Improving Language Models through Context

Dong-Ho Lee

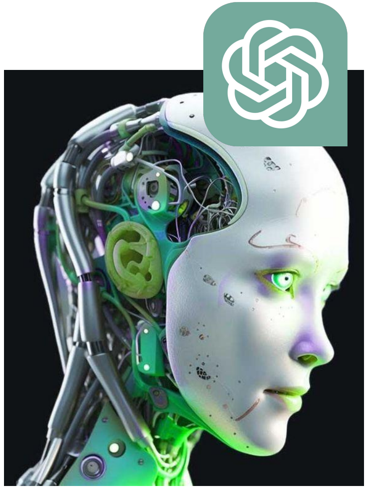
Thesis Defense (April 28, 2025)

Professor Jay Pujara (Chair)
Professor Xiang Ren
Professor Robin Jia
Professor Fred Morstatter
Professor Meisam Razaviyayn

Ambiguity in Human-AI Communication

Is this positive or negative sentiment?

Final exam is the best gift on my birthday.



Is this positive or negative sentiment?

Final exam is the best gift on my birthday

The sentiment of the statement is **positive**. The use of "best gift" suggests a positive connotation, implying that the speaker finds joy or humor in the situation, even if it's slightly ironic.

What if we have **context** ?

Is this sarcasm or not ?

Final exam is the best gift on my birthday

Yes, this is likely sarcasm. The phrase suggests a mismatch between the expectation of a birthday gift being something enjoyable and the reality of having a final exam, which most people would not consider a "gift." This contrast creates the sarcastic tone.

Context

Is this positive or negative sentiment?

Final exam is the best gift on my birthday

Flipped !

The sentiment is **negative**, despite the phrasing appearing positive at face value. The sarcasm in the statement conveys dissatisfaction or frustration, implying that having a final exam on one's birthday is far from ideal or enjoyable.

What are the **examples of helpful context** for AI?

Is this positive or negative sentiment?

Final exam is the best gift on my birthday.

Human Explanation

The speaker uses sarcasm to express frustration, implying negative sentiment about having a final exam on their birthday.

What are the **examples of helpful context** for AI?

Is this positive or negative sentiment?

Final exam is the best gift on my birthday.

Human Explanation

The speaker uses sarcasm to express frustration, implying negative sentiment about having a final exam on their birthday.

Task Examples

Positive example:
Spending time with my family was the best gift on my birthday.

Negative example:
Getting laid off is the best present on my anniversary.

What are the examples of helpful context for AI?

Is this positive or negative sentiment?

Final exam is the best gift on my birthday.

Human Explanation

The speaker uses sarcasm to express frustration, implying negative sentiment about having a final exam on their birthday.

Dialogue, Self-Guided (CoT), ...

Task Examples

Positive example:
Spending time with my family was the best gift on my birthday.

Negative example:
Getting laid off is the best present on my anniversary.

Research Questions

- **What specific types of context** can be used to improve model predictions?
- **For which tasks** does context effectively enhance performance, and for which does it have limited or no benefit?
- Does explicitly **training language models with contextual information** improve their overall performance?
- Can language models autonomously **generate context** to improve their own outputs?

Goals

- **Context-aware inference in language models.**

  Can LMs effectively use context during inference to improve performance on various tasks?

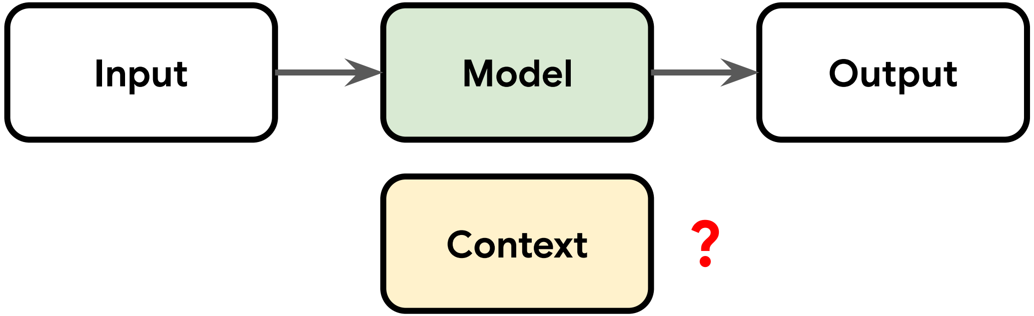
- **Contextual supervision for language model training**

  Does incorporating context during training enhance model behavior?

- **Language models as self-refining context generators**

  Can LMs generate and refine context autonomously, to improve their own downstream outputs?

Context Framework



[FewNER] Good Examples Make A Faster Learner: Simple Demonstration-based Learning for Low-Resource NER., Lee et al., 2022
[Data-Creation] Making Large Language Models Better Data Creators., Lee et al., 2023
[TKG-LLM] Temporal Knowledge Graph Forecasting Without Knowledge Using In-Context Learning., Lee et al., 2023
[STAR] STAR: A SImple Training-free Approach for Recommendations using Large Language Models., Lee et al., 2024
[NormVio-RT] Analyzing Norm Violations in Live-Stream Chat., Lee et al., 2023
[TriggerNER] TriggerNER: Learning with Entity Triggers as Explanations for Named Entity Recognition., Lee et al., 2020
[LEAN-LIFE] LEAN-LIFE: A Label-Efficient Annotation Framework Towards Learning from Explanation., Lee et al., 2020
[AutoTriggER] AutoTriggER: Label-Efficient and Robust Named Entity Recognition with Auxiliary Trigger Extraction., Lee et al., 2023
[XMD] XMD: An End-to-End Framework for Interactive Explanation-Based Debugging of NLP Models., Lee et al., 2023
[REALTALK] REALTALK: A 21-Day Real-World Dataset for Long-Term Conversation, Lee et al., 2025
[QUEST] What is a Good Question? Utility Estimation with LLM-based Simulations., Lee et al., 2025

PhD Journey

Different Types of Context

Input → Model → Output

Context ?

Task Examples

FewNER (ACL 2022)
Data-Creation (EMNLP 2023)
TKG-LLM (EMNLP 2023)

Dialogue Context

NormVio-RT (EMNLP 2023)
LoCoMo (ACL 2024)
REALTALK (2025)

Explanation as Context

TriggerNER (ACL 2020)
LEAN-LIFE (ACL 2020 Demo)
AutoTriggER (EACL 2023)
XMD (ACL 2023 Demo)

Data Context

STAR (2025)

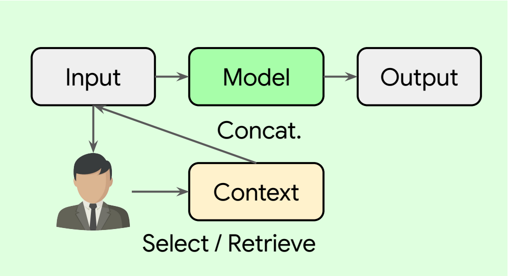
Model-Generated Context

QUEST (2025)

[FewNER] Good Examples Make A Faster Learner: Simple Demonstration-based Learning for Low-Resource NER., Lee et al., 2022
[Data-Creation] Making Large Language Models Better Data Creators., Lee et al., 2023
[TKG-LLM] Temporal Knowledge Graph Forecasting Without Knowledge Using In-Context Learning., Lee et al., 2023
[STAR] STAR: A SImple Training-free Approach for Recommendations using Large Language Models., Lee et al., 2024
[NormVio-RT] Analyzing Norm Violations in Live-Stream Chat., Lee et al., 2023
[TriggerNER] TriggerNER: Learning with Entity Triggers as Explanations for Named Entity Recognition., Lee et al., 2020
[LEAN-LIFE] LEAN-LIFE: A Label-Efficient Annotation Framework Towards Learning from Explanation., Lee et al., 2020
[AutoTriggER] AutoTriggER: Label-Efficient and Robust Named Entity Recognition with Auxiliary Trigger Extraction., Lee et al., 2023
[XMD] XMD: An End-to-End Framework for Interactive Explanation-Based Debugging of NLP Models., Lee et al., 2023
[REALTALK] REALTALK: A 21-Day Real-World Dataset for Long-Term Conversation, Lee et al., 2025
[QUEST] What is a Good Question? Utility Estimation with LLM-based Simulations., Lee et al., 2025

PhD Journey

Context Frameworks



Inference with Context

Task Examples

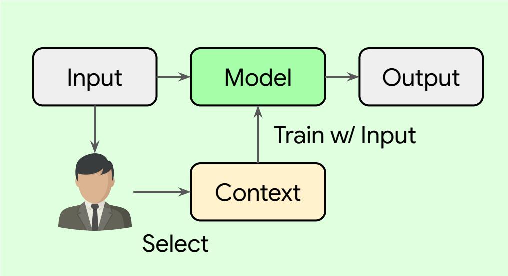
FewNER (ACL 2022)
Data-Creation (EMNLP 2023)
TKG-LLM (EMNLP 2023)

Data Context

STAR (2025)

Dialogue Context

REALTALK (2025)



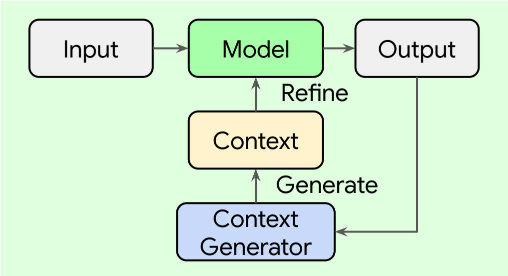
Training with Context

Explanation as Context

TriggerNER (ACL 2020)
LEAN-LIFE (ACL 2020 Demo)
AutoTriggER (EACL 2023)
XMD (ACL 2023 Demo)

Dialogue Context

NormVio-RT (EMNLP 2023)



Generating Context

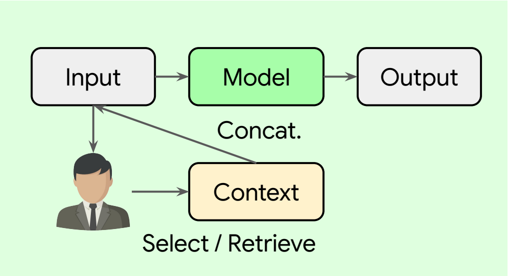
Model Generated Context

QUEST (2025)

[FewNER] Good Examples Make A Faster Learner: Simple Demonstration-based Learning for Low-Resource NER., Lee et al., 2022
[Data-Creation] Making Large Language Models Better Data Creators., Lee et al., 2023
[TKG-LLM] Temporal Knowledge Graph Forecasting Without Knowledge Using In-Context Learning., Lee et al., 2023
[STAR] STAR: A SImple Training-free Approach for Recommendations using Large Language Models., Lee et al., 2024
[NormVio-RT] Analyzing Norm Violations in Live-Stream Chat., Lee et al., 2023
[TriggerNER] TriggerNER: Learning with Entity Triggers as Explanations for Named Entity Recognition., Lee et al., 2020
[LEAN-LIFE] LEAN-LIFE: A Label-Efficient Annotation Framework Towards Learning from Explanation., Lee et al., 2020
[AutoTriggER] AutoTriggER: Label-Efficient and Robust Named Entity Recognition with Auxiliary Trigger Extraction., Lee et al., 2023
[XMD] XMD: An End-to-End Framework for Interactive Explanation-Based Debugging of NLP Models., Lee et al., 2023
[REALTALK] REALTALK: A 21-Day Real-World Dataset for Long-Term Conversation., Lee et al., 2025
[QUEST] What is a Good Question? Utility Estimation with LLM-based Simulations., Lee et al., 2025

PhD Journey

Context Frameworks



Inference with Context

Task Examples

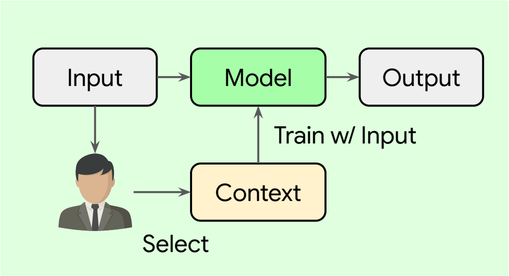
FewNER (ACL 2022)
Data-Creation (EMNLP 2023)
TKG-LLM (EMNLP 2023)

Data Context

STAR (2025)

Dialogue Context

REALTALK (2025)



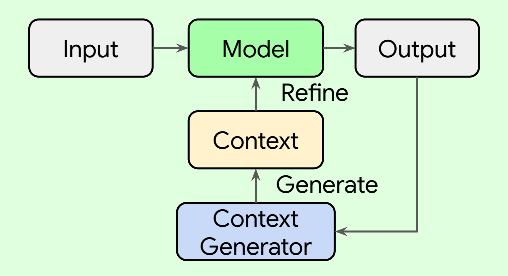
Training with Context

Explanation as Context

TriggerNER (ACL 2020)
LEAN-LIFE (ACL 2020 Demo)
AutoTriggER (EACL 2023)
XMD (ACL 2023 Demo)

Dialogue Context

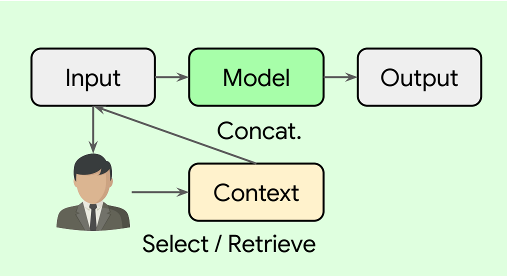
NormVio-RT (EMNLP 2023)



Generating Context

Model Generated Context

QUEST (2025)



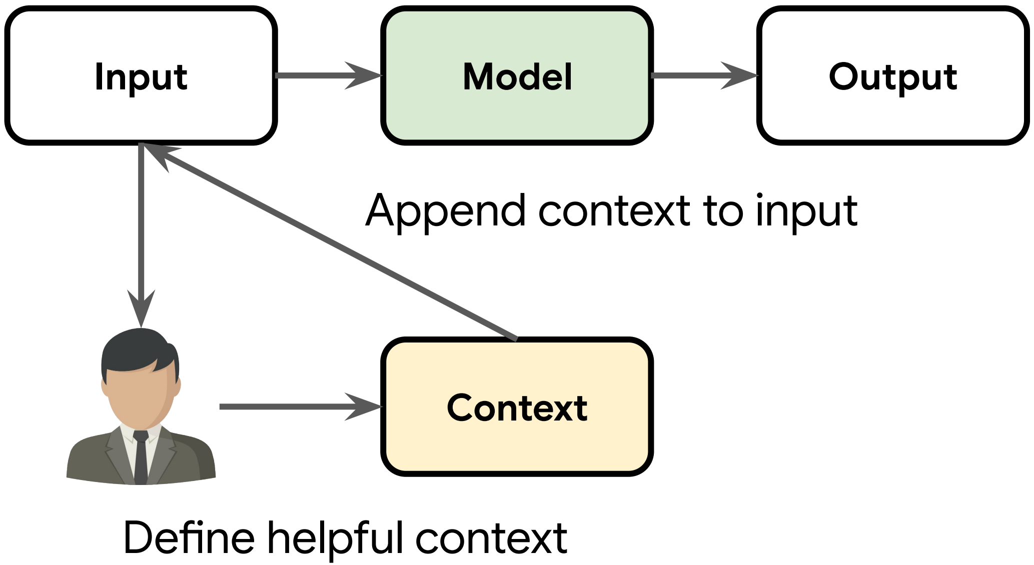
RQ1. Can models learn from context?

