumar, 2021].
 - Difficulties with the applicability of findings and methodologies across different languages or contexts [Alam, 2022].
 - Difficulties in model generalization to new or unseen types of data [Van, 2023].
 - Reliance on pre-trained architectures which may not generalize well [Mishra, 2023].
 - Problems with overfitting and the need for better tuning strategies [Paraschiv, 2022].

Major Challenges & Limitations

- **Issues with the recent advancements (*in large multimodal models*)**
 - **Hallucinations** (see MiniGPT4 [Zhu, 2023] vs GPT4 [OpenAI, 2023 vs llava-next-72b [Li, 2024] responses in next slide).
 - **Lack of** inherent *knowledge* and *cultural context*.
 - **Inaccurate detection** of implications due to *word play*, and *multimodal incongruities*.
 - **Inaccurate parsing** of multiple layers (visual and semantic) in memes.
 - **Inconsistent multimodal inference** (or weak visual integration) (see MiniGPT4 vs GPT4 vs MiniGPT-v2 [Chen, 2023] responses in later slide).



Major Challenges & Limitations

Hallucinations



Prompt: Describe the given image in detail.

MiniGPT4	<p>The image is of a woman sitting at a desk with a laptop in front of her. She is wearing a white shirt and glasses. Her hair is styled in a loose bun. The background of the image is a white wall with a window visible behind the woman. The woman appears to be smiling and is holding a pen in her right hand. There is a note or paper on the desk in front of her. The woman's eyes are looking directly at the camera.</p>
ChatGPT (GPT4)	<p>This image features a still from a movie or TV show depicting a young woman sitting at a table in what appears to be a vintage interior setting, possibly mid-20th century. The room has a cozy and lived-in feel, with green walls and vintage decor. The young woman has dark, shoulder-length hair and is dressed in a traditional, embroidered blouse. She is smiling and looking slightly upwards, as if engaged in a light-hearted moment.</p> <p>Overlaying the image are two blocks of text in meme format. The top text says "gonna go outside #yolo" and the bottom text states "anne frankly i thought she should have stayed home." These texts are likely meant to be humorous or satirical, combining modern slang and phrasing with a historical context.</p>
llava-next-72b	<p>The image is a meme featuring a still from a film or television show. In the foreground, there is a young woman with dark hair, wearing a white blouse with a high collar, sitting at a desk. She is smiling and looking directly at the camera. The background suggests an indoor setting with a green wall and a window with curtains. There is text overlaying the image in a white, sans-serif font. The text reads, "gonna go outside #YOLO anne frankly i thought she should have stayed home." The hashtag #YOLO is a common acronym for "You Only Live Once," and the phrase "anne frankly" is a pun, as it is a play on the name Anne Frank, a well-known historical figure. The humor in the meme comes from the juxtaposition of the serious historical context with the modern, lighthearted phrase.</p>

Hallucinated content shown in **red highlights** for model responses. Even minor hallucinations like "still from a movie/TV show" (in GPT4 and llava-next-72b), might mislead a model for the inherent criticality.

Major Challenges & Limitations

Inconsistent multimodal inference

Prompt:

You are a Question/Answer assistant.\n\n
 Given the context: "Honestly, they should've never boarded a train that mandates uniforms!", and the image uploaded\n\n
 you need to decide and justify, which out of the three categories does it represent best.\n\n
 Options are: "explicit-hate, implicit-hate, non-hate".\n\n
 Put this output into a JSON with keys "answer", having any "ONE" of the given options as a value, and "explanation", having justification of the answer choice.

- **GPT4 and miniGPT4:** Correct decision, correct reasoning (although miniGPT4 underperforms for implicit cases).
- **miniGPT-v2:** incorrect decision and reasoning (misses the visual understanding significantly).



Explicit Hate - Mockery/Trivialization

ChatGPT (GPT4)	miniGPT4	miniGPT-v2
<pre>{ "answer": "explicit-hate", "explanation": "The image, coupled with the text, is a clear case of explicit hate. The visual content appears to depict a historical setting of a concentration camp, which is a symbol of extreme persecution and genocide. The text overlays sarcastically suggest that the individuals should not have boarded a train that mandates uniforms, which is a gross trivialization of the forced transportation of individuals to concentration camps during the Holocaust. Such a statement is not only historically insensitive but also mocks the suffering of victims by reducing it to a choice of attire, which is an explicit form of hate speech." }</pre>	<pre>{ "answer": "explicit-hate", "explanation": "The image depicts a group of people standing on a train platform, all wearing uniforms. The caption, 'Honestly, they should've never boarded a train that mandates uniforms!' implies that the people in the image are being forced to wear uniforms, which is a form of hate. Therefore, this image represents explicit-hate." }</pre>	<pre>{ "answer": "non-hate", "explanation": "The image shows a meme about passengers boarding a train, and the text 'honestly, they should've never boarded a train that mandates uniforms!' is in an ironic tone. The meme is about the ridiculousness of a specific choice of attire for transportation, which has nothing to do with how one's body looks or how safe they are." }</pre>

Comparison of the responses for a joint - *answer prediction* and *explanation generation* for multimodal hate, using the prompt shown on left side of the slide.

