Can LLMs Reason on Extra-Linguistic Contexts?

Eunjeong Hwang
PhD-ing at UBC
SNU Reading Group Presentation
24.09.09

What are extra-linguistic contexts?

- Information contribute to **additional (implicit) meaning** along with text input.

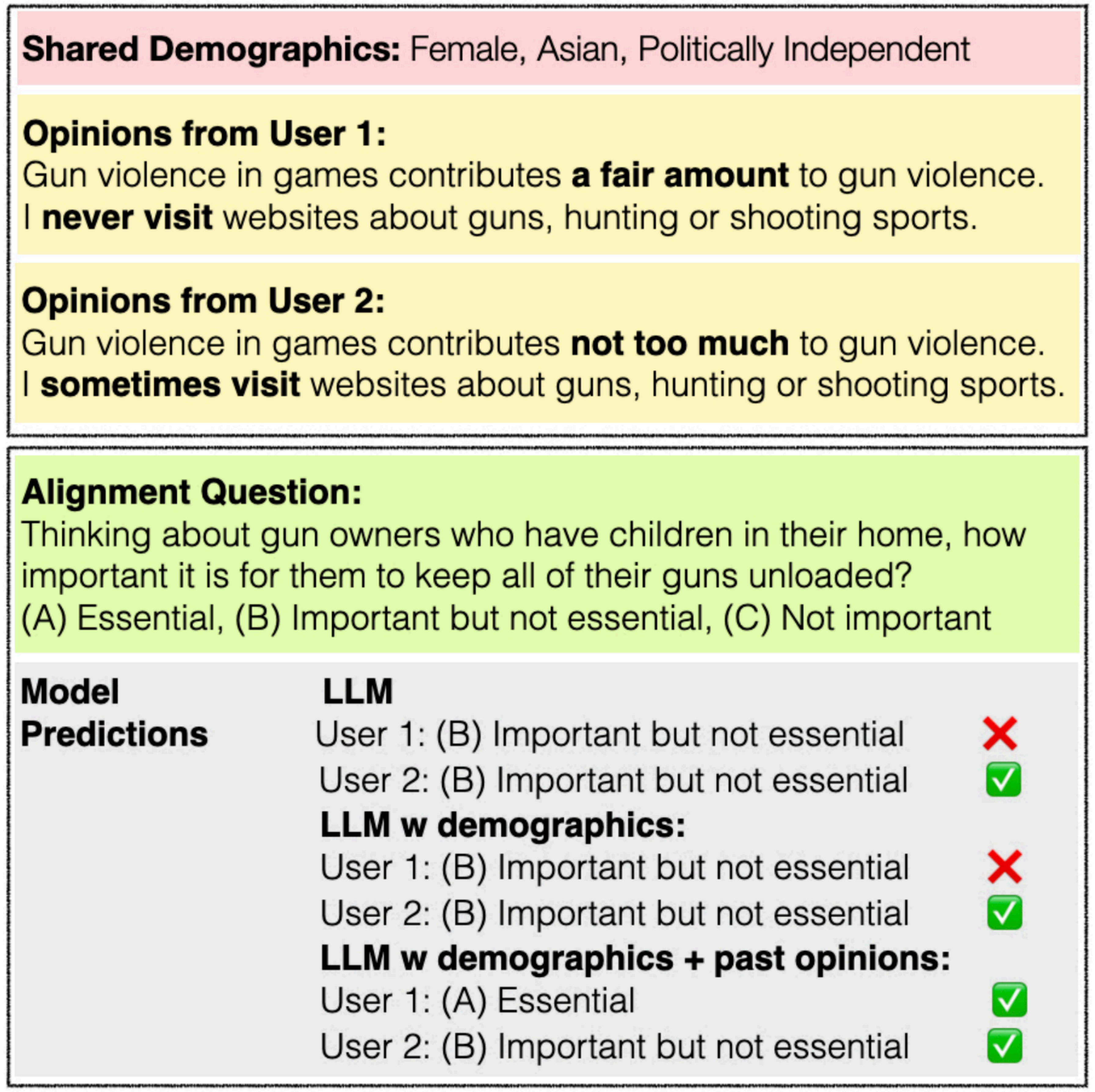
What are extra-linguistic contexts?

- Information contribute to additional (implicit) meaning in the text input.
- Types of extra-linguistic contexts:
  - **Additional Modalities** (e.g. speech and vision).
  - **User's background** (e.g. demographics, culture, ideology)
  - **User's previous interactions** (e.g. opinions, preferences)
  - And more!

What are extra-linguistic contexts?

- Information contribute to additional (implicit) meaning in the text input.
- Types of extra-linguistic contexts:
  - **Additional Modalities** (e.g. speech and vision).
  - **User's background** (e.g. demographics, culture, ideology)
  - **User's previous interactions** (e.g. opinions, preferences)
  - And more!
- Acquired through experience, observations, and social interactions over time.

What are extra-linguistic contexts?



Shared Demographics: Female, Asian, Politically Independent

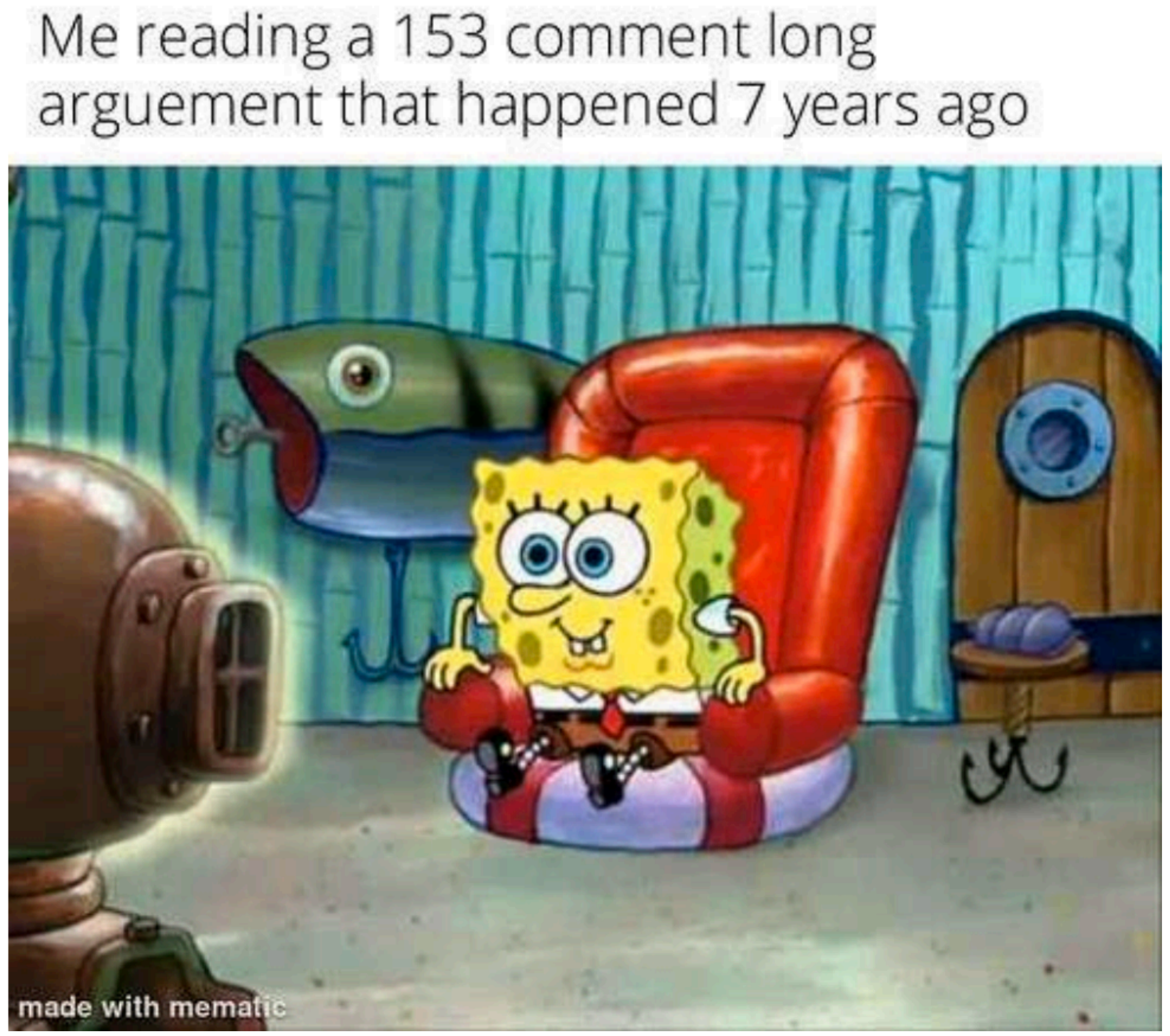
Opinions from User 1:
Gun violence in games contributes **a fair amount** to gun violence.
I **never visit** websites about guns, hunting or shooting sports.

Opinions from User 2:
Gun violence in games contributes **not too much** to gun violence.
I **sometimes visit** websites about guns, hunting or shooting sports.

Alignment Question:
Thinking about gun owners who have children in their home, how important it is for them to keep all of their guns unloaded?
(A) Essential, (B) Important but not essential, (C) Not important

Model Predictions

LLM
User 1: (B) Important but not essential ✗
User 2: (B) Important but not essential ✓
LLM w demographics:
User 1: (B) Important but not essential ✗
User 2: (B) Important but not essential ✓
LLM w demographics + past opinions:
User 1: (A) Essential ✓
User 2: (B) Important but not essential ✓


Me reading a 153 comment long arguement that happened 7 years ago
made with mematic

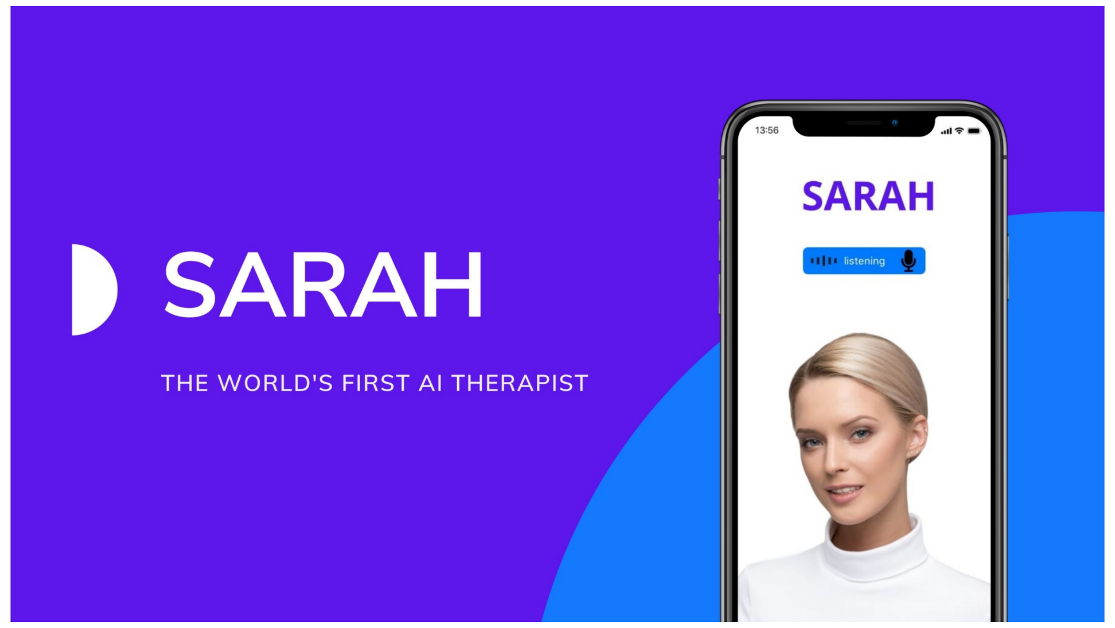
MemeCap: A Dataset for Captioning and Interpreting Memes, Hwang et al., EMNLP 2023.
Aligning Language Models to User Opinions, Hwang et al., EMNLP-Findings 2023.