Can LMs effectively use context during inference to improve performance on various tasks?

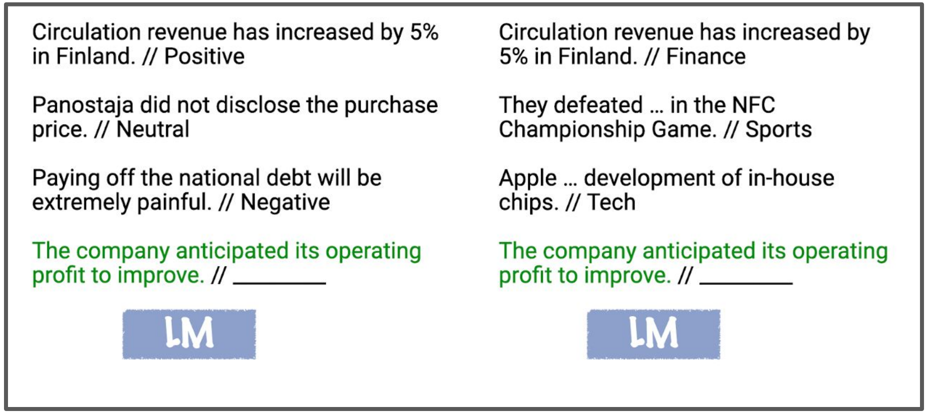
Temporal Knowledge Graph Forecasting Without Knowledge using In-context Learning.,
Dong-Ho Lee*, Kian Ahrabian*, Woojeong Jin, Fred Morstatter, Jay Pujara., EMNLP 2023

REALTALK: A 21-Day Real-World Dataset for Long-Term Conversation.,
Dong-Ho Lee*, Adyasha Maharana*, Jay Pujara, Xiang Ren, Francesco Barbieri., In submission to ACL 2025

Can LMs learn from **context** to solve the task ?



In-context Learning



ICL with task examples
Classification task with 1-shot example

Language models are few-shot learners., Brown et al., 2020
Rethinking the role of demonstrations: What makes in-context learning work?., Min et al., 2023

Beyond In-context Learning with Task Examples

Extrapolation

Circulation revenue has increased by 5% in Finland. // Positive

Panostaja did not disclose the purchase price. // Neutral

Paying off the national debt will be extremely painful. // Negative

The company anticipated its operating profit to improve. // ________

LM

Circulation revenue has increased by 5% in Finland. // Finance

They defeated ... in the NFC Championship Game. // Sports

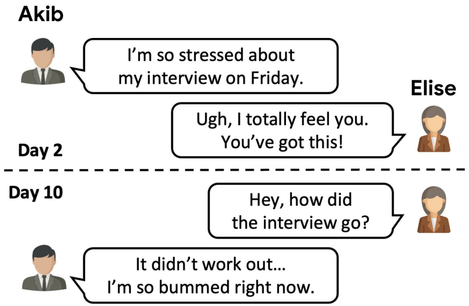
Apple ... development of in-house chips. // Tech

The company anticipated its operating profit to improve. // ________

LM

ICL with task examples
Classification task with 1-shot example

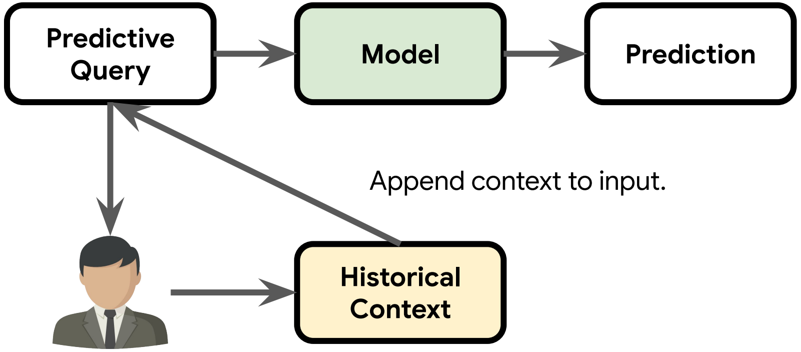
Super Bowl champion in 2000 is St Louis.
Super Bowl champion in 2001 is Baltimore.
Super Bowl champion in 2002 is New England.
Super Bowl champion in 2003 is Tampa Bay.
...
Super Bowl champion in 2020 is Kansas City.
Super Bowl champion in 2021 is Tampa Bay.
Super Bowl champion in 2022 is Los Angeles.
Super Bowl champion in 2023 is Kansas City.
Super Bowl champion in 2024 is ________



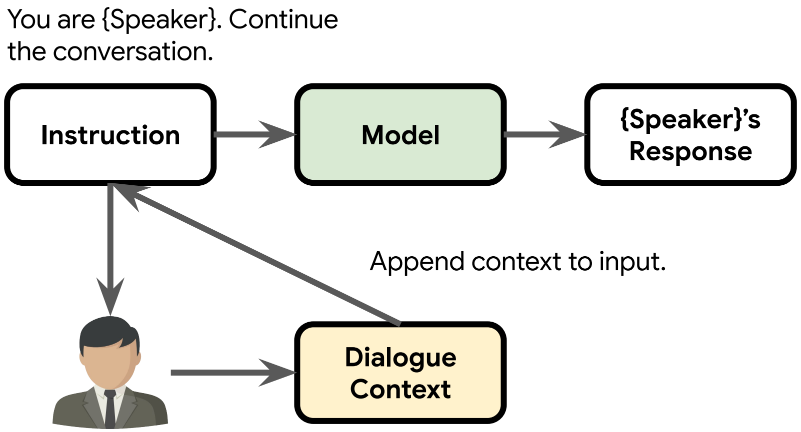
You are Elise. Continue the conversation.

Contextual Extrapolation

Preview: Inference with Context



Define helpful context.



Structural Extrapolation

Can LMs learn **patterns** from **historical context** to predict event ?

Semantic Extrapolation

Can LMs learn **user persona** from **dialogue context** to simulate user ?

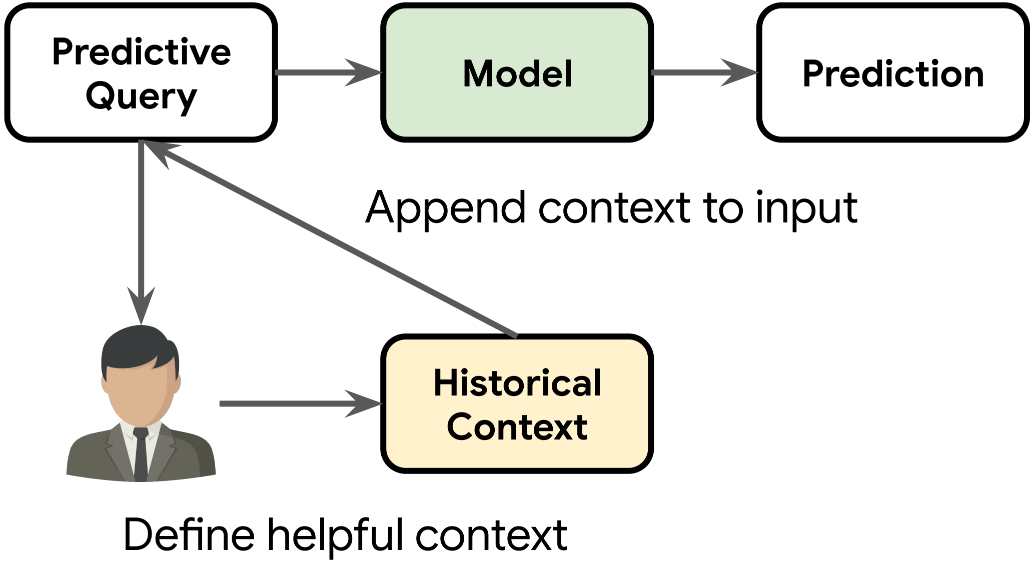
Temporal Knowledge Graph Forecasting Without Knowledge using In-context Learning., **Dong-Ho Lee***, Kian Ahrabian*, Woojeong Jin, **Fred Morstatter, Jay Pujara.**, EMNLP 2023

REALTALK: A 21-Day Real-World Dataset for Long-Term Conversation **Dong-Ho Lee**, Adyasha Maharana*, **Jay Pujara, Xiang Ren**, Francesco Barbieri., In submission to ACL 2025

RQ1-1. Can LMs learn patterns from historical context to predict event?

Temporal Knowledge Graph Forecasting Without Knowledge using In-context Learning.,
Dong-Ho Lee*, Kian Ahrabian*, Woojeong Jin, Fred Morstatter, Jay Pujara., EMNLP 2023

Can LMs learn patterns from historical context to predict event ?



Can LMs learn patterns from historical context to predict event ?

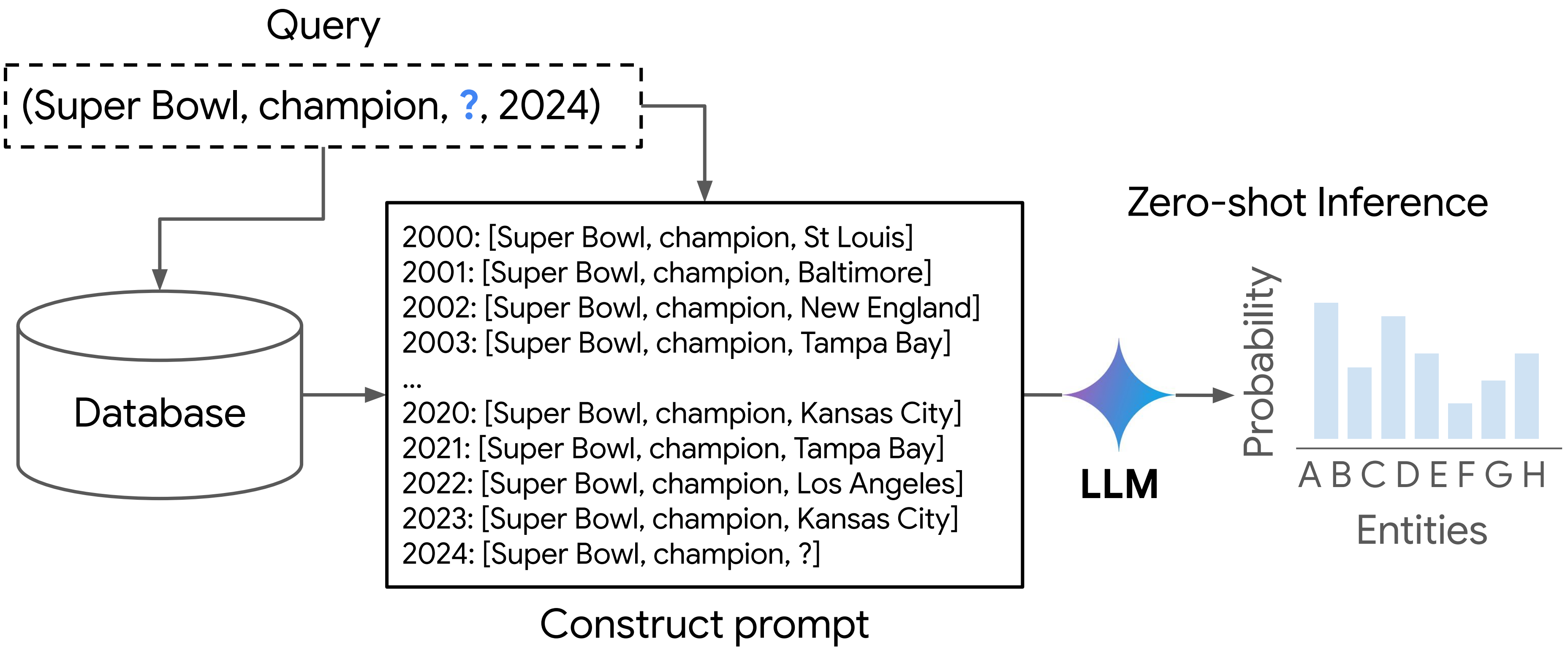
History

Super Bowl champion in 2000 is St Louis.
Super Bowl champion in 2001 is Baltimore.
Super Bowl champion in 2002 is New England.
Super Bowl champion in 2003 is Tampa Bay.
...
Super Bowl champion in 2020 is Kansas City.
Super Bowl champion in 2021 is Tampa Bay.
Super Bowl champion in 2022 is Los Angeles.
Super Bowl champion in 2023 is Kansas City.
Super Bowl champion in 2024 is ___________

Kansas City...?

Gemini

Methodology



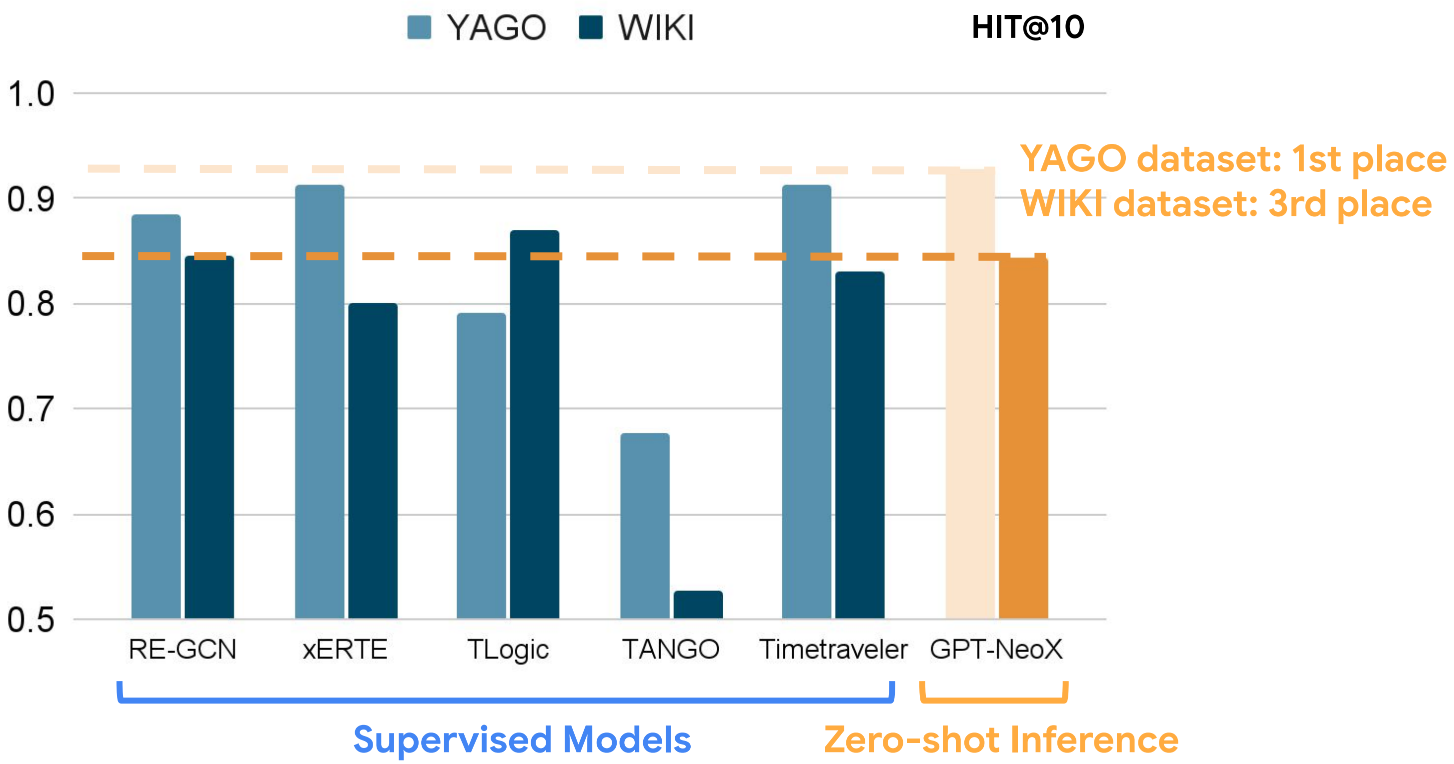
LMs learn patterns in the context.