Major Challenges & Limitations

- **Data Handling and Quality Issues**

- Challenges in handling **noisy data from OCR** [Nandi, 2022; Mahran,2022].
- Limitations due to **unbalanced classes** and specific dataset characteristics (e.g., *limited size, language specificity*) [Paraschiv, 2022].
- Issues with **high dimensionality and computational demand** [Kumar, 2021].
- Rich set of features and meta-data, characterizing memes, needs datasets large enough to enable **generalization** [Al-Natour, 2021], which is typically **scarce**.
- Keyword-based platform-dependent memes collection **biases** the sample space, **over-representing** typical memetic characteristics [Fairchild, 2020].



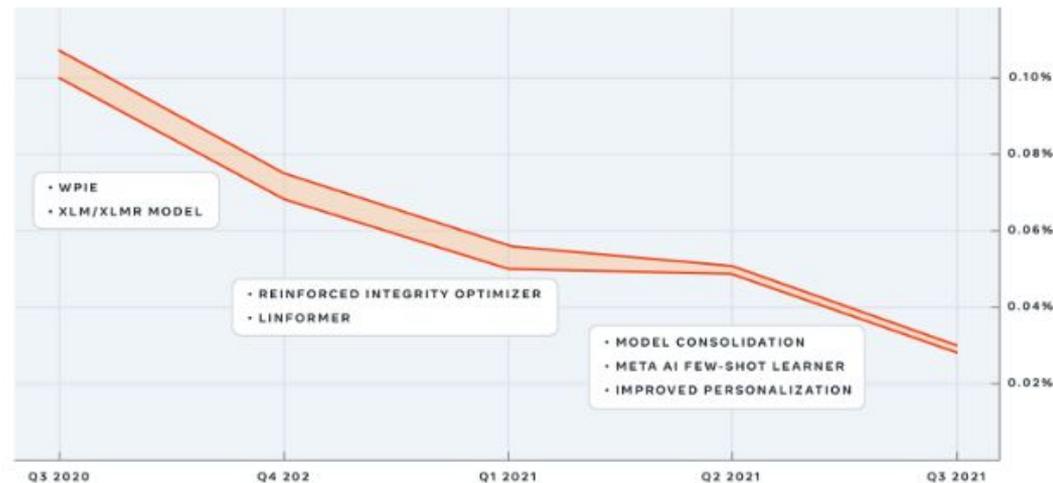
Comparison plots for the top-5 harmfully referenced entities, for their harmful/non-harmful referencing [Sharma et al., 2022].

Major Challenges & Limitations

- **Rapid evolution**

- Harmful memes evolve **quickly**, while devising safety passages over the security frameworks.
- Humans generally good at understanding new harmful concepts quickly, which AI systems **struggle** with [Wang et al., 2020].
- Crucial to advance **multimodal FSL** for quick adaptation [Tsimpoukelli et al., 2021].

Hate speech prevalence



The percentage of views of violating content people see — before and after FSL was rolled out.

Major Challenges & Limitations

- Identifying real instigators of harm

- Poe's law: Differentiation between online satire and extremism **not possible** without critical understanding of the **user's intent** [Greene, 2019].
- Regulatory bodies **struggle** to distinguish *anti-democracy extremists* from *anti-democratic alt-right factions* [Askanius, 2021].
- Caution advised while associating alt-right with culture. Alt-right is a historical phenomenon that *leverages* culture as a tool for its propagation [Dafaure, 2020].



Future Forecasting



CHARACTERIZING
VEHICLES OF HARM



CROSS-CULTURAL
STUDIES



DATA-DRIVEN IN
ADDITION TO
THEORETICAL



RICH METADATA



MEMETIC
MODERATION



MODEL
IMPROVEMENTS AND
EXTENSIONS



DATA HANDLING
ENHANCEMENTS



BETTER TRAINING AND
OPTIMIZATION



APPLICATIONS AND
GENERALIZATIONS



COMPUTATIONAL
EFFICIENCY AND
ACCESSIBILITY