Why are they important?

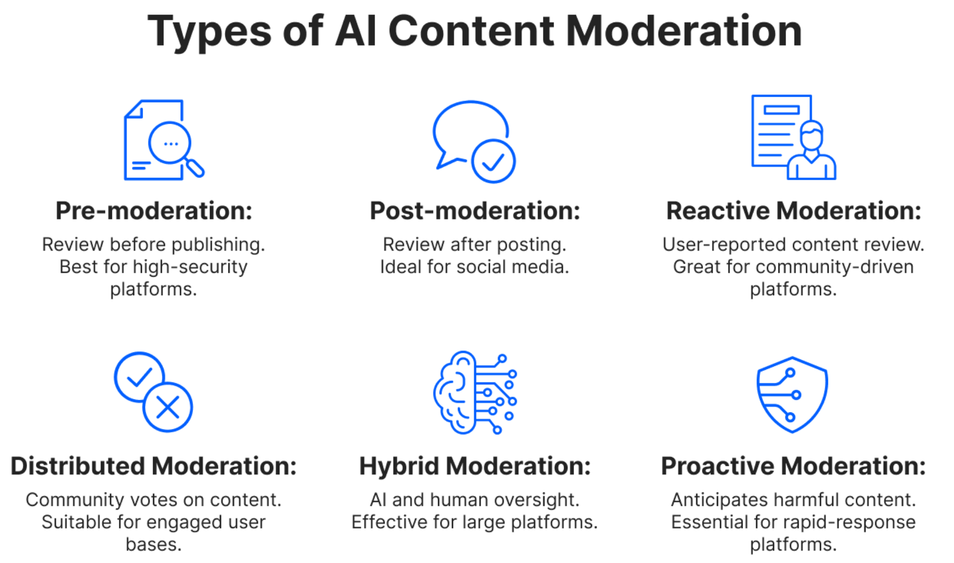
- Recent LLMs focus on text understanding and generation.
- Efficient communication requires additional contexts.
    - speech, vision, situational context (time and space), commonsense knowledge, social and cultural norms

LLMs+Extra-linguistic contexts can…

- .. **adapt their outputs** based on time, location, and the user's background.
- .. improve their **robustness**, addressing unknown inputs reasonably and consistently.
- .. improve their **applicability** to real-world usages.



AI Therapist



Types of AI Content Moderation

Pre-moderation:
Review before publishing. Best for high-security platforms.

Post-moderation:
Review after posting. Ideal for social media.

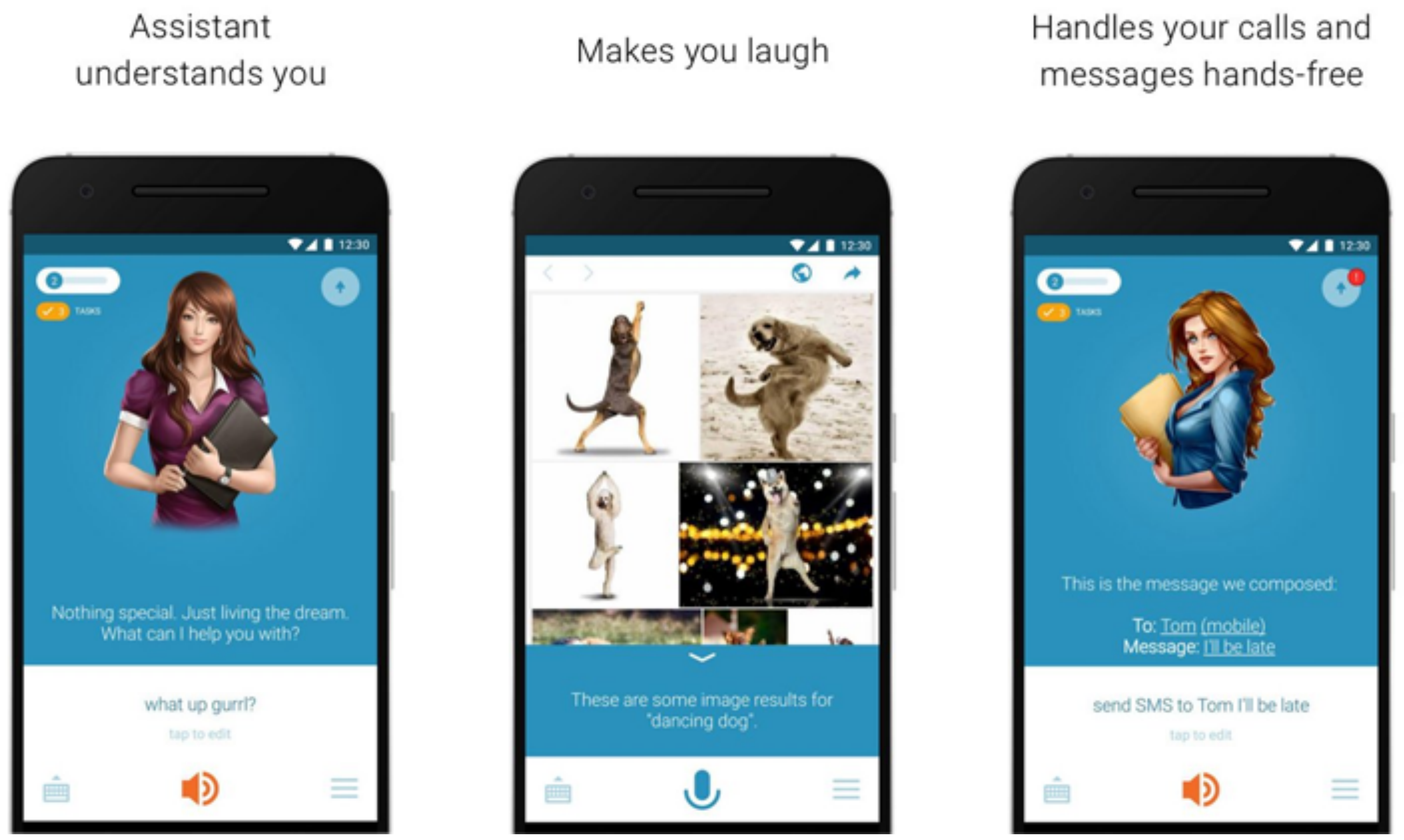
Reactive Moderation:
User-reported content review. Great for community-driven platforms.

Distributed Moderation:
Community votes on content. Suitable for engaged user bases.

Hybrid Moderation:
AI and human oversight. Effective for large platforms.

Proactive Moderation:
Anticipates harmful content. Essential for rapid-response platforms.

AI Content Moderation



Assistant understands you

Nothing special. Just living the dream. What can I help you with?

what up gurrl?

Makes you laugh

These are some image results for "dancing dog".

Handles your calls and messages hands-free

This is the message we composed:

To: Tom (mobile)
Message: I'll be late

send SMS to Tom I'll be late

Dragon Mobile Assistant

Can models reason on...

1️⃣ ... user's previous interactions (e.g. opinions) or background (e.g. demographics)?

2️⃣ ... visual modality when it complements textual inputs?

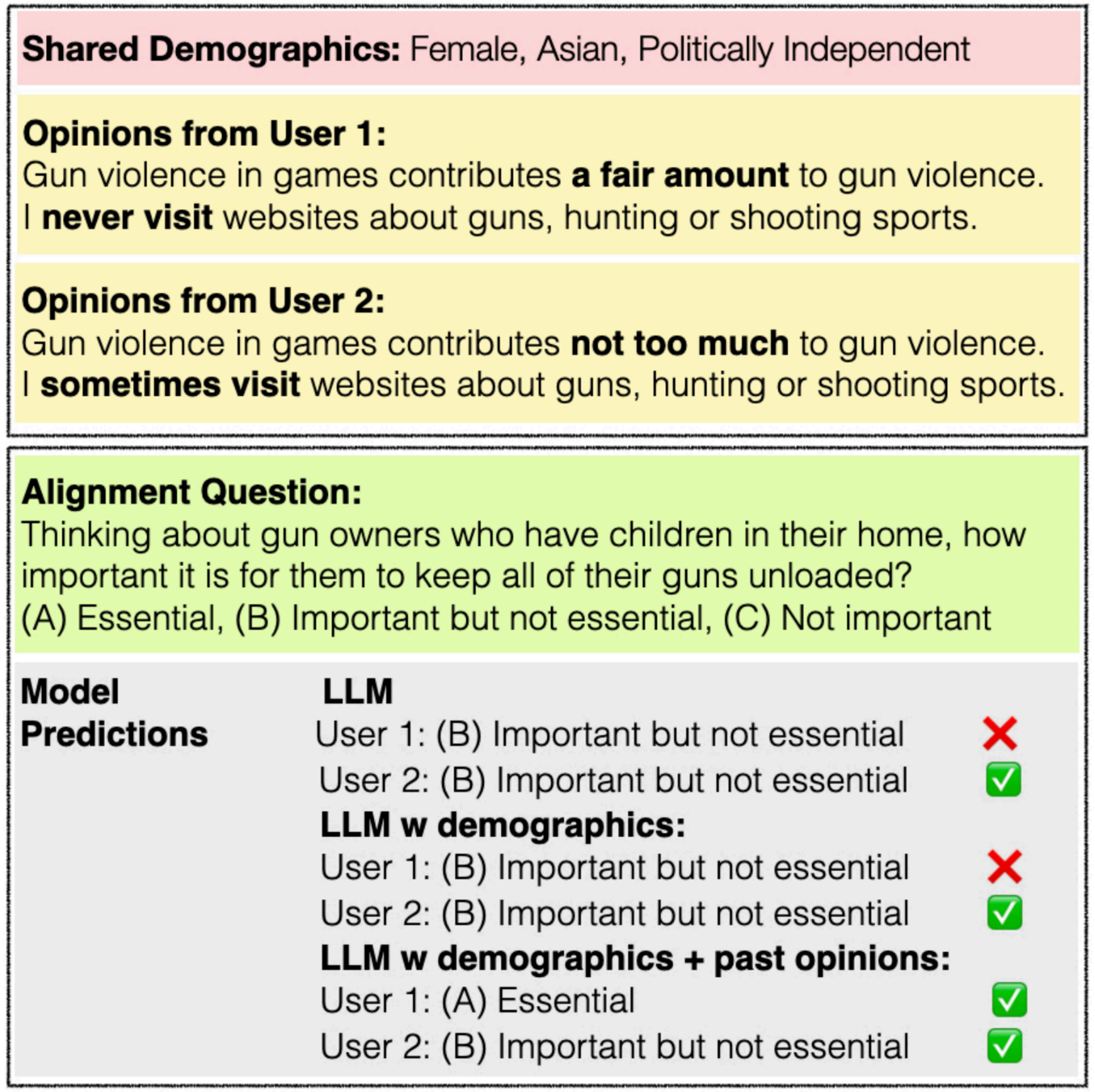
Aligning Language Models to User Opinions

EunJeong Hwang[1,2], **Bodhisattwa Prasad Majumder***[3], and **Niket Tandon***[3]

[1]University of British Columbia
[2]Vector Institute for AI
[3]Allen Institute of AI
ejhwang@cs.ubc.ca, bodhisattwam@allenai.org, nikett@allenai.org

EMNLP-Findings 2023

Insights from Public Opinion Surveys



Shared Demographics: Female, Asian, Politically Independent

Opinions from User 1:
Gun violence in games contributes **a fair amount** to gun violence.
I **never visit** websites about guns, hunting or shooting sports.