Because:

- **Finding 1.** LMs can outperform supervised models without training by using context.
- **Finding 2.** LMs use context effectively without relying on semantic priors.
- **Finding 3.** LMs learn diverse patterns from the context beyond just frequency and recency.

Finding 1

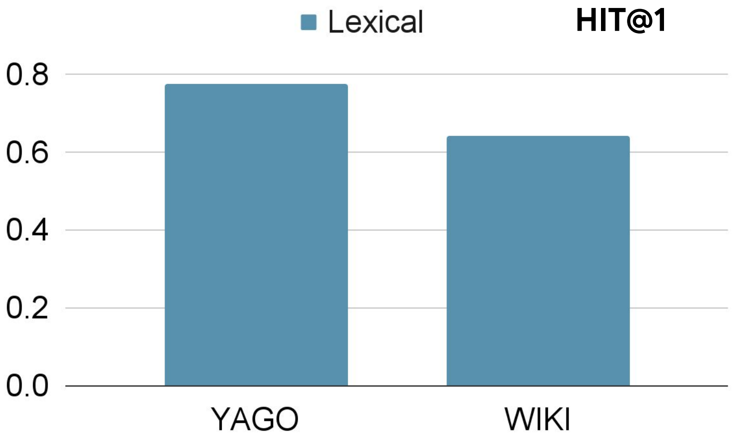
LMs can outperform supervised models without training.



Finding 2

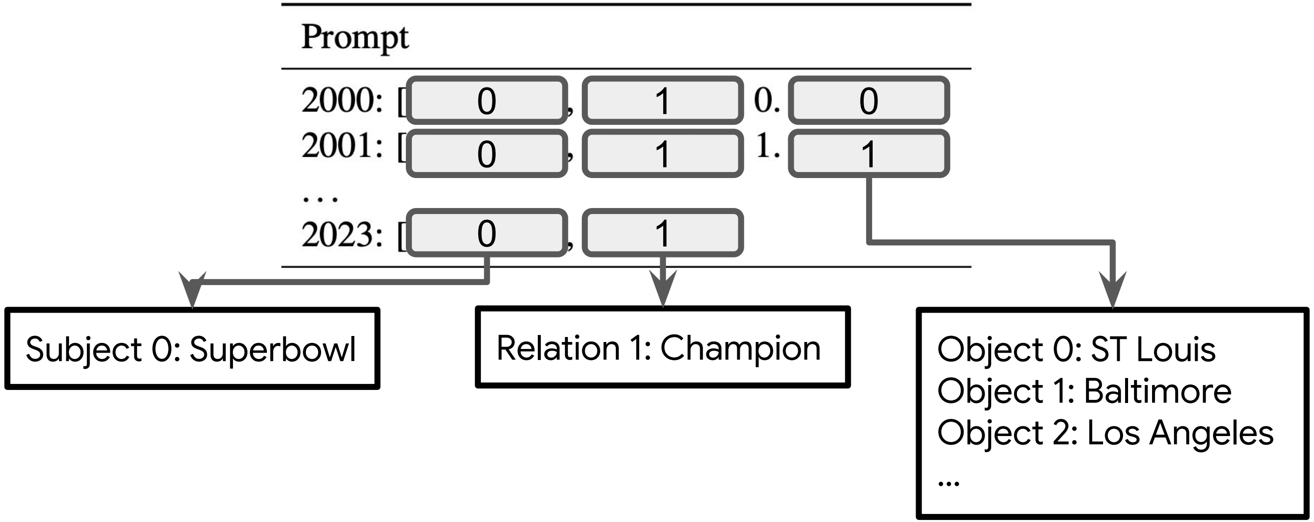
LMs use context effectively without relying on semantic priors.

Category	Prompt
Lexical $\mathcal{L}(\cdot)$	2000: [Superbowl, Champion, 0. St Louis] 2001: [Superbowl, Champion, 1. Baltimore] … 2023: [Superbowl, Champion,



Finding 2

LMs use context effectively without relying on semantic priors.



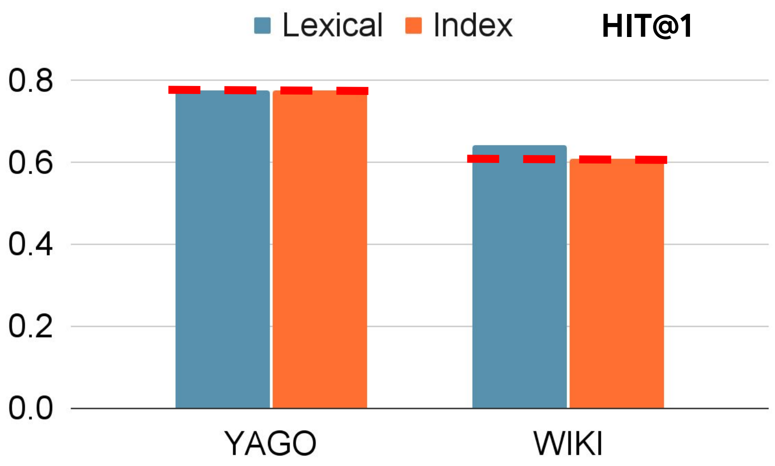
What if we **mask certain words that have semantic content** (e.g., names, event titles, relation labels) and replace them with arbitrary indices that have no inherent meaning?

Finding 2

LMs use context effectively without relying on semantic priors.

Category	Prompt
Lexical $\mathcal{L}(\cdot)$	2000: [Superbowl, Champion, 0. St Louis] 2001: [Superbowl, Champion, 1. Baltimore] . . . 2023: [Superbowl, Champion,
Index $\mathcal{I}(\cdot)$	2000: [0, 0, 0. 0] 2001: [0, 0, 1. 1] . . . 2023: [0, 0,

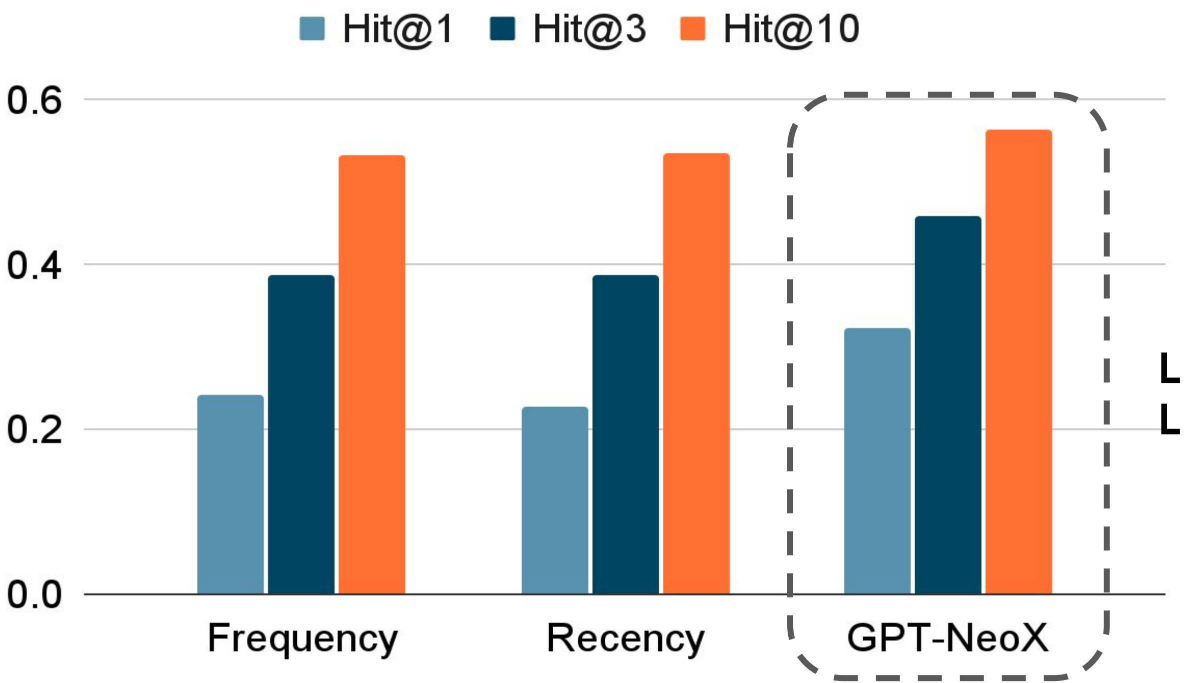
Remove semantic information



Results for “**Lexical**” and “**Index**” are similar.

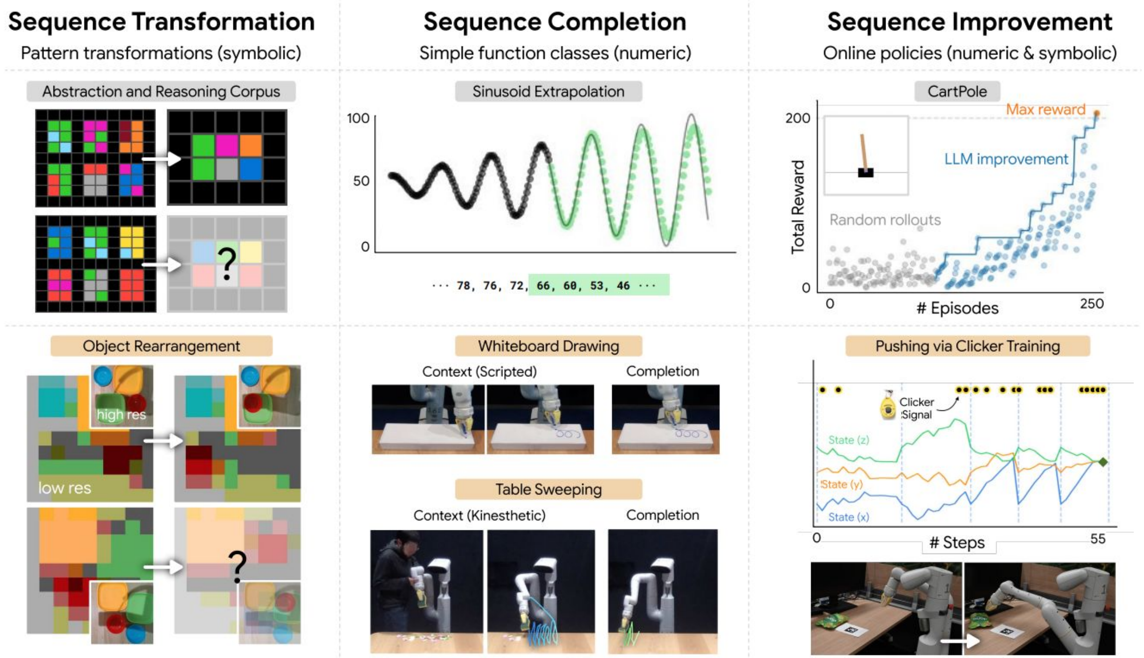
Finding 3

LMs learn diverse contextual patterns beyond just frequency and recency.



LM > Frequency
LM > Recency

Similar Findings (Structural Extrapolation) from Other Paper



Structural Extrapolation

Model identifies and extends patterns in sequential input without necessarily understanding their semantic meaning.

LLMs can serve as general sequence modelers, driven by in-context learning.

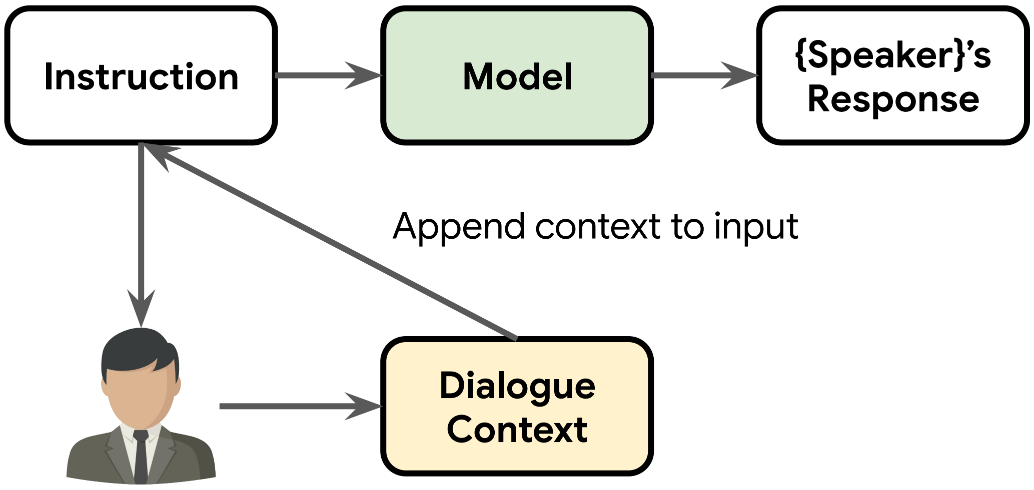
LLM as General Pattern Machines., Mirchandani et al., 2023

RQ1-2. Can LMs learn user persona from dialogue context to simulate user?

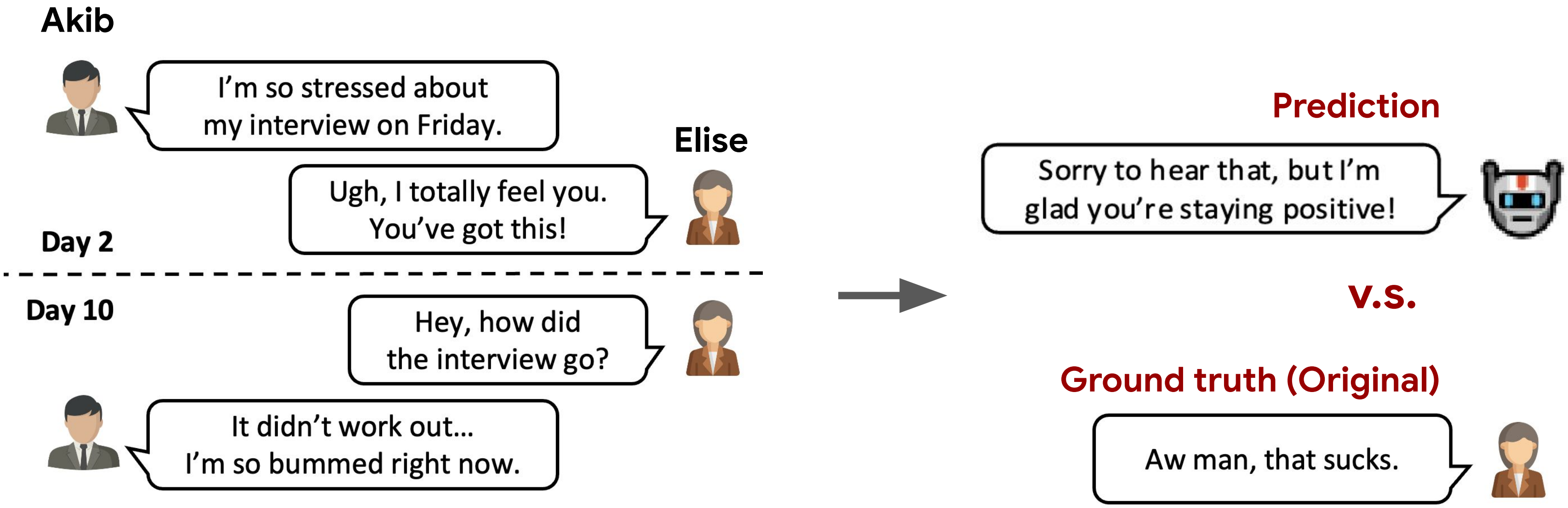
REALTALK: A 21-Day Real-World Dataset for Long-Term Conversation
Dong-Ho Lee, Adyasha Maharana, Jay Pujara, Xiang Ren, Francesco Barbieri., In submission to ACL 2025

Can LMs learn user persona from dialogue context to simulate user ?

You are {Speaker}. Continue the conversation.

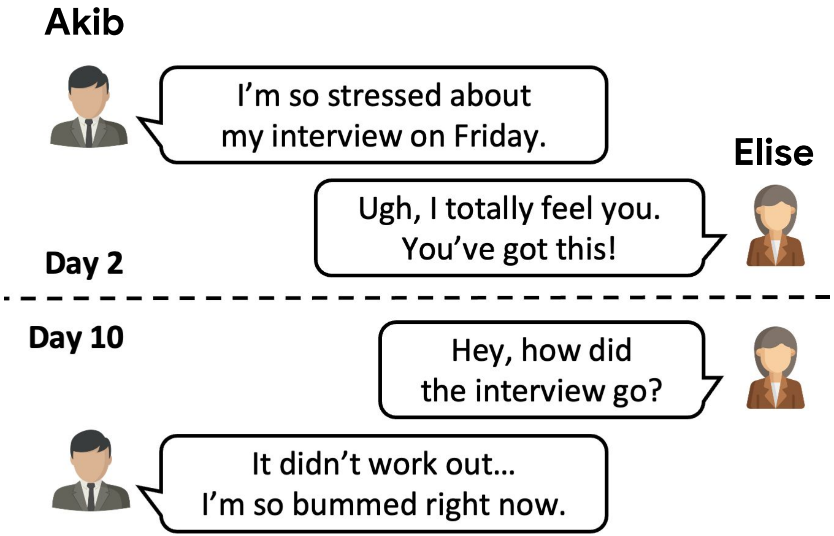


Can LMs learn user persona from dialogue context to simulate user ?



You are Elise. Continue the conversation.

Can LMs learn user persona from dialogue context to simulate user ?



We need a long conversation where the persona remains consistent throughout.

Longitudinal Dialogues in the Wild

A 21-day Real-World Dataset for Long-Term Conversation

Dataset	Dialogue Participants	# Turns / $\mathcal{C}$	# Session / $\mathcal{C}$	# Tokens / $\mathcal{C}$	Multimodal	Collection
MemoryBank (Zhong et al., 2024)	Human-AI	3.7	10	257.8	✗	LLM-simulated
LongMemEval (Wu et al., 2024)	Human-AI	9.8	50.2	1,572.3	✗	LLM-simulated
SODA (Kim et al., 2023a)	Human-Human	7.6	1	122.4	✗	LLM-simulated
Conversation Chronicles (Jang et al., 2023b)	Human-Human	58.5	5	1,054.7	✗	LLM-simulated
LoCoMo (Maharana et al., 2024)	Human-Human	588.2	27.2	13,377.2	✓	LLM-simulated
MPChat (Ahn et al., 2023)	Human-Human	2.8	1	53.3	✓	Reddit
MMDialog (Feng et al., 2022)	Human-Human	4.6	1	72.5	✓	Social media
Daily Dialog (Li et al., 2017)	Human-Human	7.9	1	114.7	✗	Crowdsourcing
MSC (Xu et al., 2022)	Human-Human	53.3	4	1,225.9	✗	Crowdsourcing
REALTALK	Human-Human	894.4	21.9	17,109.8	✓	Crowdsourcing

The longest crowd-sourced human-human conversation

Evaluating Very Long-Term Conversational Memory of LLM Agents.,
Adyasha Maharana, **Dong-Ho Lee**, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, Yuwei Fang., ACL 2024

REALTALK: A 21-Day Real-World Dataset for Long-Term Conversation
Dong-Ho Lee, Adyasha Maharana, Jay Pujara, Xiang Ren, Francesco Barbieri., In submission to ACL 2025

Longitudinal Dialogues in the Wild

A 21-day Real-World Dataset for Long-Term Conversation

Name	Job	City	Age
Emi	College student	New York	20
Elise	College student	Houston	21
Kevin	College student	Houston	18
Paola	College student	Cambridge	21
Nebraas	Vet technician	New York	24
Nicolas	College student	New York	23
Vanessa	Vet technician	New York	24
Mohammed	College student	New York	23
Syed	College student	New York	23
Fahim	College student	New York	19

Chat	# days	# words	# words / day	Images	Topics
Emi + Elise	21	17341	826	35	6
Elise + Kevin	21	16040	764	46	7
Kevin + Paola	16	11057	691	35	7
Paola + Emi	16	11511	719	37	6
Nicolas + Nebraas	21	16902	805	37	7
Vanessa + Nicolas	21	18005	857	39	8
Vanessa + Nebraas	21	16343	778	43	8
Akib + Muhammed	21	17191	819	25	6
Fahim + Akib	21	18089	861	27	7
Fahim + Muhammed	21	16951	807	23	8

- Two people talk each other for 21 days through **Whatsapp**.
- Each participant involves in two conversations.

Research Question

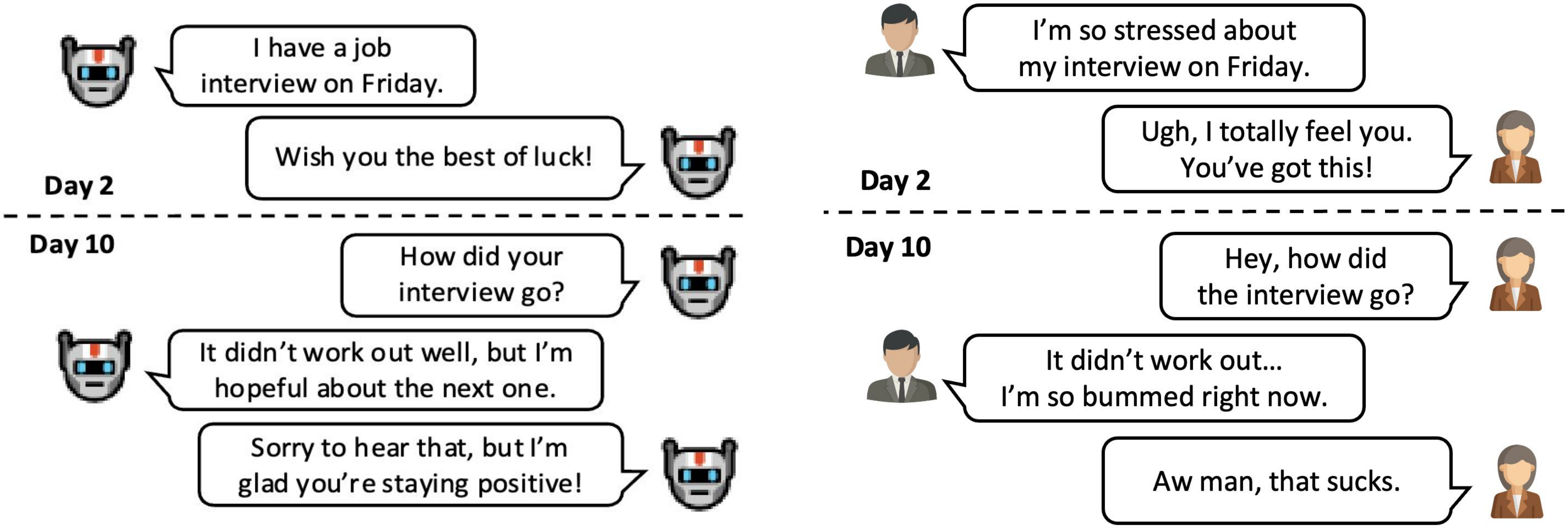
Data Analysis

1. How do authentic human dialogues differ from LLM-simulated ones?
2. Does an individual maintain a consistent persona when talking to other people?

Persona Simulation (Semantic Extrapolation)

3. Can models accurately simulate an individual's unique persona?

How do authentic human dialogues differ from LLM-simulated ones?



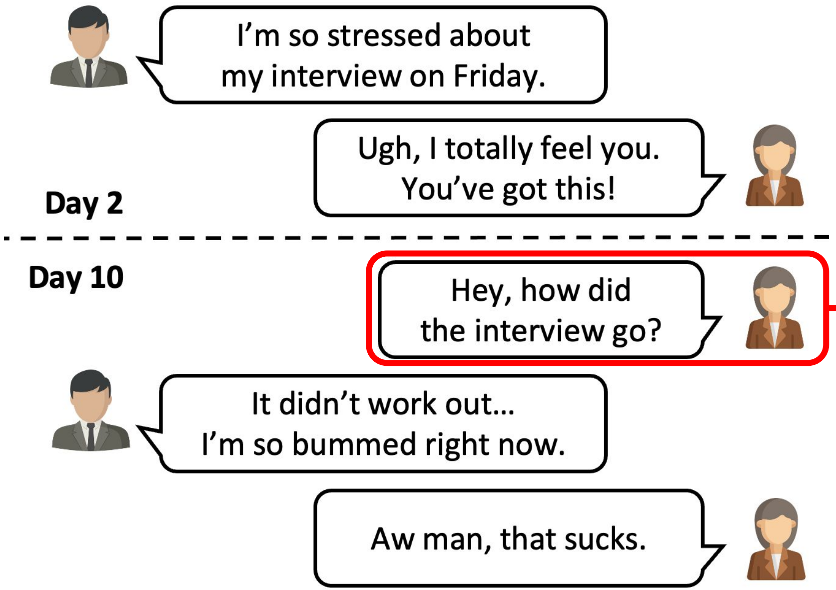
LLM-Simulated Dialogue (LoCoMo) | Real-world Dialogue (REALTALK)

Evaluating Very Long-Term Conversational Memory of LLM Agents.,
Adyasha Maharana, **Dong-Ho Lee**, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, Yuwei Fang., ACL 2024

REALTALK: A 21-Day Real-World Dataset for Long-Term Conversation
Dong-Ho Lee*, Adyasha Maharana*, **Jay Pujara, Xiang Ren**, Francesco Barbieri., In submission to ACL 2025

How do authentic human dialogues differ from LLM-simulated ones?

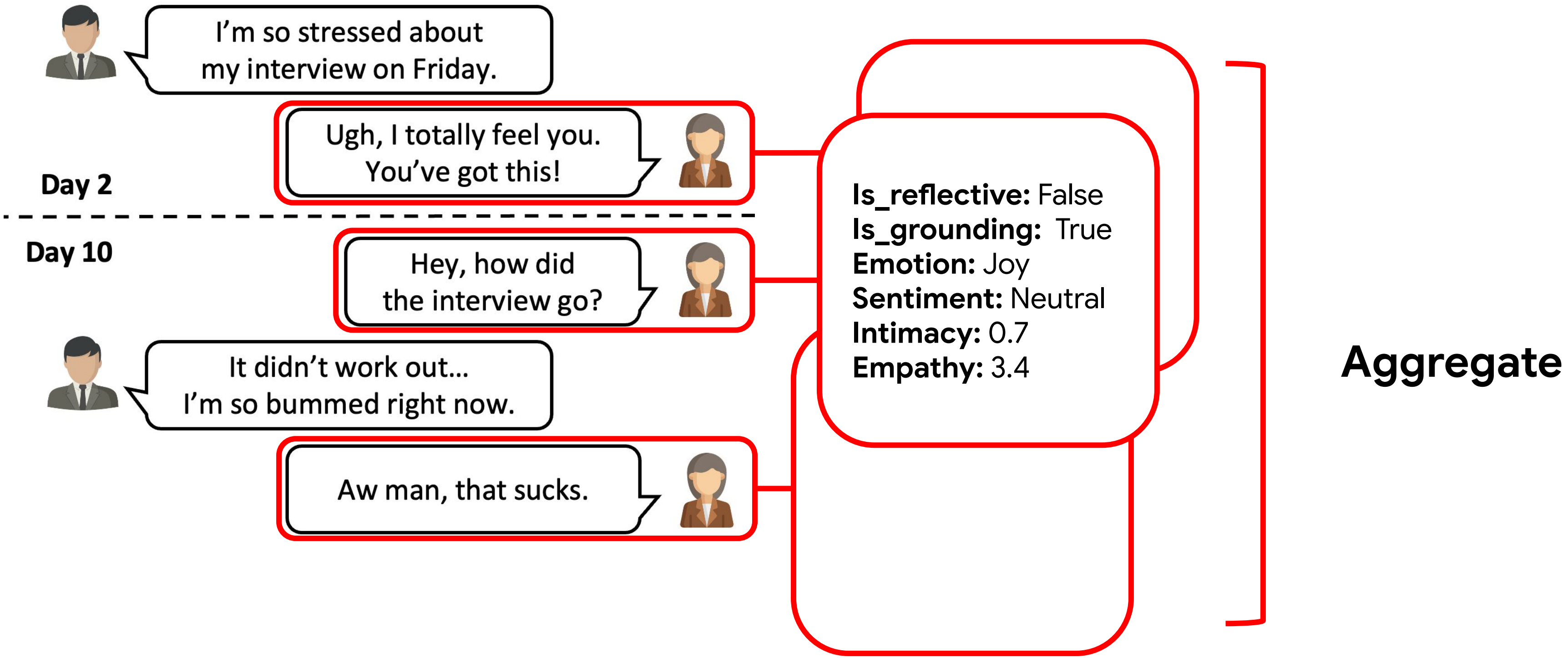
Message-level Emotional Intelligence



Is_reflective: False	{True, False}
Is_grounding: True	{True, False}
Emotion: Joy	{Joy, Anger, ...} → 11 labels
Sentiment: Neutral	{Positive, Neutral, Negative}
Intimacy: 0.7	0~1
Empathy: 3.4	0~6

How do authentic human dialogues differ from LLM-simulated ones?