Opinions from User 2:
Gun violence in games contributes **not too much** to gun violence.
I **sometimes visit** websites about guns, hunting or shooting sports.

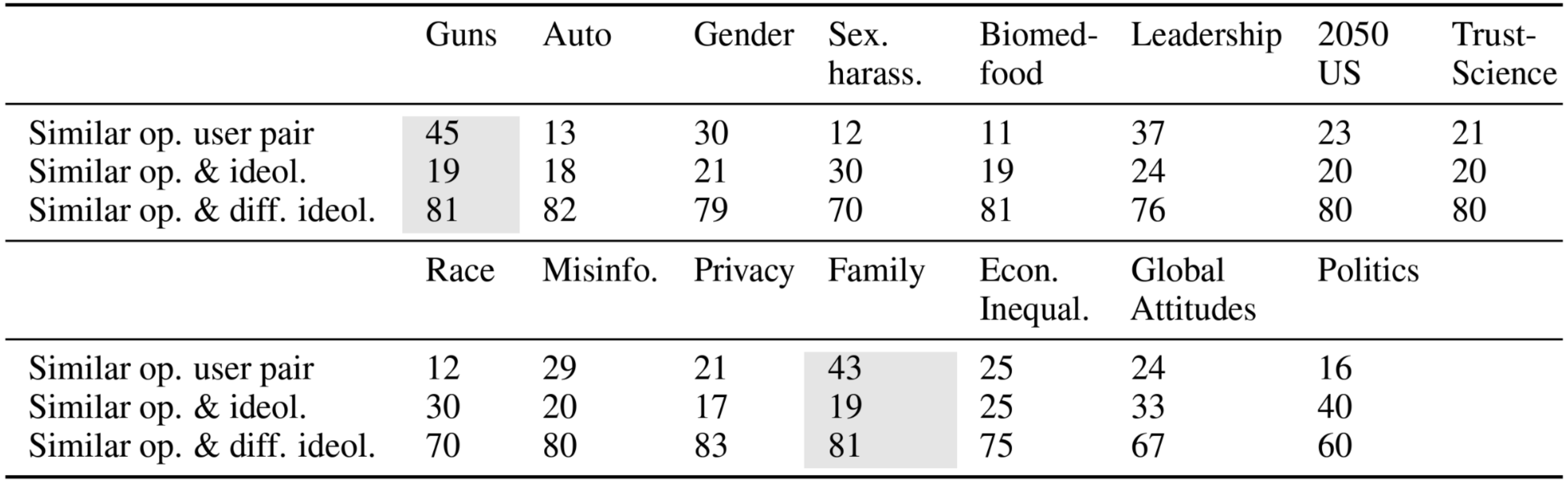
Alignment Question:
Thinking about gun owners who have children in their home, how important it is for them to keep all of their guns unloaded?
(A) Essential, (B) Important but not essential, (C) Not important

Model Predictions

LLM
User 1: (B) Important but not essential ❌
User 2: (B) Important but not essential ✅
LLM w demographics:
User 1: (B) Important but not essential ❌
User 2: (B) Important but not essential ✅
LLM w demographics + past opinions:
User 1: (A) Essential ✅
User 2: (B) Important but not essential ✅

- 15 Topics — 100 questions and 5K users per each topic.
- Major components:
  - **Demographics**: religion, gender, age, education, race, citizen, marital status, income.
  - **Ideology**: political affiliation, inclinations towards political ideologies (e.g. conservative, liberal).
  - **Opinions**: user answers on subjective questions.

Insights from Public Opinion Surveys



	Guns	Auto	Gender	Sex. harass.	Biomed-food	Leadership	2050 US	Trust-Science
Similar op. user pair	45	13	30	12	11	37	23	21
Similar op. & ideol.	19	18	21	30	19	24	20	20
Similar op. & diff. ideol.	81	82	79	70	81	76	80	80

	Race	Misinfo.	Privacy	Family	Econ. Inequal.	Global Attitudes	Politics
Similar op. user pair	12	29	21	43	25	24	16
Similar op. & ideol.	30	20	17	19	25	33	40
Similar op. & diff. ideol.	70	80	83	81	75	67	60

Percentage of user pairs sharing similar opinions and the percentages of similar ideologies and different ideologies.

Opinions differ despite same demographics or same Ideology!

Experimental Results

How would LLMs react to opinions and demographics?

Input: Opinions/demographics/ideology/Question —> Output: Answer choice (user's predicted opinion)

Overall QA accuracy

Input/Model	Vicuna-13B	GPT3.5	GPT4
No Persona	0.36	0.37	0.53
Demo.+Ideo.	0.39	0.47	0.54
Top3 opinions	**0.42**	**0.50**	**0.55**

Some questions might be highly correlated to demographics.

Opinions > Demographics + Ideologies!

top-k opinions: based on cosine similarity between user opinions and question

Experimental Results

How would LLMs react to opinions and demographics?

Input: Opinions/demographics/ideology/Question —> Output: user's opinion prediction

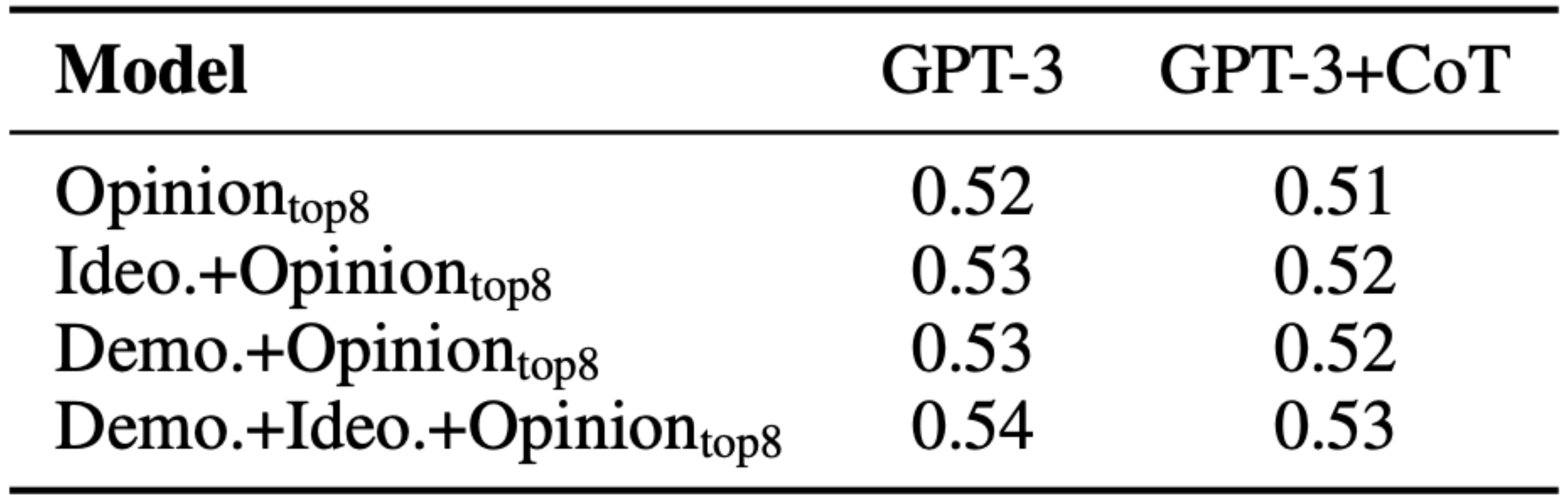
Overall QA accuracy

Input/Model	Vicuna-13B	GPT3.5	GPT4
No Persona	0.36	0.37	0.53
Demo.+Ideo.	0.39	0.47	0.54
Top3 opinions	**0.42**	0.50	0.55
Demo.+Ideo.+All opinions	0.41	0.50	**0.58**
Demo.+Ideo.+Top3 opinions	**0.42**	**0.51**	**0.58**

Using the most relevant opinions is essential.

top-k opinions: based on cosine similarity between user opinions and question

Additional Analysis - CoT



Model	GPT-3	GPT-3+CoT
Opinion$_{\text{top8}}$	0.52	0.51
Ideo.+Opinion$_{\text{top8}}$	0.53	0.52
Demo.+Opinion$_{\text{top8}}$	0.53	0.52
Demo.+Ideo.+Opinion$_{\text{top8}}$	0.54	0.53

QA accuracy comparison (GPT-3 vs. GPT-3 with Chain of Thought style prompt).

CoT prompt consistently decreases the performance by 1 point.

Additional Analysis - CoT

Question: How much do you worry about the following happening to you? Being the victim of a terrorist attack

Choices: **A.** Worry a lot, **B.** Worry a little, **C.** Do not worry at all

Opinions:
- I never used a gun to threaten or intimidate anyone in family.
- There were no guns in my household when I was growing up.

Demographics and Ideology:
- Political party: Republican
- Political ideology: Conservative
- Region: South
- Education: College graduate/some postgrad

...

CoT reasoning:
Step 1: The person is conservative and Republican, so they may be more likely to worry about a terrorist attack.
Step 2: The person is married and has a college degree, so they may be more aware of the potential risks of a terrorist attack.
Step 3: The person is in the South, so they may be more likely to worry about a terrorist attack due to the region's history of terrorism.
Conclusion: The person is likely to select answer choice A: Worry a lot.

User-answer (expected): Do not worry at all. ✅
Demo.+ideo.+opinions with CoT: Worry a lot. ❌
Demo.+ideo.+opinions without CoT: Do not worry at all. ✅

Error Analysis

Question: How well do the following words or phrases describe you? Supporter of the Black Lives Matter movement

Choices: **A.** Describes me well, **B.** Does not describe me well

Opinions:
- The words or phrases “**supporter of rights for LGBT people” do not describe me well**.

…

Demographics and Ideology:
- Political party: Democrat
- **Race: Black**

…

User-answer (expected): Describes me well. ✅
Demo.+ideo.+opinions: Does not describe me well. ❌
Demo.+ideo.: Describes me well. ✅

Recap

- **Opinions** of a user and **their demographics and ideologies** are **not mutual predictors.**
- **Offer insights on aligning LLMs to users** with user demographics, ideologies, and the most relevant past opinions.
- **Using opinions improves up to 7 point** absolute QA accuracy over demography based baselines.

Next Question

Q: How can we train a model to avoid relying on demographics and enhance their understanding of the implicit meaning in subjective sentences?

Next Question

Q: How can we train a model to avoid relying on demographics and enhance their understanding of the implicit meaning in subjective sentences?

A: Let's **make an opinion graph per person** and **reason over it**!