Speaker-level Emotional Intelligence



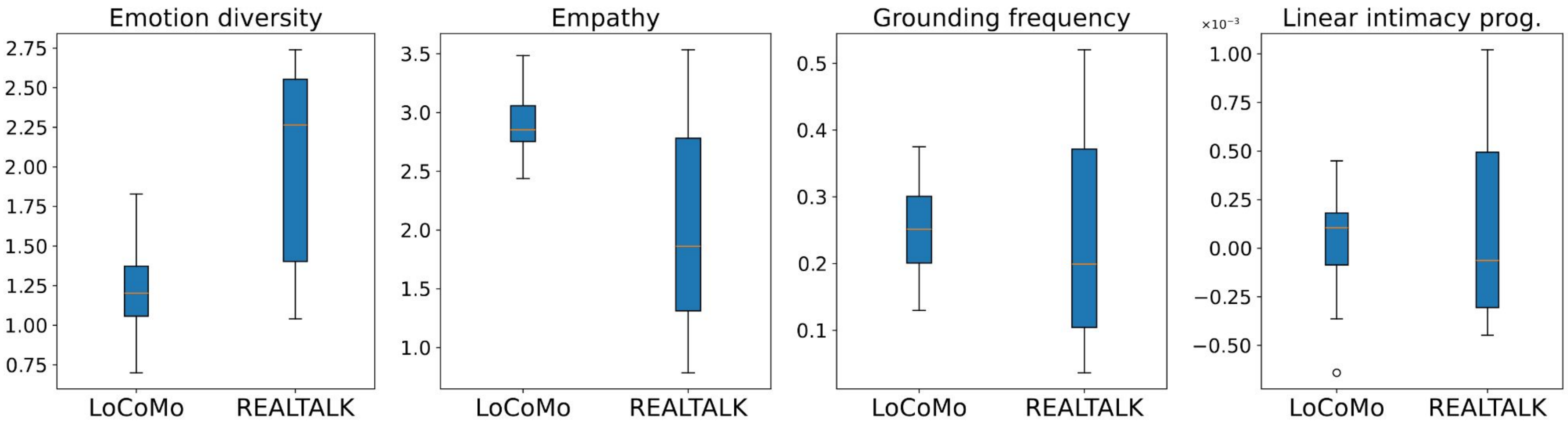
How do authentic human dialogues differ from LLM-simulated ones?

Speaker-level Emotional Intelligence

- **Reflective frequency**: Average
  - How often the speaker uses reflective language.
- **Grounding frequency**: Average
  - How often the speaker uses clarifying or follow-up questions.
- **Empathy**: Average
  - Average empathy score across all the speaker's messages.
- **Emotion / Sentiment diversity**: Entropy
  - The range of emotions or sentiments expressed in the speaker's messages.
- **Intimacy Progression**: Curve Inclination
  - How intimacy develops over time.

How do authentic human dialogues differ from LLM-simulated ones?

Speaker-level Emotional Intelligence

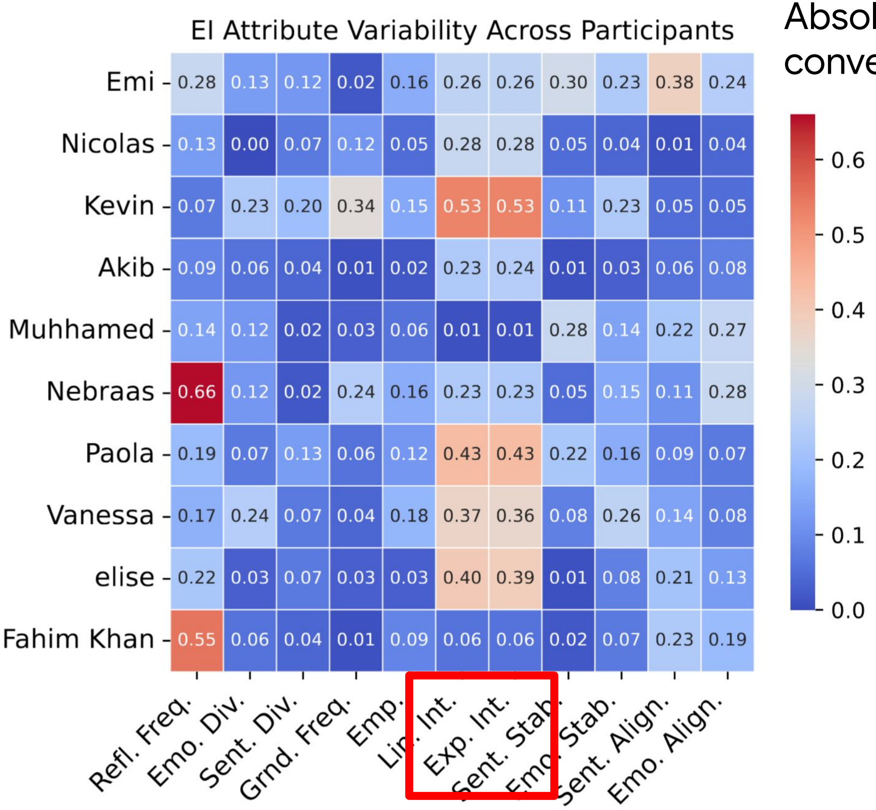


- Humans exhibit greater emotion and sentiment diversity.
- LLMs show excessive empathy.
- Humans show high variance in overall EI attributes, while LLMs are uniform.
    - (Align with other work (Lee et al., 2024))

Language Models show stable value orientations across diverse role-plays., Bruce Lee et al., 2024

Does an individual maintain a consistent persona across multiple chats?

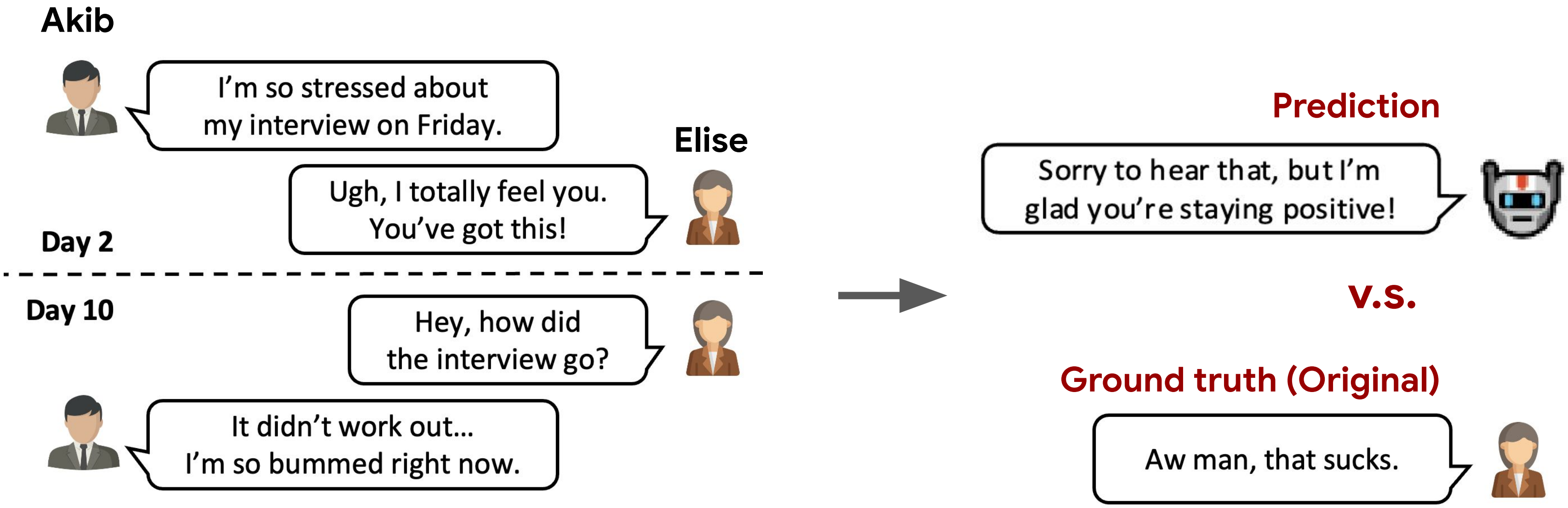
Speaker-level Emotional Intelligence



Absolute difference between two conversations from the same speaker

- Some participants maintain stable persona while others not.
- Intimacy progression shows the most variation → It is highly influenced by who they talk with.

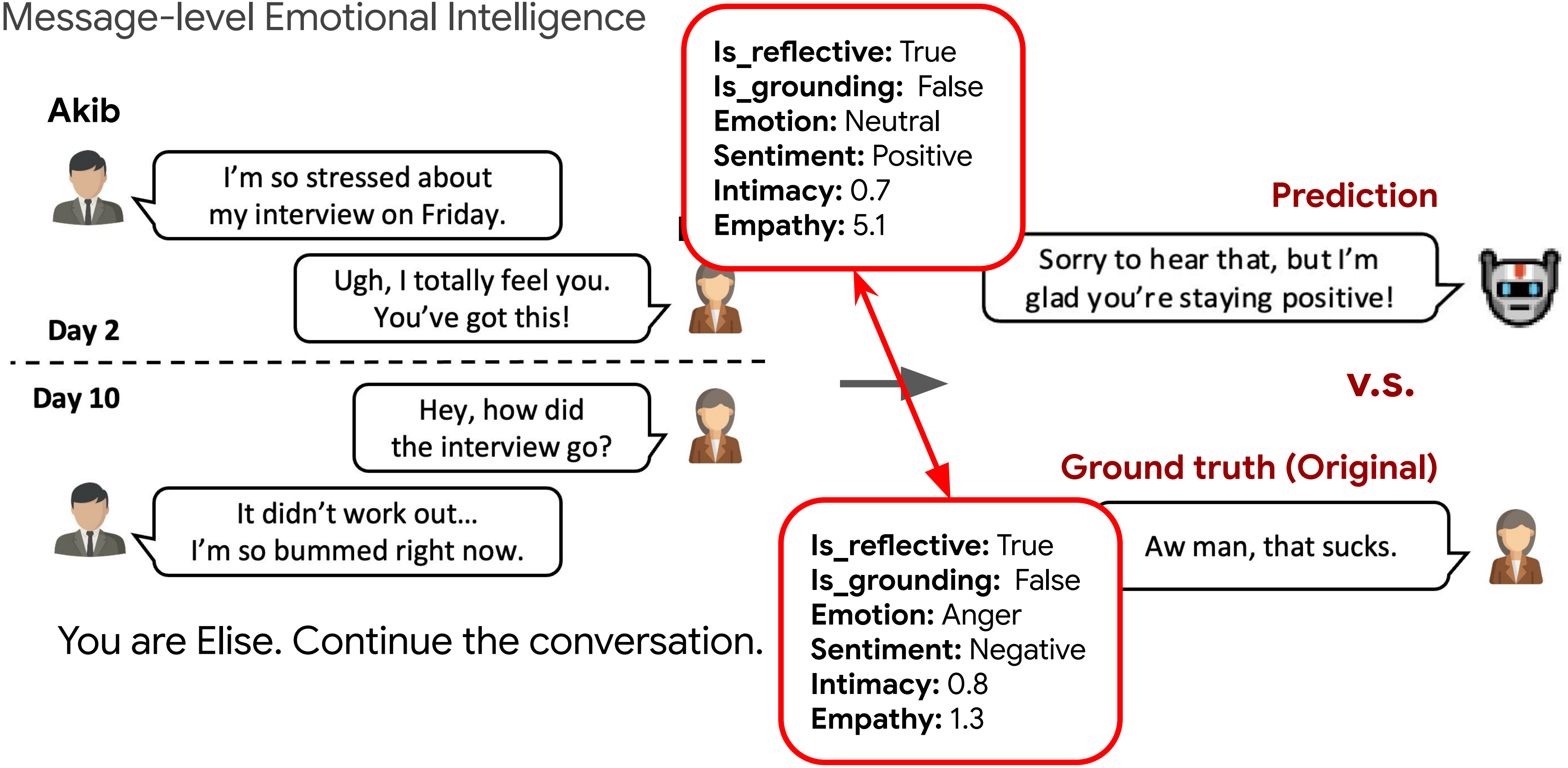
Can models accurately simulate an individual's unique persona?



You are Elise. Continue the conversation.

Can models accurately simulate an individual's unique persona?

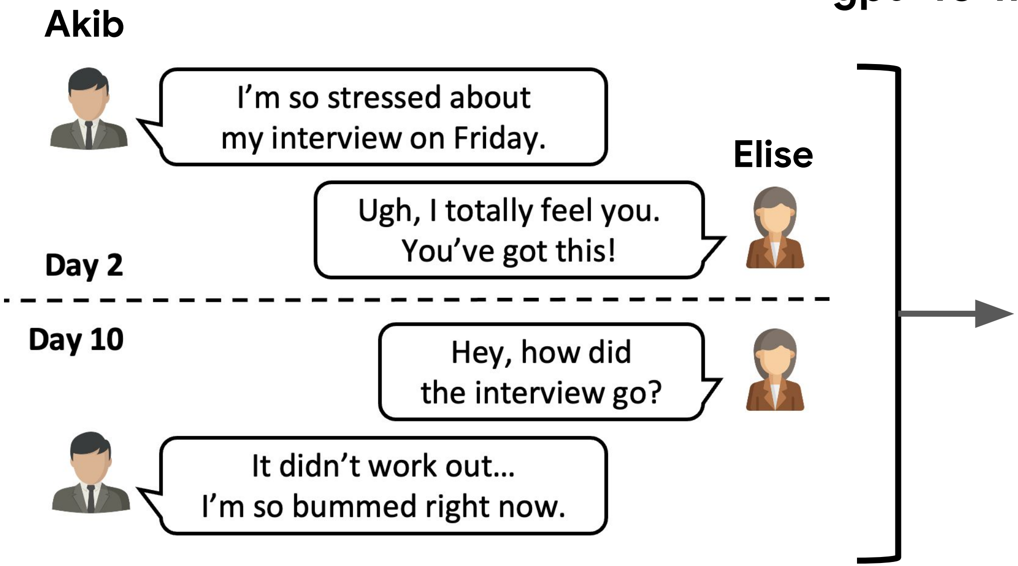
Message-level Emotional Intelligence



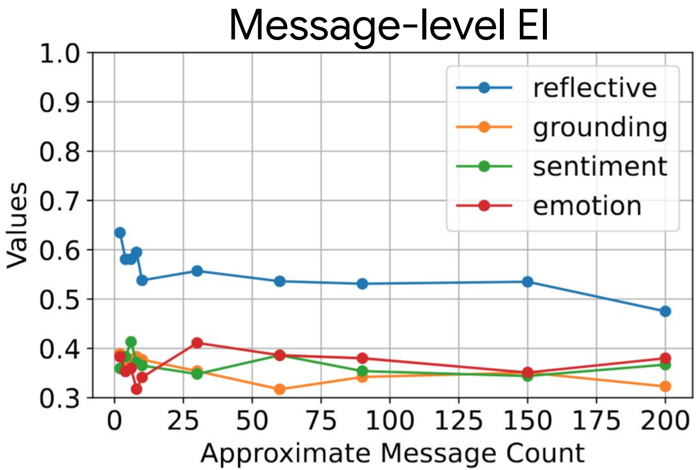
Can models accurately simulate an individual's unique persona?

In-context Learning

gpt-4o-mini



You are Elise. Continue the conversation.



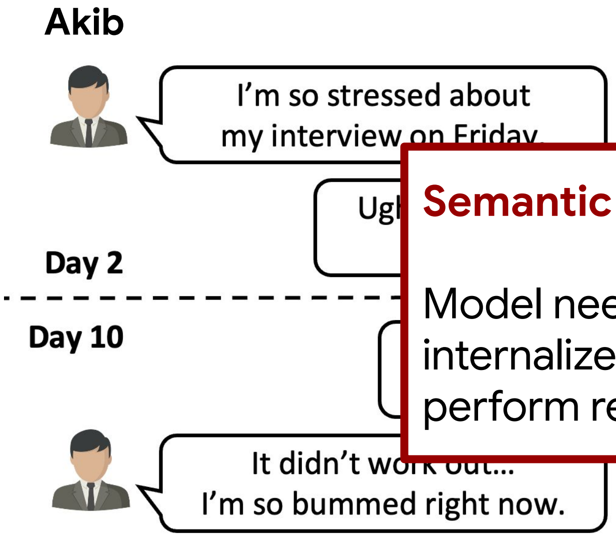
Length of conversation history

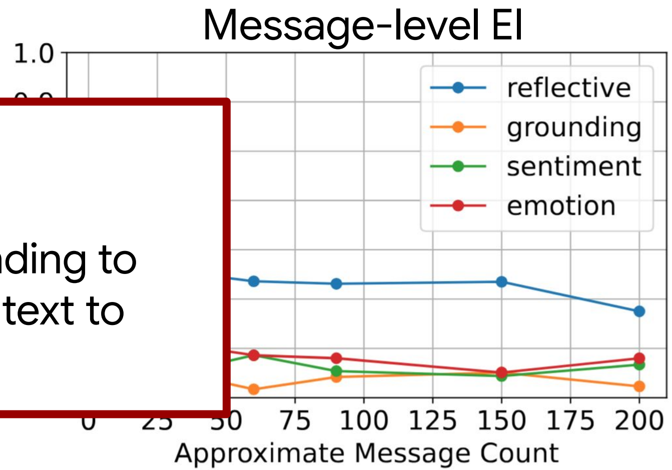
Add more context → Cannot improve the simulation performance.

Can models accurately simulate an individual's unique persona?

In-context Learning

gpt-4o-mini





Length of conversation history

You are Elise. Continue the conversation.

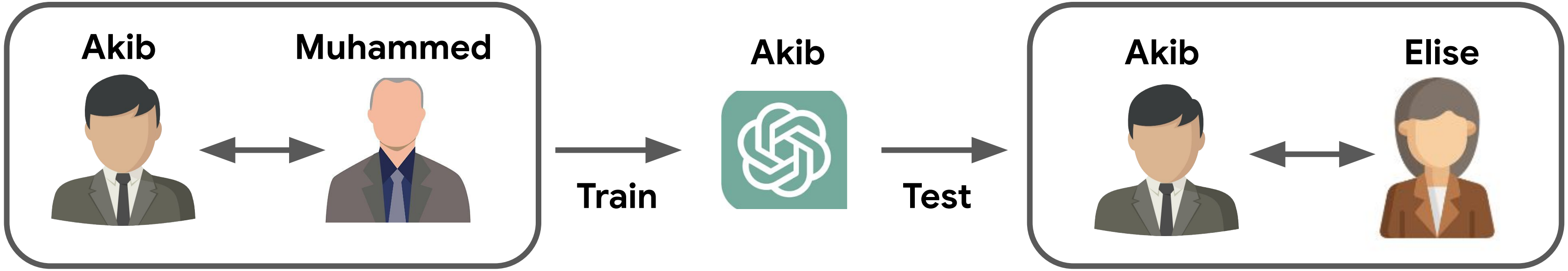
Semantic Extrapolation

Model needs semantic understanding to internalize and interpret prior context to perform reasoning or simulation.

Model may struggle with semantic extrapolation.

Can models accurately simulate an individual's unique persona?

Bonus: Fine-tune



What if we train Akib's messages exclusively?

Can models accurately simulate an individual's unique persona?

Bonus: Fine-tune

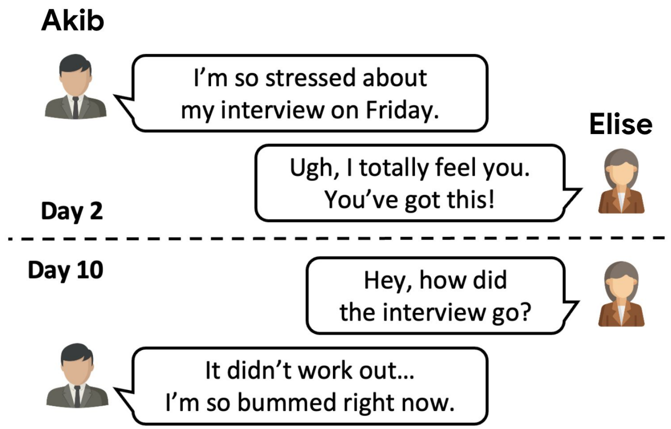
The mean performance across 10 distinct persona fine-tuning simulations (OpenAI Fine-tuning API w/ gpt-4o-mini)

	Content Similarity		Message-level EI (Emotional Intelligence)					
	Lexical ↑	Semantic ↑	Reflective ↑	Grounding ↑	Sentiment ↑	Emotion ↑	Intimacy ↓	Empathy ↓
w/o fine-tune	0.136	0.76	0.623	0.396	0.529	0.427	**0.061**	1.8
w/ fine-tune	**0.138**	**0.778**	**0.769**	**0.619**	**0.587**	**0.464**	0.073	**1.238**

- Fine-tuning captures speaker's style.
  - emotion, sentiment, empathy, reflective, grounding
- Fine-tuning cannot improve content similarity.
- Fine-tuning does not improve intimacy, as intimacy depends on who they talk with.

Conclusion: Inference with Context

Super Bowl champion in 2000 is St Louis.
Super Bowl champion in 2001 is Baltimore.
Super Bowl champion in 2002 is New England.
Super Bowl champion in 2003 is Tampa Bay.
...
Super Bowl champion in 2020 is Kansas City.
Super Bowl champion in 2021 is Tampa Bay.
Super Bowl champion in 2022 is Los Angeles.
Super Bowl champion in 2023 is Kansas City.
Super Bowl champion in 2024 is __________



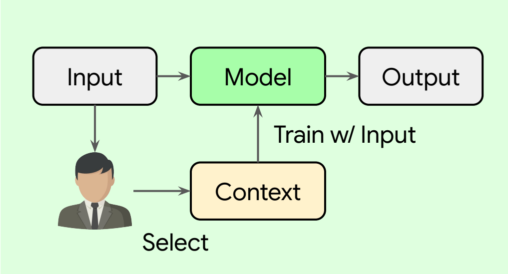
You are Elise. Continue the conversation.

Yes, LMs learn **patterns** from **historical context.**

No, LMs do not learn **user persona** from **dialogue context.**

Evaluating Very Long-Term Conversational Memory of LLM Agents.,
Adyasha Maharana, **Dong-Ho Lee**, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, Yuwei Fang., ACL 2024

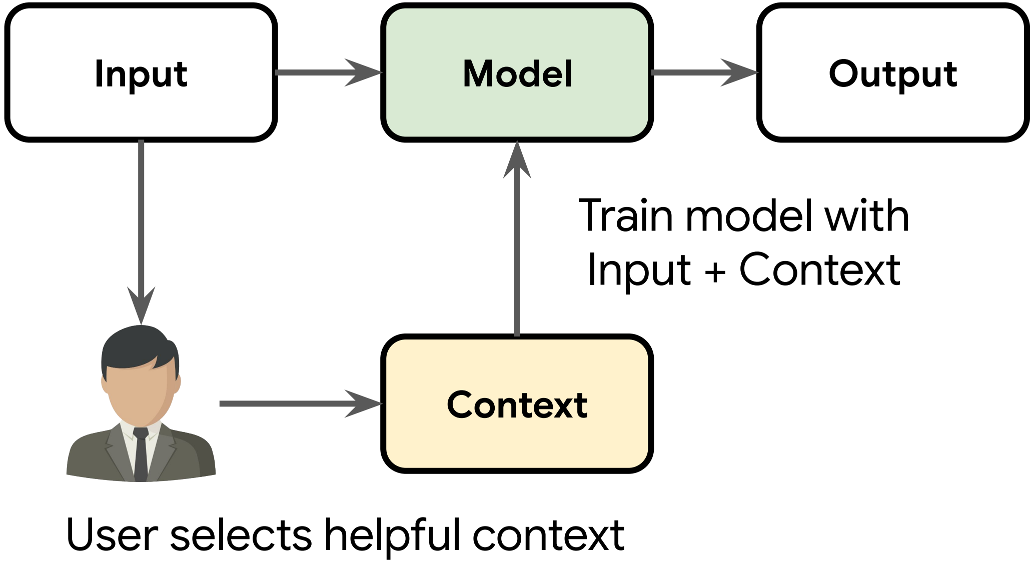
REALTALK: A 21-Day Real-World Dataset for Long-Term Conversation
Dong-Ho Lee, Adyasha Maharana, Jay Pujara, Xiang Ren, Francesco Barbieri., In submission to ACL 2025



RQ2. Does training models with context improve their performance?

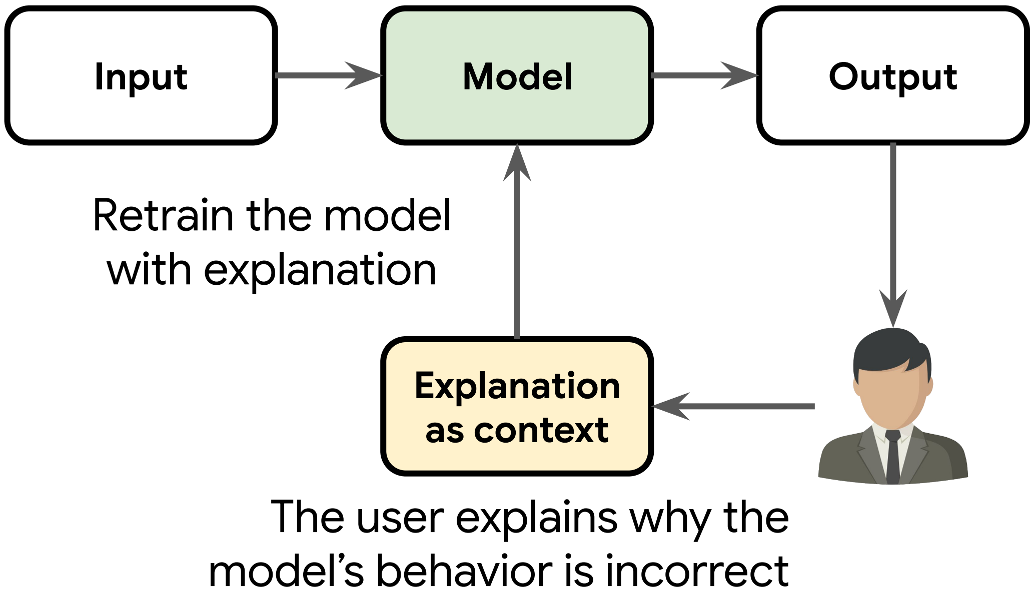
XMD: An End-to-End Framework for Interactive Explanation-Based Debugging of NLP Models
Dong-Ho Lee, Akshen Kadakia, Brihi Joshi, Aaron Chan, Ziyi Liu, Kiran Narahari, Takashi Shibuya, Ryosuke Mitani, Toshiyuki Sekiya, **Jay Pujara, Xiang Ren.**, ACL 2023 Demo

Does **training models with context** improve their performance?



Preview: Train LMs with Context

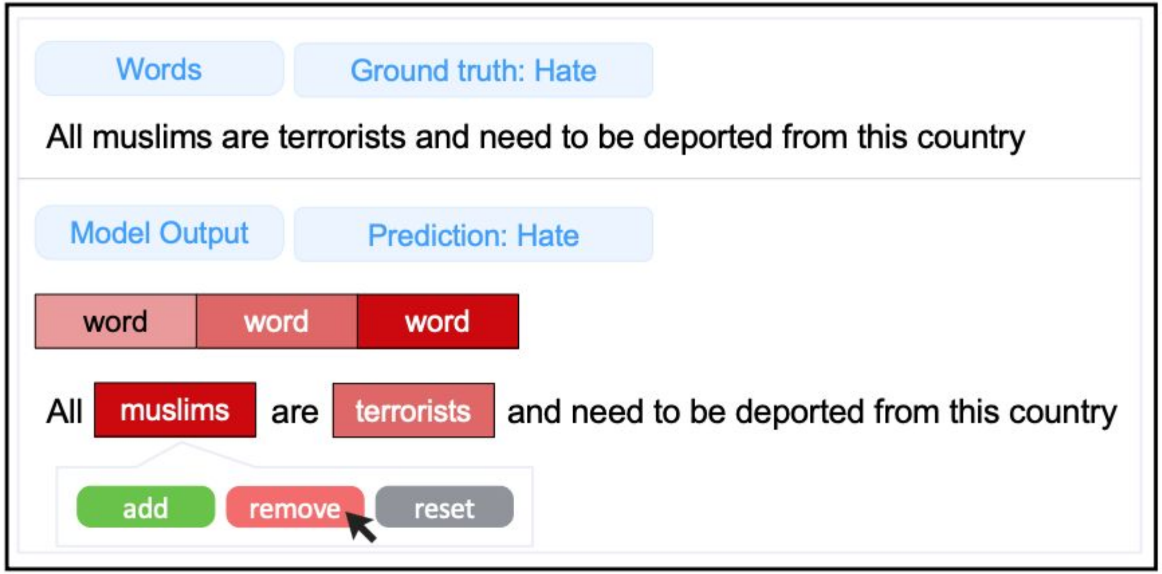
Does human explanation help debugging the model?



XMD: An End-to-End Framework for Interactive Explanation-Based Debugging of NLP Models
Dong-Ho Lee, Akshen Kadakia, Brihi Joshi, Aaron Chan, Ziyi Liu, Kiran Narahari, Takashi Shibuya, Ryosuke Mitani, Toshiyuki Sekiya, **Jay Pujara, Xiang Ren.**, ACL 2023 Demo

*Slide contains harmful contents (Example is from the existing dataset.)