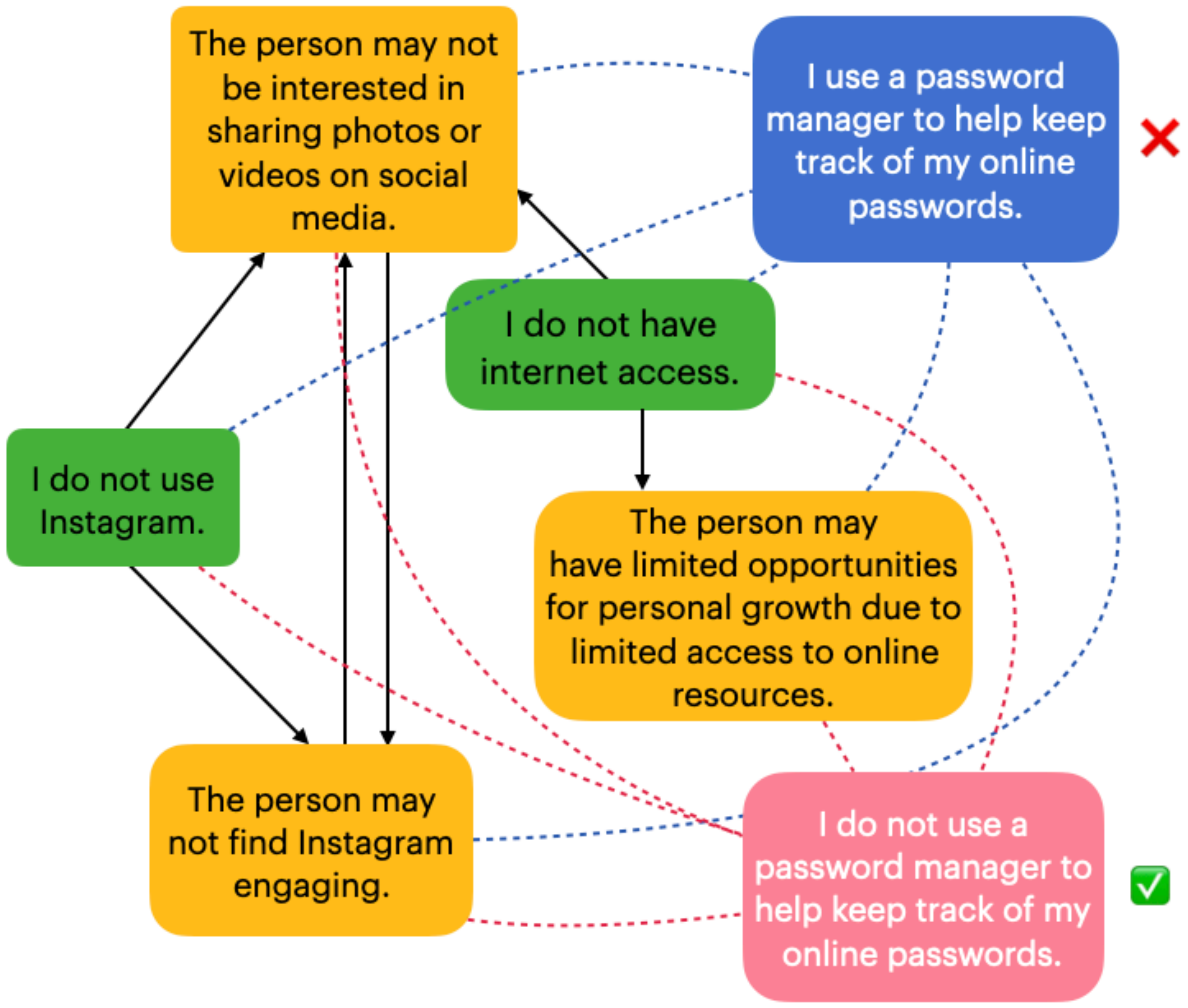
A Graph per Persona: Reasoning about Subjective Natural Language Descriptions

EunJeong Hwang[1,2], **Vered Shwartz**[1,2], **Dan Gutfreund**[3], **and Veronika Thost**[3]
[1] University of British Columbia [2] Vector Institute for AI
[3] MIT-IBM Watson AI Lab, IBM Research
`{ejhwang,vshwartz}@cs.ubc.ca`,
`dgutfre@us.ibm.com`, `veronika.thost@ibm.com`

ACL-Findings 2024

Example

Input:
Past opinions/preferences



The person may not be interested in sharing photos or videos on social media.

I use a password manager to help keep track of my online passwords. ✗

I do not have internet access.

I do not use Instagram.

The person may have limited opportunities for personal growth due to limited access to online resources.

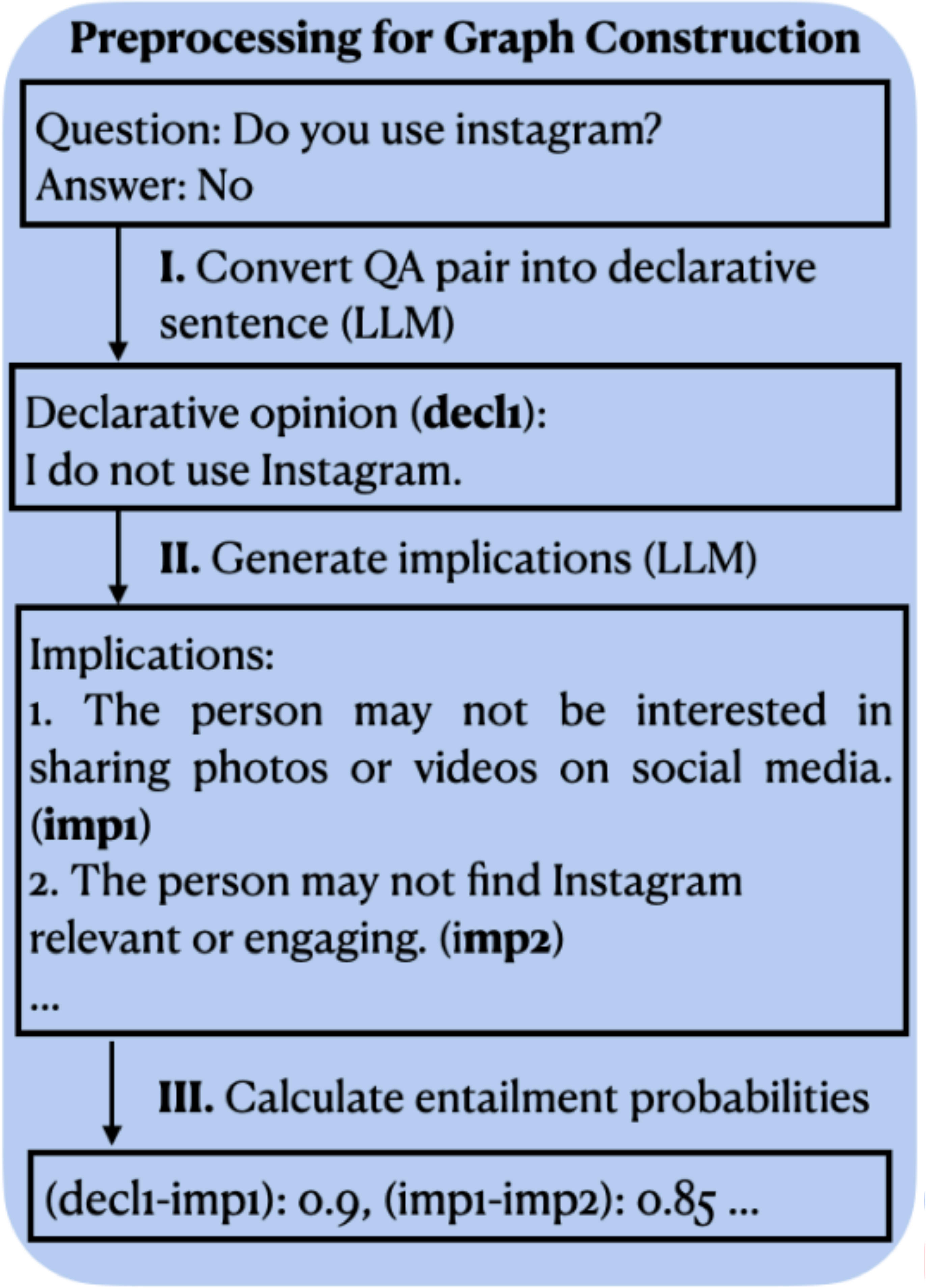
The person may not find Instagram engaging.

I do not use a password manager to help keep track of my online passwords. ✅

Output:
Current opinions/preferences

Method

Step1: Generate Implications & Identify Entailment Relationships



Preprocessing for Graph Construction

Question: Do you use instagram?
Answer: No

I. Convert QA pair into declarative sentence (LLM)

Declarative opinion (**decl1**):
I do not use Instagram.

II. Generate implications (LLM)

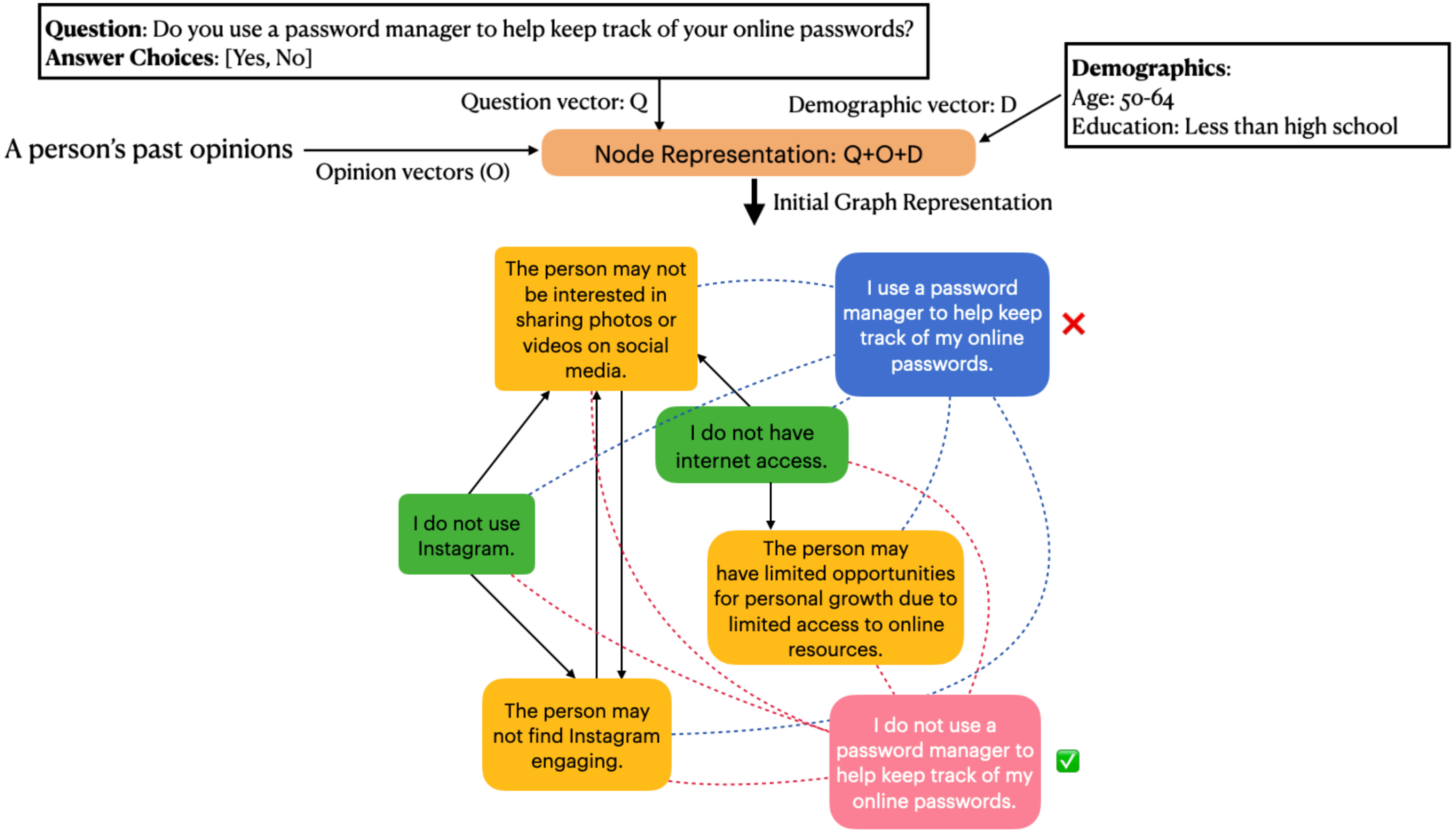
Implications:
1. The person may not be interested in sharing photos or videos on social media. (**imp1**)
2. The person may not find Instagram relevant or engaging. (**imp2**)
...