Preview: Train LMs with Context

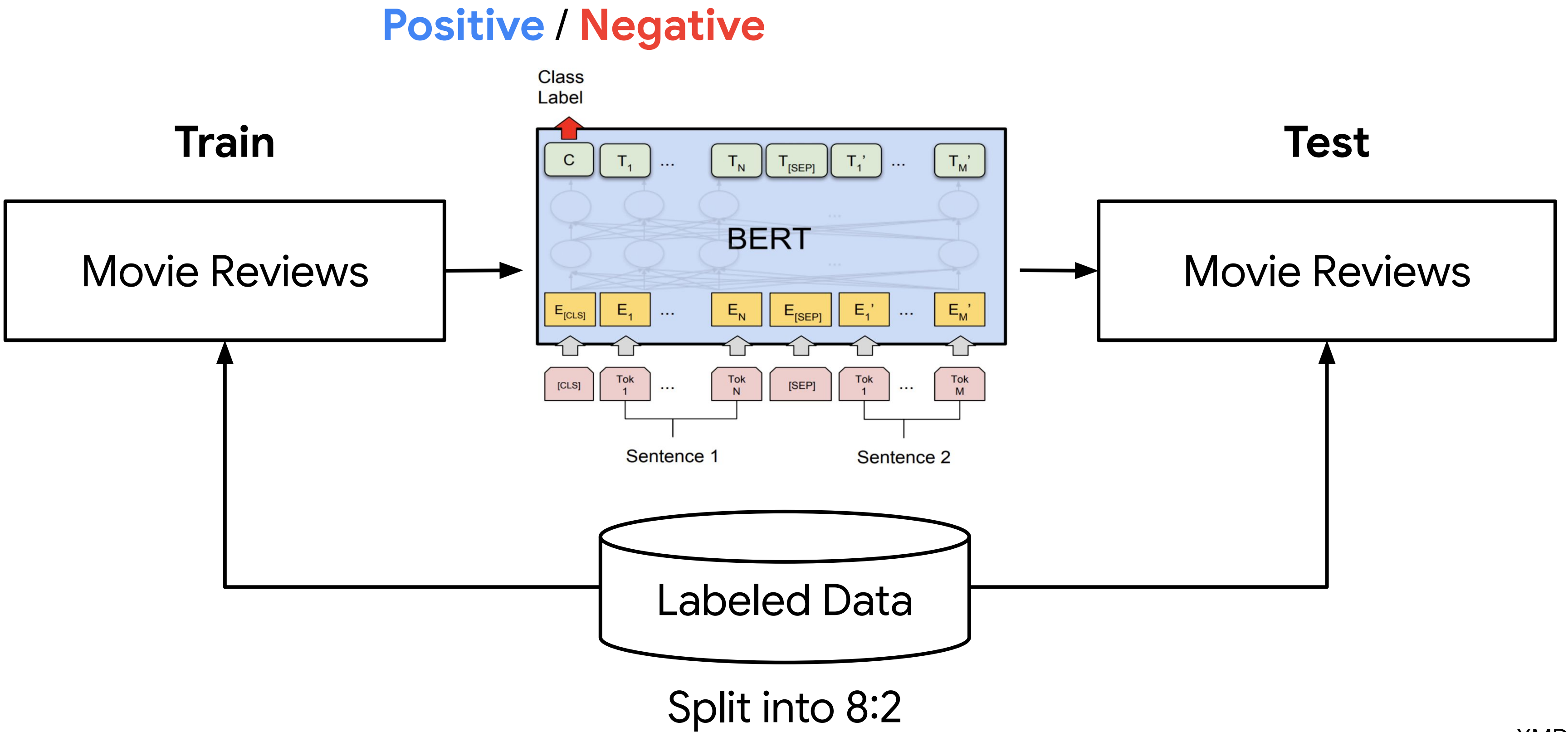


Human explanation helps **debugging & improving** the model.

XMD: An End-to-End Framework for Interactive Explanation-Based Debugging of NLP Models
Dong-Ho Lee, Akshen Kadakia, Brihi Joshi, Aaron Chan, Ziyi Liu, Kiran Narahari, Takashi Shibuya, Ryosuke Mitani, Toshiyuki Sekiya, **Jay Pujara, Xiang Ren.**, ACL 2023 Demo

LMs perform well on ID Test set

ID: Identically Distributed



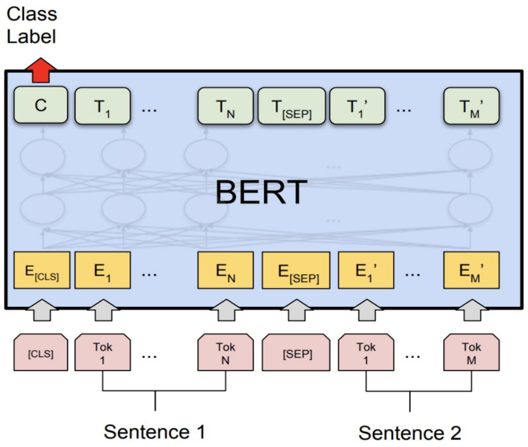
LMs perform well on ID Test set

ID: Identically Distributed

Positive / Negative

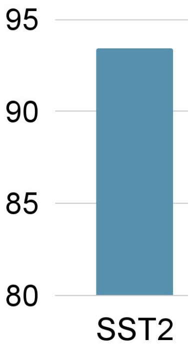
SST2 Train

Movie Reviews (Rotten Tomatoes)



SST2 Test

Movie Reviews (Rotten Tomatoes)



~ 95% accuracy

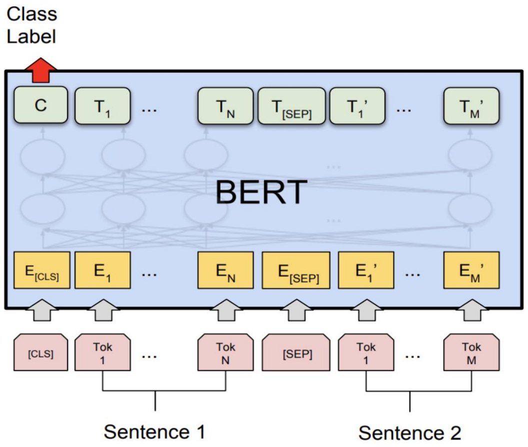
LMs perform well on OOD Test set?

OOD: Out-of-Distribution

Positive / Negative

SST2 Train

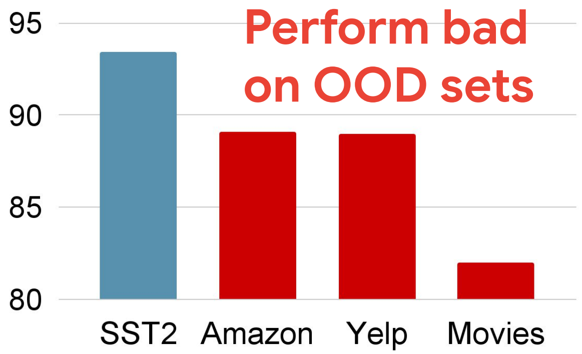
Movie Reviews
(Rotten Tomatoes)



Amazon Reviews

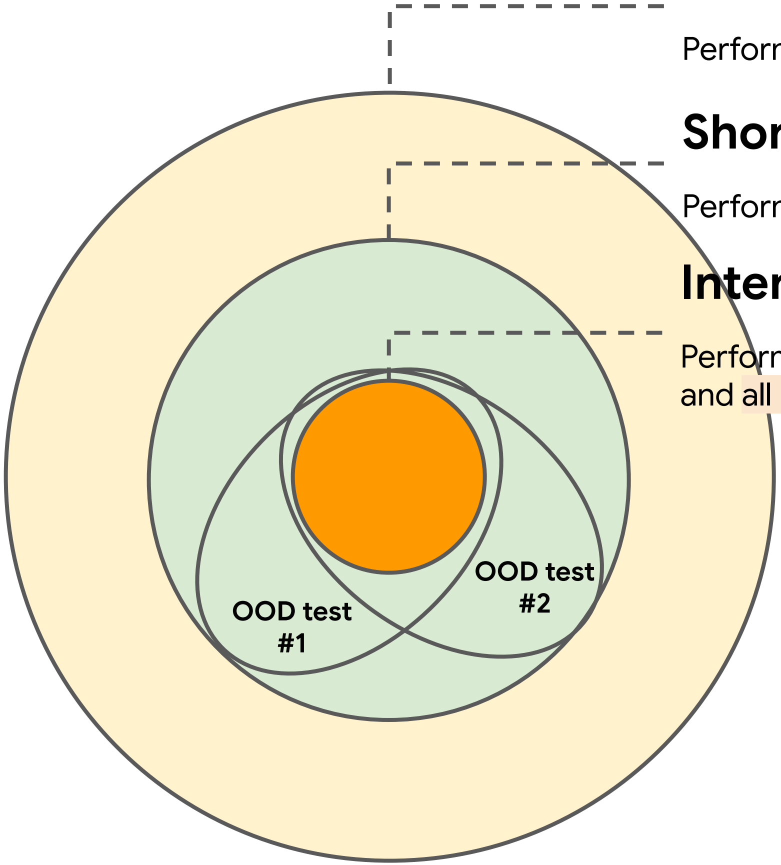
Yelp Reviews

Movie Reviews
(IMDB)



Bias in NLP Model

Shortcut Learning



Training Solutions

Perform well on training set.

Shortcut Solutions

Perform well on training set & i.d. test set

Intended Solutions

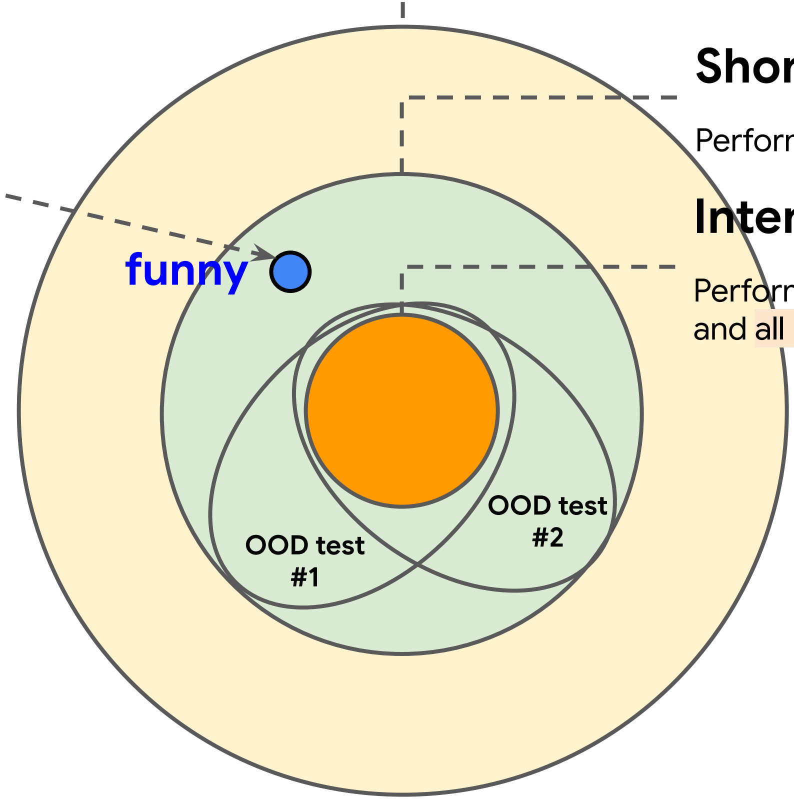
Perform well on training set & i.d. test set and all relevant test sets.

Shortcut Learning in Deep Neural Networks., Geirhos et al., 2020

Bias in NLP Model

Shortcut Learning

Rich veins of **funny** stuff in this movie! (**Positive**)
Is pretty **funny**. (**Positive**)
Very **funny** film (**Positive**)



Training Solutions

Perform well on training set.

Shortcut Solutions

Perform well on training set & i.d test set

Intended Solutions

Perform well on training set & i.d test set and all relevant test sets.

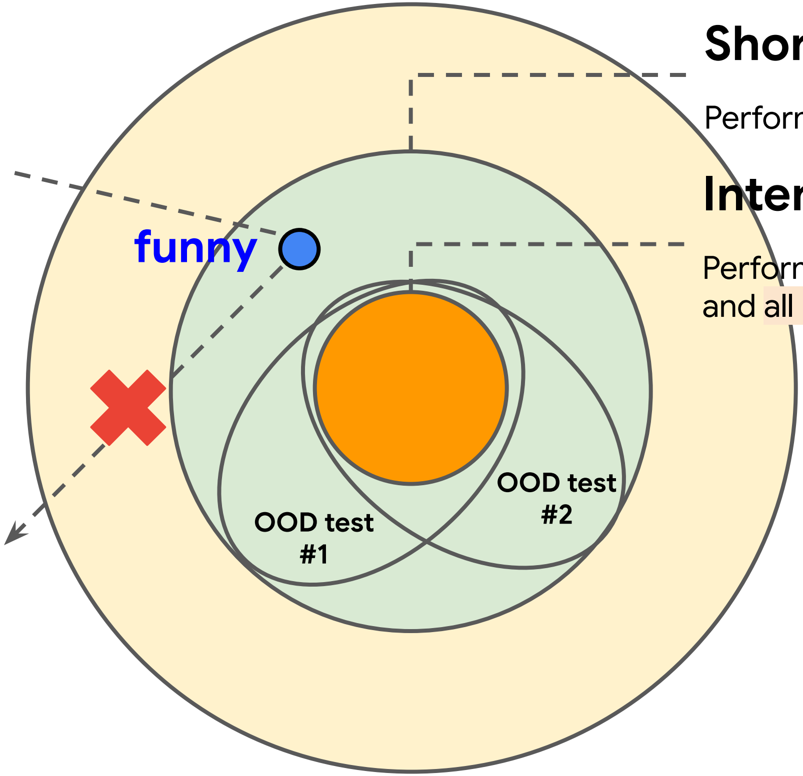
Shortcut Learning in Deep Neural Networks., Geirhos et al., 2020

Bias in NLP Model

Shortcut Learning

Rich veins of **funny** stuff in this movie! (**Positive**)
Is pretty **funny**. (**Positive**)
Very **funny** film (**Positive**)

$40 million failure with failed attempts at being **funny** (**Negative**)



Training Solutions

Perform well on training set.

Shortcut Solutions

Perform well on training set & i.d test set

Intended Solutions

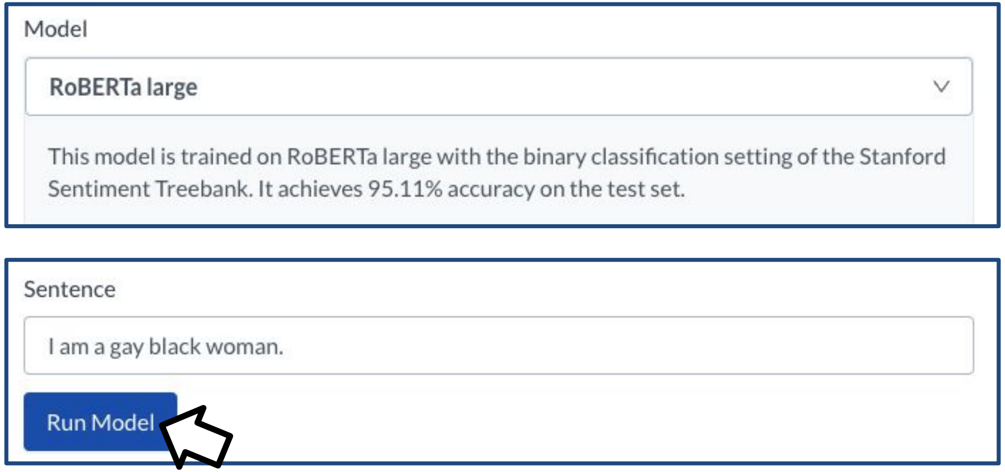
Perform well on training set & i.d test set and all relevant test sets.