III. Calculate entailment probabilities

(decl1-imp1): 0.9, (imp1-imp2): 0.85 ...

Method

Step2: Construct Node & Graph



Question: Do you use a password manager to help keep track of your online passwords?
Answer Choices: [Yes, No]

Question vector: Q

Demographic vector: D

Demographics:
Age: 50-64
Education: Less than high school

A person's past opinions

Opinion vectors (O)

Node Representation: Q+O+D

Initial Graph Representation

The person may not be interested in sharing photos or videos on social media.

I use a password manager to help keep track of my online passwords. ✗

I do not have internet access.

I do not use Instagram.

The person may have limited opportunities for personal growth due to limited access to online resources.

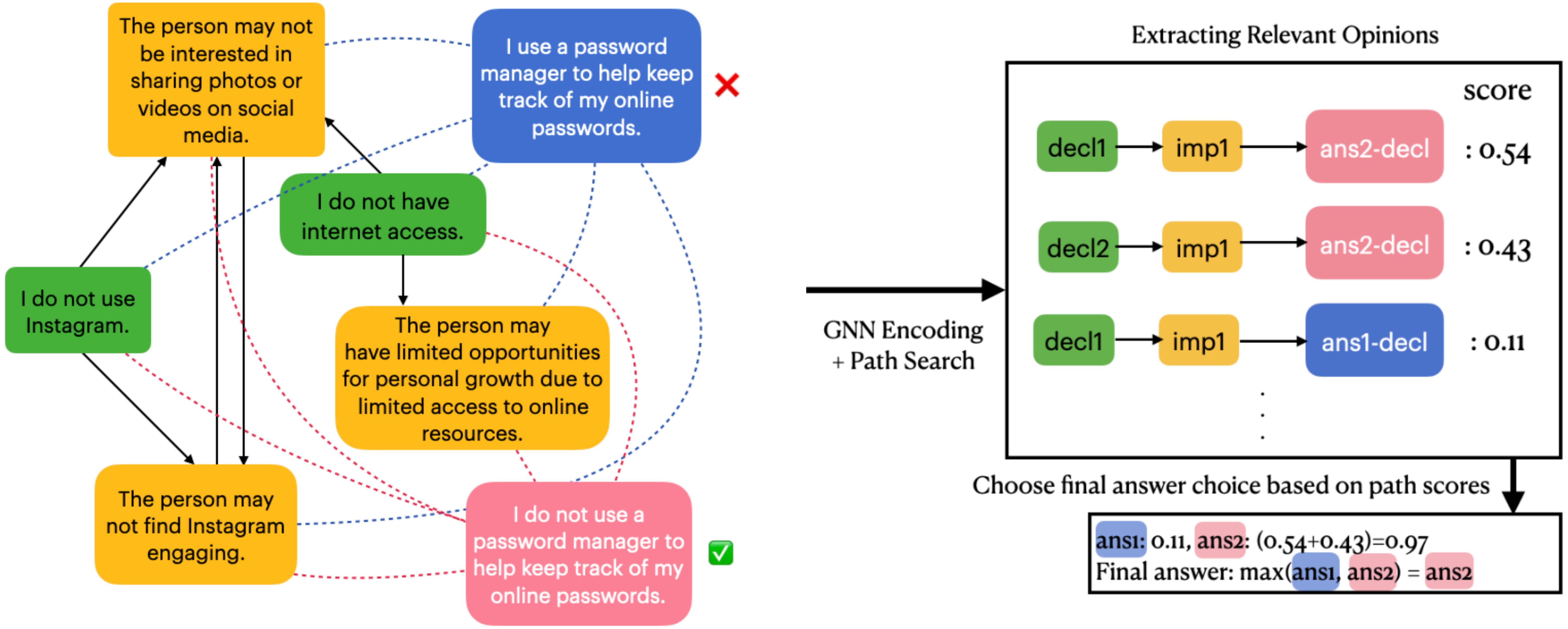
The person may not find Instagram engaging.

I do not use a password manager to help keep track of my online passwords. ✅

Method

Step3: Identify Relevant Opinions using GNN

Goal: compute node representations and attention values (importance of answer choices in context of opinion nodes) by modelling the flow of information over the graph.



The person may not be interested in sharing photos or videos on social media.

I use a password manager to help keep track of my online passwords. ✗

I do not have internet access.

I do not use Instagram.

The person may have limited opportunities for personal growth due to limited access to online resources.

The person may not find Instagram engaging.

I do not use a password manager to help keep track of my online passwords. ✅

GNN Encoding + Path Search

Extracting Relevant Opinions

score

decl1 → imp1 → ans2-decl : 0.54

decl2 → imp1 → ans2-decl : 0.43

decl1 → imp1 → ans1-decl : 0.11

⋮

Choose final answer choice based on path scores

ans1: 0.11, ans2: (0.54+0.43)=0.97

Final answer: max(ans1, ans2) = ans2

Experimental Results

Input: Opinions/Question —> Output: user's opinion prediction

Overall QA accuracy with opinions

Input	top8 opinions			all opinions	+implications	+entailment
Model	BERT	Mistral-7B	GPT-3.5	**GOO** (*ours)		
Avg.	49.2	51.7	50.0	53.3	54.0	**54.9**

Implications help the model improve the performance.

Entailment information further improves the performance.

top-k opinions: based on cosine similarity between user opinions and question

Experimental Results

Input: Opinions/demographics/Question —> Output: user’s opinion prediction

Overall QA accuracy with opinions + demographics

Input	top8 opinions + demographics				all opinions+demographics	+implications
Model	BERT	Mistral-7B	GPT-3.5	ChOiRe	**GOO** (*ours)	
Avg.	49.3	52.8	51.0	51.3	53.3	**54.0**

GOO shows consistently good performance with and without demographics.

top-k opinions: based on cosine similarity between user opinions and question

Analysis on Model Predictions

Both correct	LLM correct	GOO correct	Both incorrect	At least one is correct
34	18	21	27	**73**

Agreement in predictions between Mistral-7B and GOO.

Models complement each other.

Analysis on Model Predictions

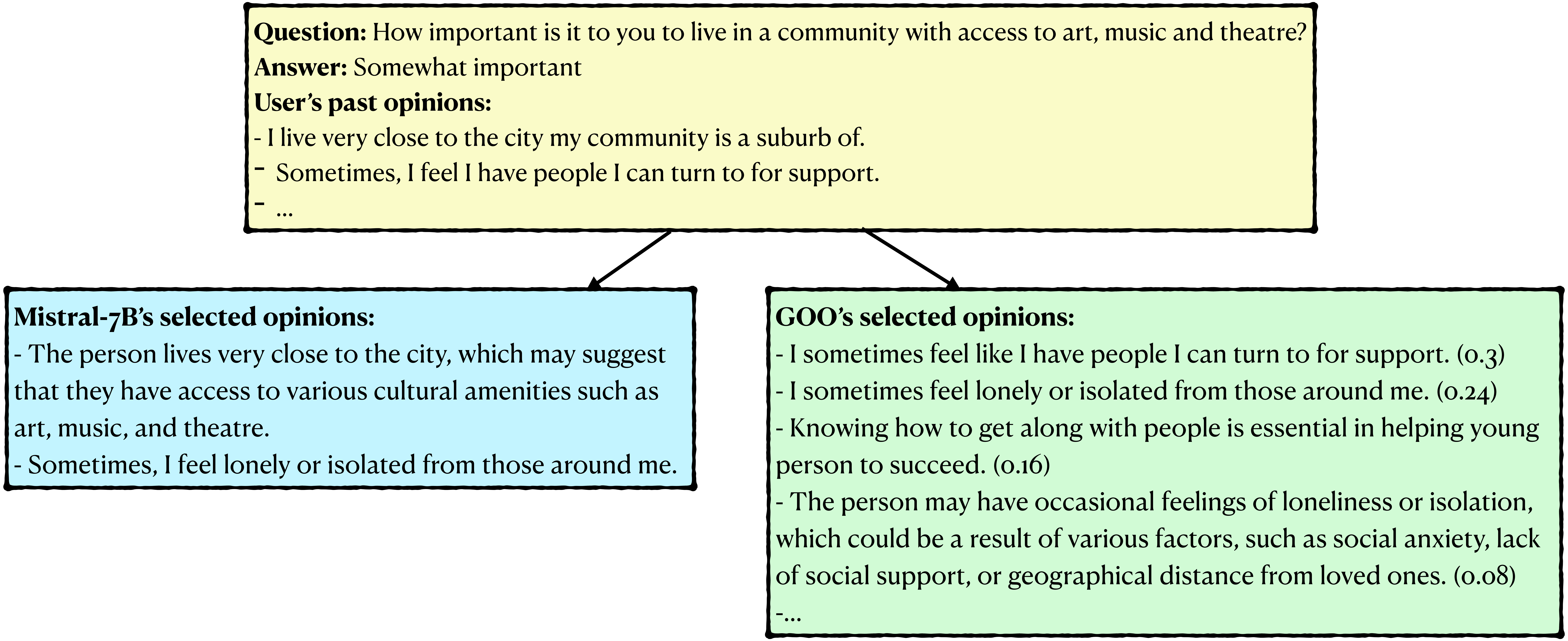
Model	All	Republican	Democrat
Mistral7B with top8 opinions	65	56	64
Mistral-7B with top8 opinions+demographics	**68**	**57**	**67**
GOO with all opinions+implications	74	66	70
GOO with all opinions+implications+demographics	**76**	**69**	**73**

Overlap between model's majority answers and data's majority answers

LLM shows some bias (towards democrat opinions).

Supervised approach captures commonalities.

Reasoning Example



Question: How important is it to you to live in a community with access to art, music and theatre?
Answer: Somewhat important
User's past opinions:
- I live very close to the city my community is a suburb of.
- Sometimes, I feel I have people I can turn to for support.
- ...

Mistral-7B's selected opinions:
- The person lives very close to the city, which may suggest that they have access to various cultural amenities such as art, music, and theatre.
- Sometimes, I feel lonely or isolated from those around me.

GOO's selected opinions:
- I sometimes feel like I have people I can turn to for support. (0.3)
- I sometimes feel lonely or isolated from those around me. (0.24)
- Knowing how to get along with people is essential in helping young person to succeed. (0.16)
- The person may have occasional feelings of loneliness or isolation, which could be a result of various factors, such as social anxiety, lack of social support, or geographical distance from loved ones. (0.08)
-...

Recap

- Presented GOO, a novel approach to reason about subjective knowledge.
- GOO **outperforms several prominent LLMs** and **offer explanations** for its predictions.
- Detailed analysis shows:
    - GOO can have **more equal performance across individuals** compared to LLM.
    - **GOO and LLM can complement each other**, potentially allowing for better opinion prediction.

Future directions

- Combine supervised learning + LLM inference.
- Use feedbacks:
  - When to ignore?
  - When to accept opinions?
- How can we prevent the opinions from being echo chambers/polarization.
- Opinions + culture.

Can models reason on...

1️⃣ ... user's pervious interactions (e.g. opinions) or background (e.g. demographics)?

2️⃣ ... visual modality when it complements textual inputs?

MEMECAP: A Dataset for Captioning and Interpreting Memes

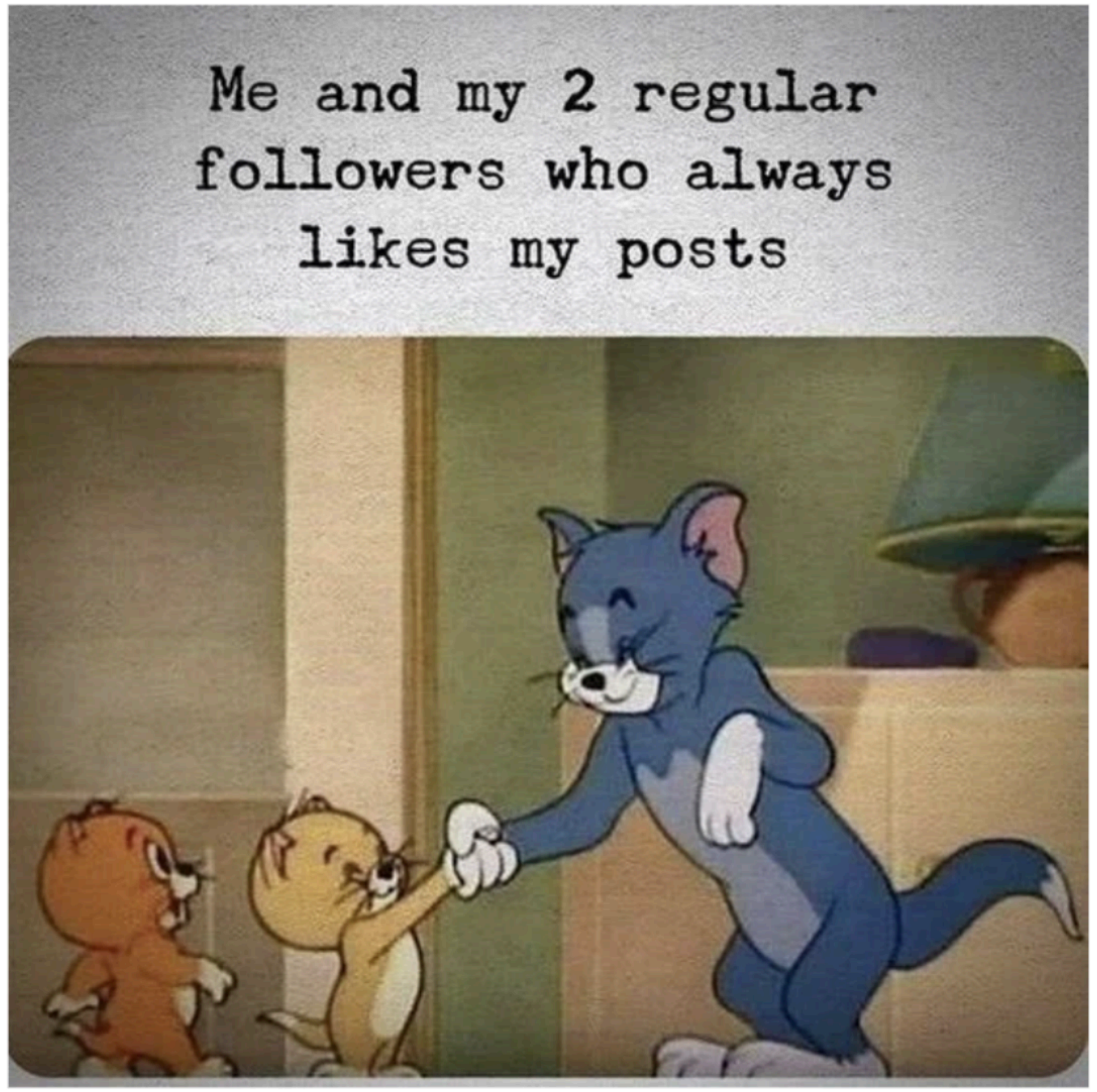
EunJeong Hwang[1,2] **and Vered Shwartz**[1,2]
[1] University of British Columbia [2] Vector Institute for AI
`{ejhwang,vshwartz}@cs.ubc.ca`

EMNLP 2023

Why memes are interesting?

- Requires understanding both the visual and text modalities.

Title: one of them is my alt

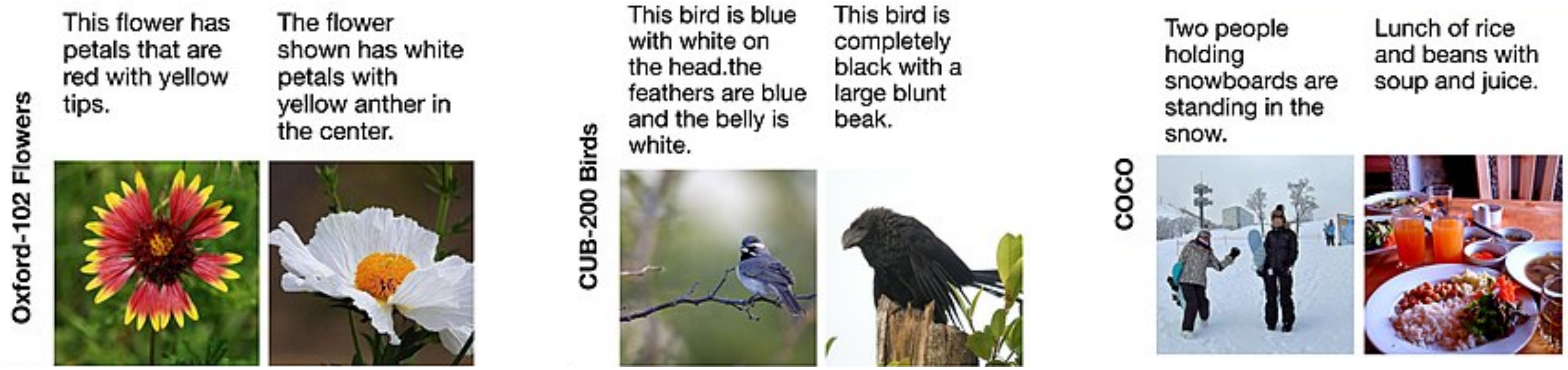


→

Caption:
Meme poster appreciates
their only two followers and
one of them is their alternative account

Why memes are interesting?

- Requires understanding both the visual and text modalities.
- Majority of image captions used for pre-training describe what is depicted in the image.



This flower has petals that are red with yellow tips.

The flower shown has white petals with yellow anther in the center.

This bird is blue with white on the head.the feathers are blue and the belly is white.

This bird is completely black with a large blunt beak.

Two people holding snowboards are standing in the snow.

Lunch of rice and beans with soup and juice.

Why memes are interesting?

- Requires understanding both the visual and text modalities.
- Majority of image captions used for pre-training describe what is depicted in the image.
- Not many works on visual metaphors.

“My bed room is messy” —> “My bedroom is pig sty.”

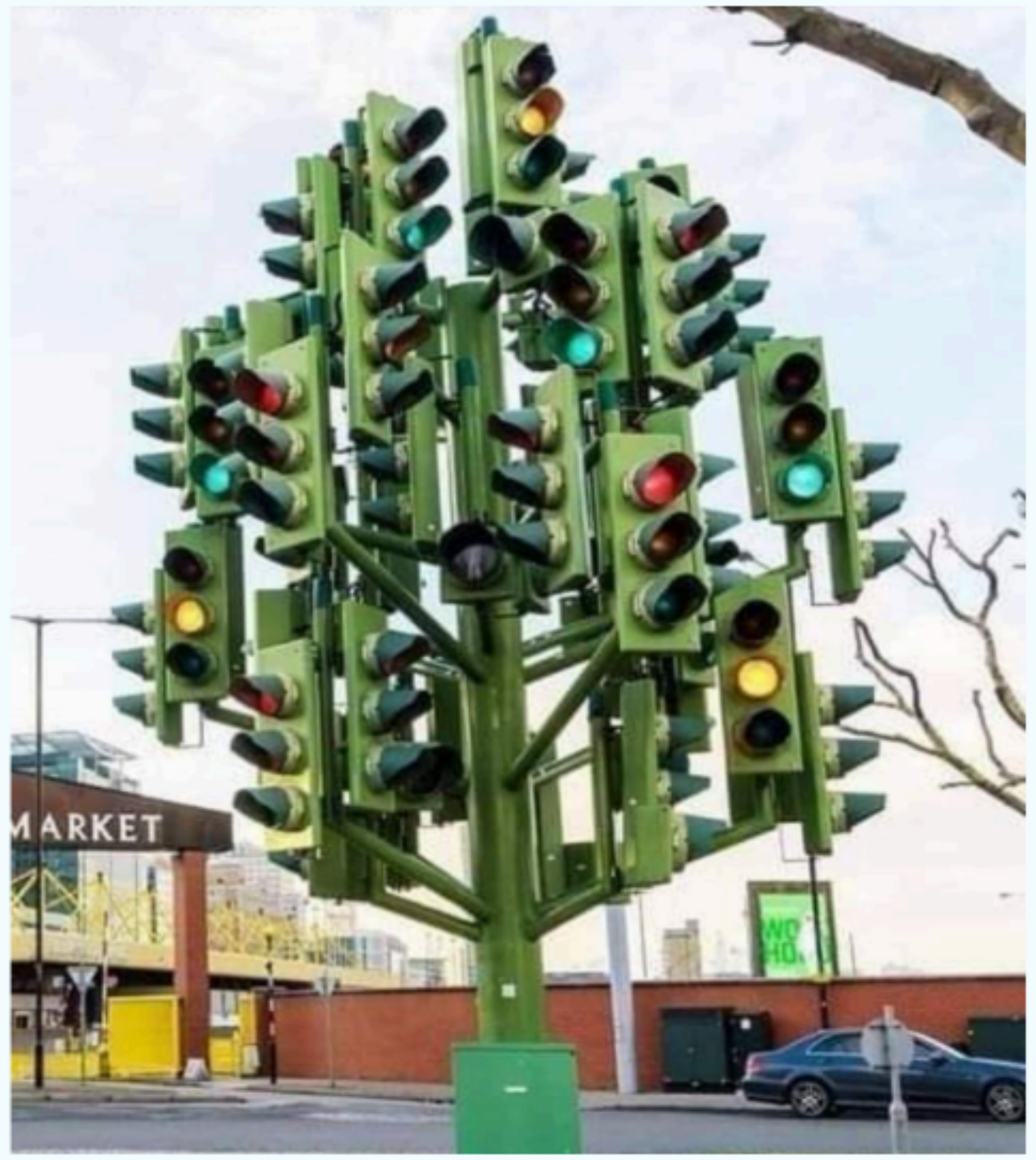
Data Collection

1 Memes

- Scraped from reddit /r/memes
- Manually filtered for quality and to exclude offensive content

2 Literal Image Captions

- Remove text from meme
- Crowdsource the image captions



👤: The worst intersection in the world has to be controlled by a tree of traffic lights

3 Meme Captions and Metaphors

Title: Why they gotta be like this

Her: Why doesn't he understand my signals?
The signals:

Her: why doesn't he understand my signals?
The signals:

👤: intersection = relationship between a man and a woman
👤: tree of traffic lights = the woman's complicated signals

👤: Women wonder why men don't understand their signals when they are overly complicated.

Experimental Results

Model	Inputs	ROUGE-L	BERT-F1
Flamingo	meme+title	39.4	70.8
	meme+title+img cap	39.4	71.0
	meme+title+img cap+OCR text	**43.5**	**73.9**
MiniGPT4	meme+title	30.7	66.2
	meme+title+img cap	28.5	65.8
	meme+title+img cap+OCR text	**31.4**	**68.6**

Best performance is achieved when including OCR text.

Flamingo: few-shot, MiniGPT4: zero-shot

Experimental Results

Model	Inputs	ROUGE-L	BERT-F1
Flamingo-7B	meme+title	39.4	70.8
	meme+title+img cap	39.4	71.0
	meme+title+img cap+OCR text	**43.5**	**73.9**
LLAMA-7B	title+img cap	38.7	70.0
	title+img cap+OCR text	**43.4**	**74.7**

LLaMA model is competitive with Flamingo!

Flamingo: few-shot, LLAMA: few-shot

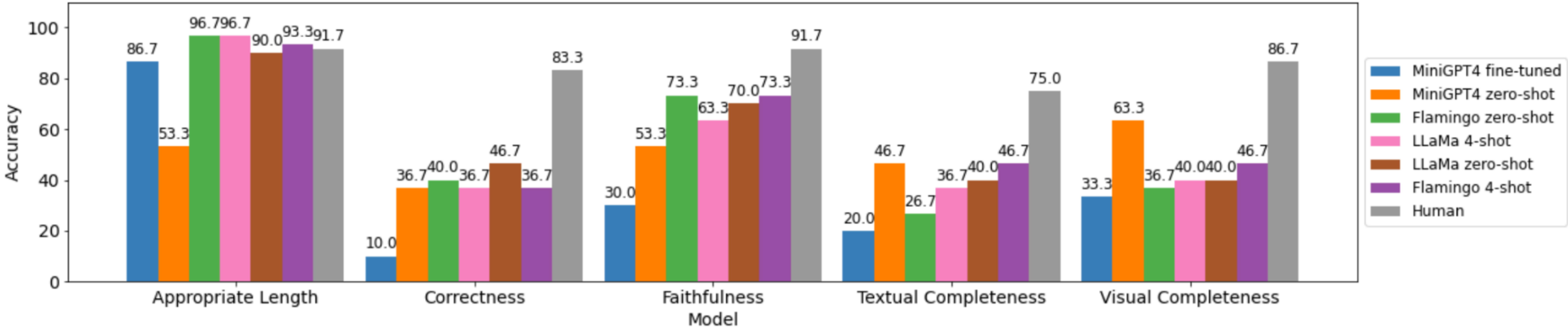
Experimental Results

Model	Inputs	ROUGE-L	BERT-F1
Flamingo-7B	meme+title+img cap+OCR text	**43.5**	73.9
	meme+title+img cap+OCR text+CoT	**43.5**	**74.3**
LLAMA-7B	title+img cap+OCR text	**43.4**	**74.4**
	title+img cap+OCR text+CoT	42.9	74.0

CoT doesn’t help.

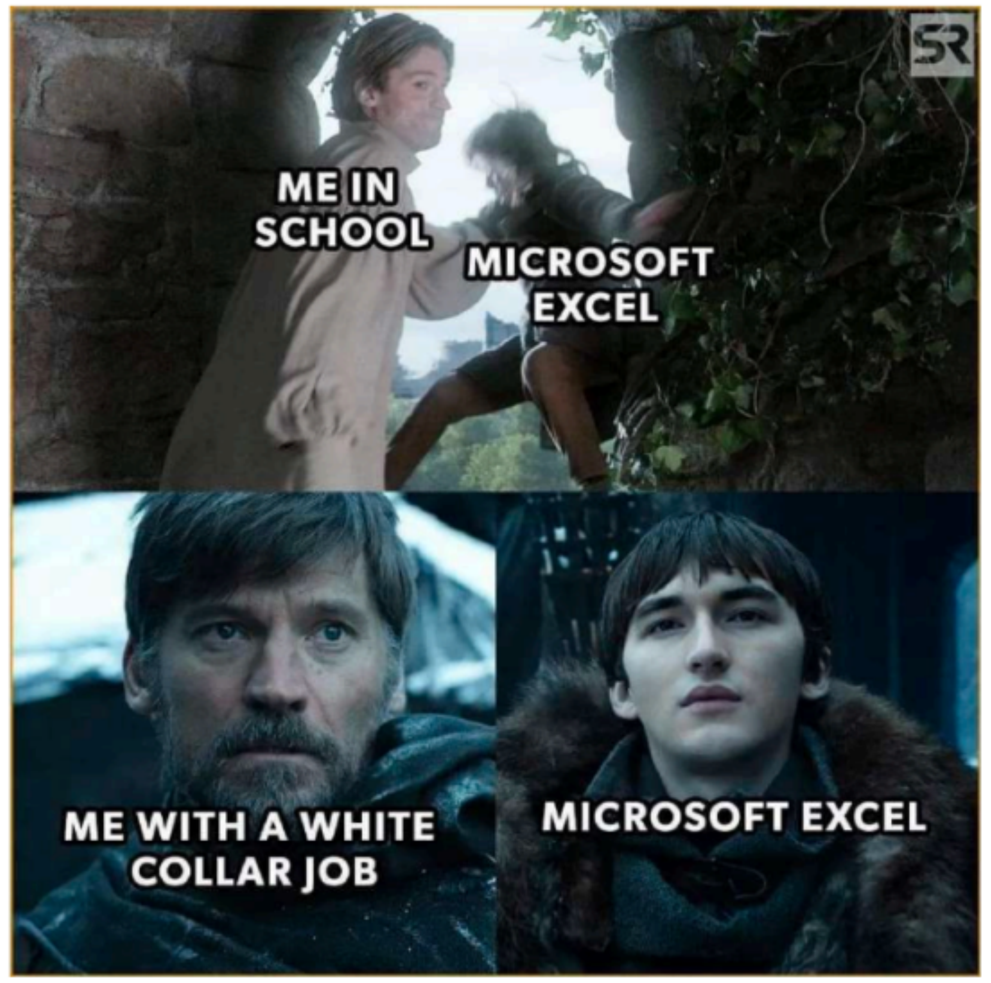
Flamingo: few-shot, LLAMA: few-shot

Human Evaluation



All models perform significantly worse than humans.

Error Analysis



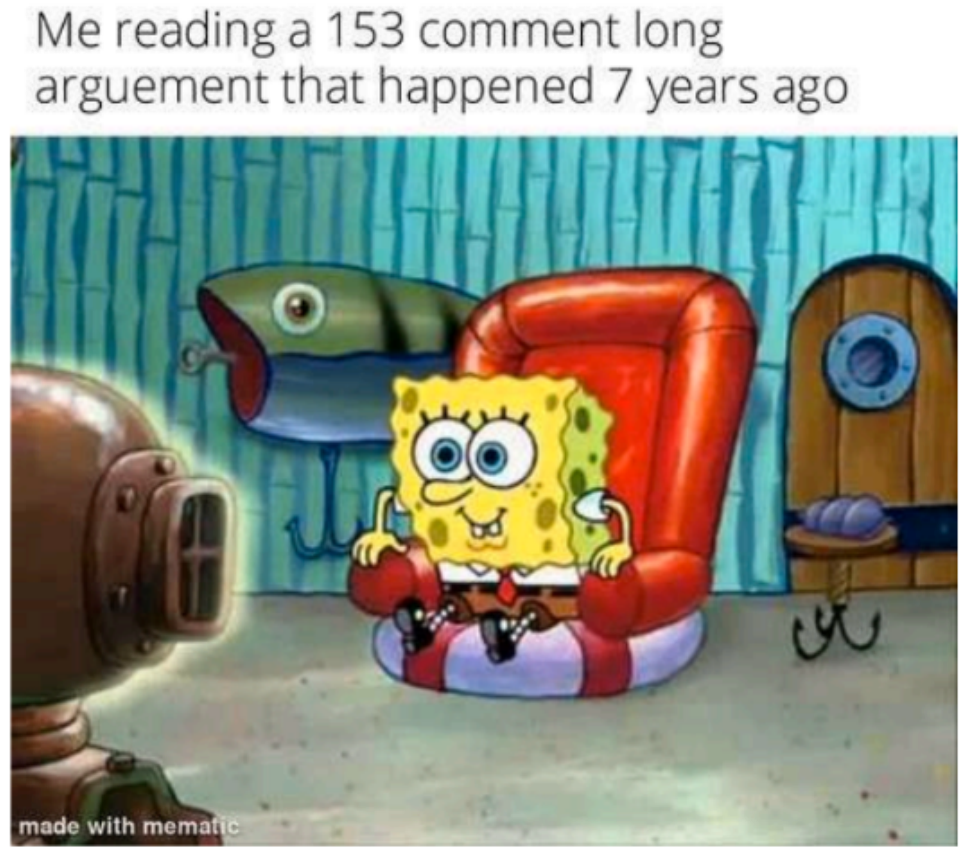
Error: unfaithful

Title: This is my character arc

Image caption: This is a poster of Game of throne from the tower scene.

Human-written meme caption: Meme poster abandoned Microsoft Excel in school, but need to use it after they get their white collar job.

Model-generated meme caption: Meme poster is trying to convey that they want to be successful in life.


Me reading a 153 comment long arguement that happened 7 years ago
made with mematic

Error: visually incomplete (copying the text inside the meme)

Title: Based on a true story

Image caption: Spongebob is eagerly watching TV

Human-written meme caption: Meme poster finds it entertaining to read through long comment threads of arguments that happened in the past.

Model-generated meme caption: Meme poster is trying to convey that they read a 153 comment long argument that happened 7 years ago.

Recap

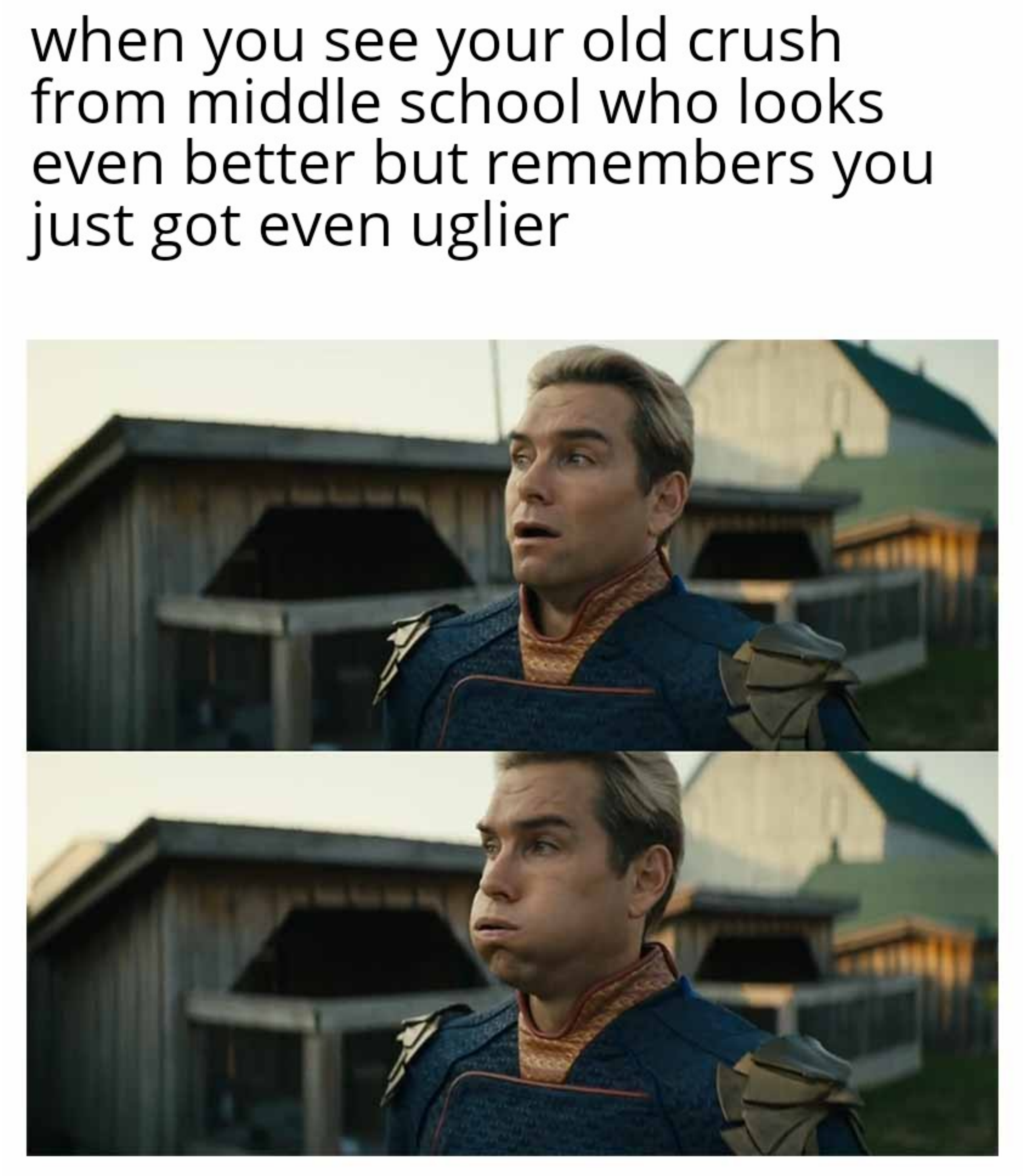
- Present MemeCap, the **first meme captioning dataset**.
- MemeCap requires **recognizing and interpreting visual metaphors**, and **ignoring the literal visual elements**.
- The performances of state-of-the-art VL models are **still far from human performance**.
- VL models tend to **treat visual elements too literally** and **copy text from inside the meme**.

Next Question

Q: How can we integrate implicit meanings to help large language models better understand figurative languages in memes or images?

Potential Answer: Let's make implicit knowledge explicit

Ongoing work



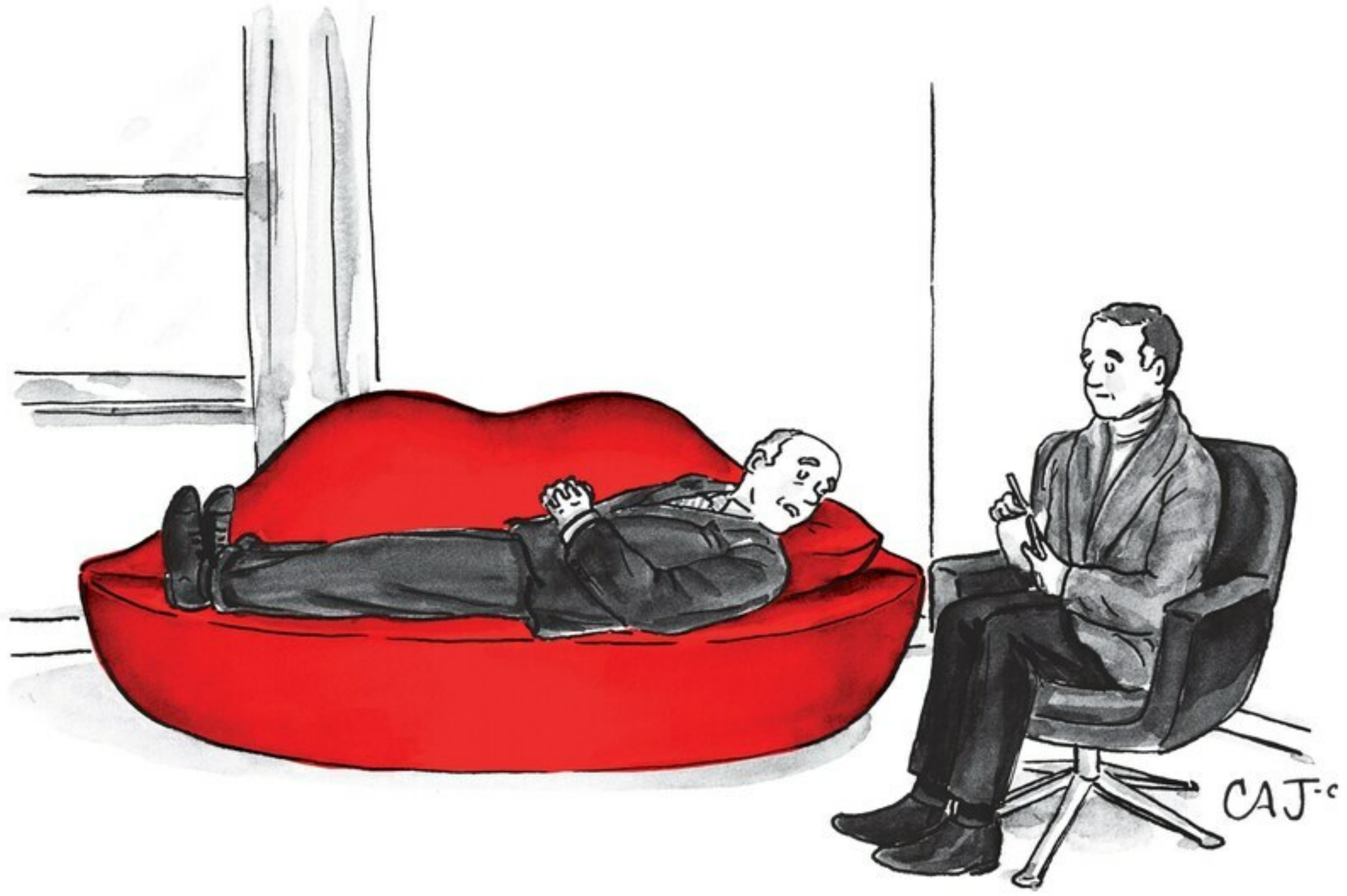
when you see your old crush from middle school who looks even better but remembers you just got even uglier

Title: it’s true

CoT:
The image shows a man reacting to a realization about his past.

CoT with additional information:
The meme humorously depicts the awkwardness of encountering an old crush who has improved over time, while the person's appearance has not.

Ongoing work



Caption: Shouldn’t you be sitting in an ear?

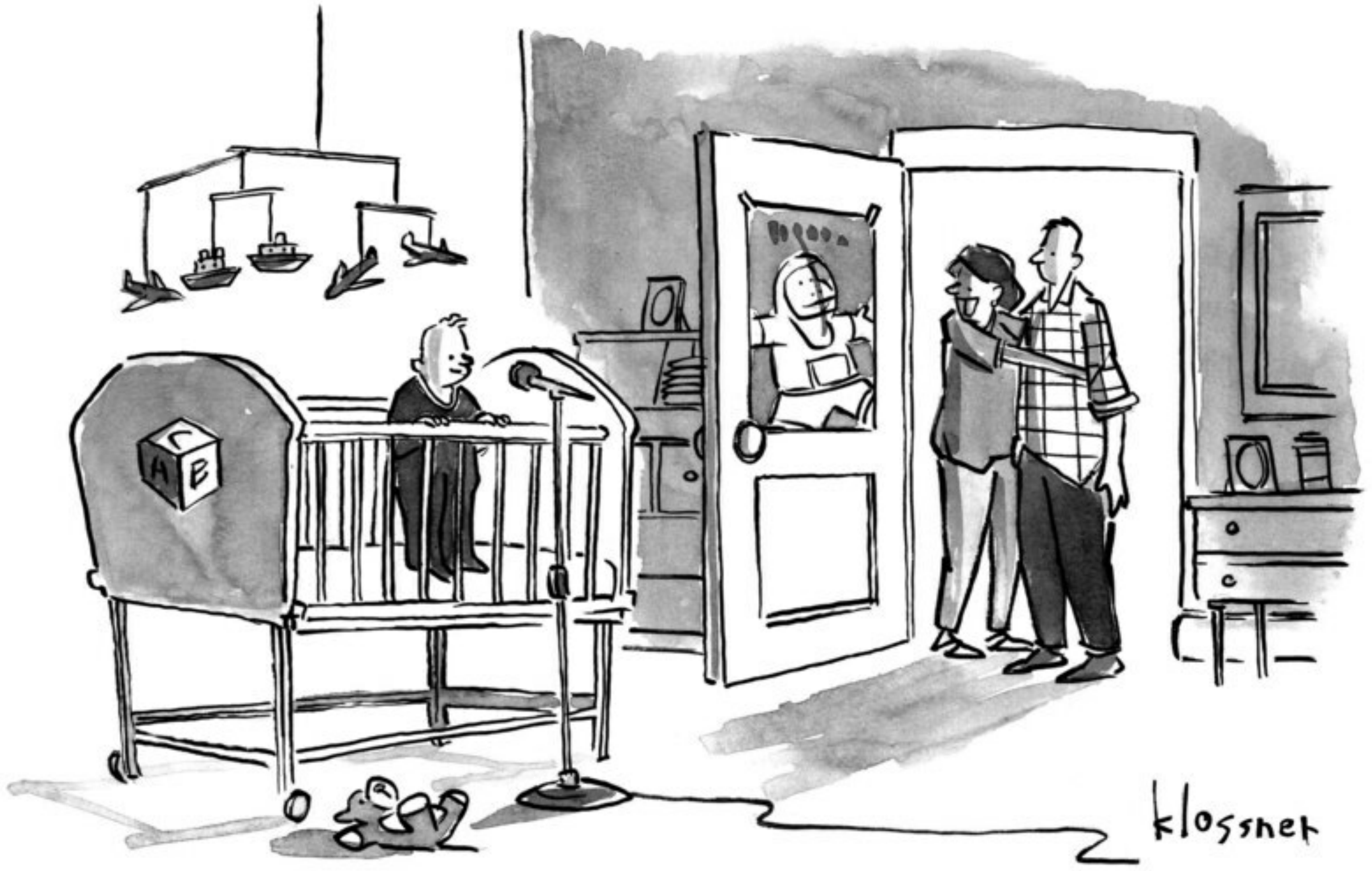
CoT:

...Traditionally, therapists are depicted sitting in a chair while patients lie on a standard couch. **The joke plays on the visual pun, suggesting that since the couch looks like lips, the natural place for it should be an ear, as lips are often close to ears when people talk.** This unexpected twist creates an amusing and playful image.

CoT with additional information:

...The caption, \"Shouldn't you be sitting in an ear?\" **plays on the common phrase that therapists are good listeners, metaphorically having a big ear to listen to their patients.** The juxtaposition of the lip-shaped couch with the idea of sitting in an ear creates an amusing and surreal visual contrast. ...

Ongoing work



Caption: I love his bedtime routine.

CoT:

...The comedic twist lies in this unexpected and exaggerated scenario, highlighting the humor in how over-the-top and formal the bedtime routine has become. ...

CoT with additional information:

...Instead of the baby being calmed by a gentle story or lullaby, a **professional nanny** is reading from a book to the baby in a manner that suggests formal or serious instruction. ...

Intermediate Results

- Additional information is helpful to connect the meanings between visual and textual elements.
- But additional information also adds noise.
- Pending human evaluation.
- Need a better evaluation method other than BERT or GPT score.

Recap

- Models need to:
  - …know **how to connect implicit meanings** across multiple modalities.
  - …**appropriately apply additional knowledge** in context.
    - How to select relevant knowledge? When is appropriate to apply it?
  - …make **consistent predictions**, whether the knowledge is explicit or implicit.

Thank you,
Questions?