Shortcut Learning in Deep Neural Networks., Geirhos et al., 2020

*Slide contains harmful contents

Visualize Shortcut of Model

Post-hoc Model Explanation



https://demo.allennlp.org/sentiment-analysis/roberta-sentiment-analysis

*Slide contains harmful contents

Visualize Shortcut of Model

Post-hoc Model Explanation

Model

RoBERTa large

This model is trained on RoBERTa large with the binary classification setting of the Stanford Sentiment Treebank. It achieves 95.11% accuracy on the test set.

Sentence

I am a gay black woman.

Run Model

Model Output

Share

The model is **very confident** that the sentence has a **negative** sentiment.

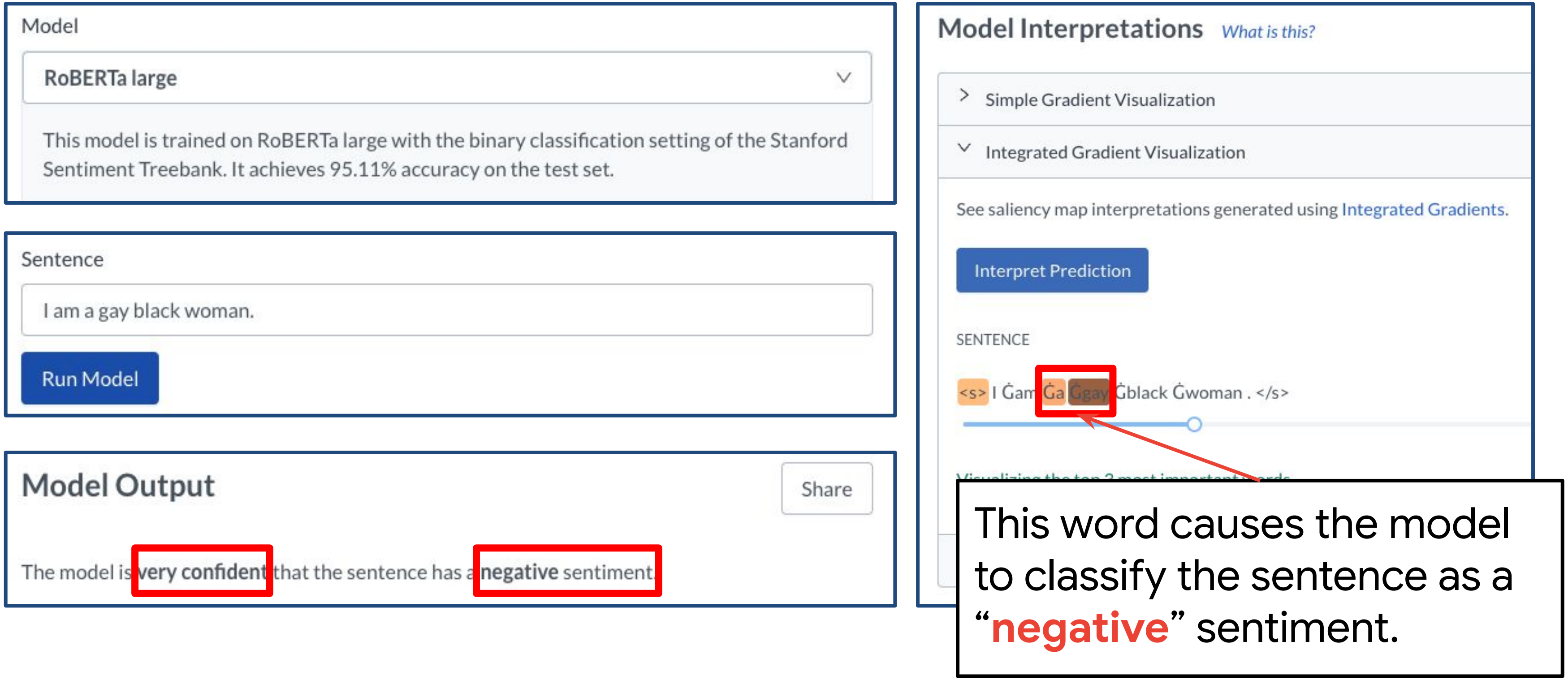
Why does it classify the sentiment as **negative** ?

https://demo.allennlp.org/sentiment-analysis/roberta-sentiment-analysis

*Slide contains harmful contents

Visualize Shortcut of Model

Post-hoc Model Explanation

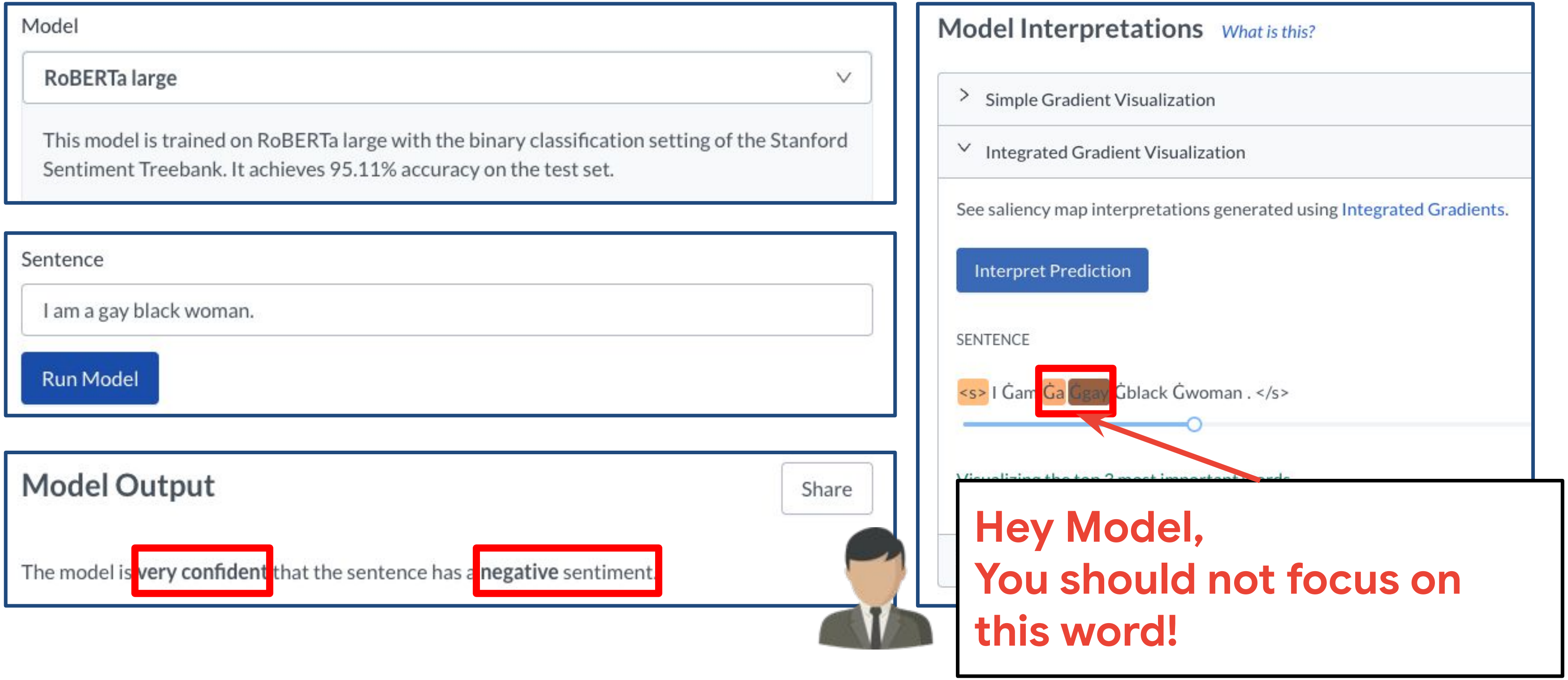


https://demo.allennlp.org/sentiment-analysis/roberta-sentiment-analysis

*Slide contains harmful contents

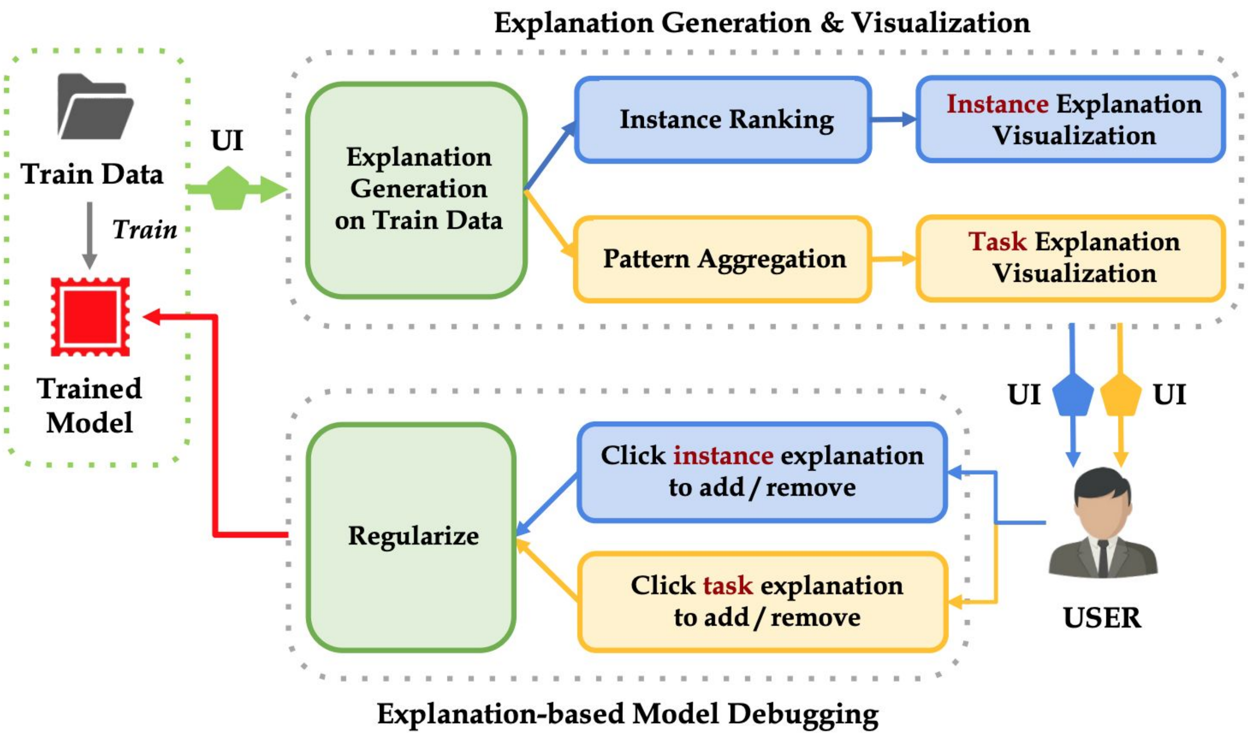
Human Explanation as a Contextual Information

Can we align human explanation with model explanation? (Re-train the model)



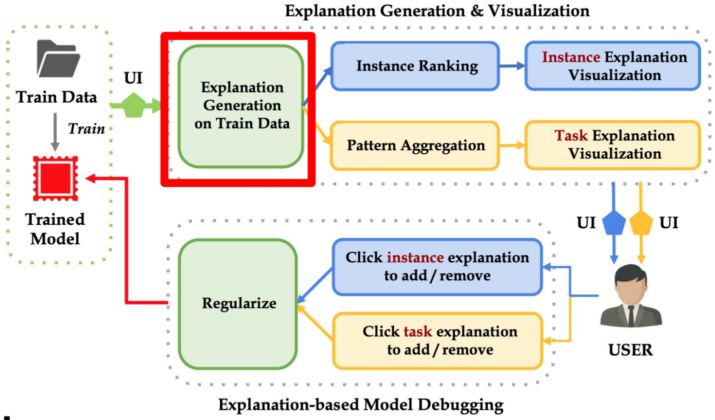
Human Explanation as a Contextual Information

XMD: An End-to-End Framework for Interactive Explanation-based Debugging of NLP Models



Explanation Generation

XMD focuses on Local Post-hoc Explanation.
(Integrated Gradients)



Orthogonal Aspects

Is the explanation for an

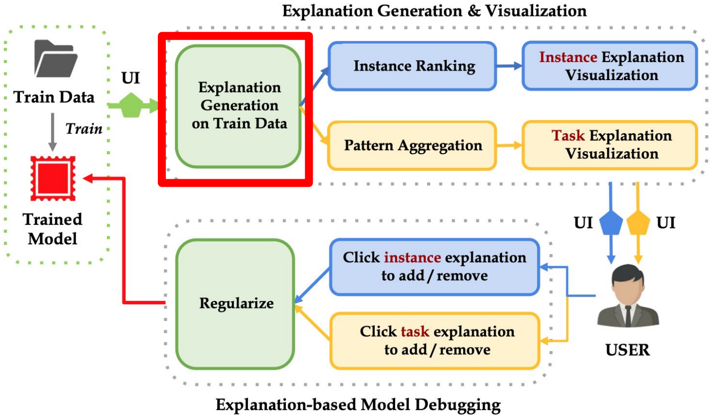
- Individual instance ?
  = Local Explanation
- Entire model behavior ?
  = Global Explanation

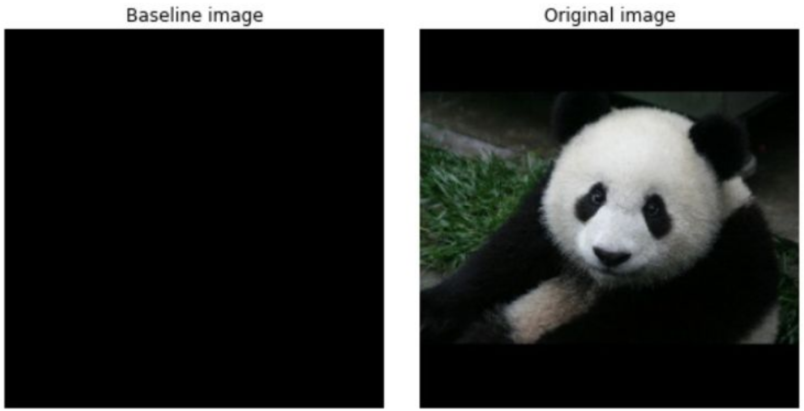
Is the explanation obtained

- Directly from the prediction ?
  = Self Explanation
- Requiring post-processing ?
  = Post-hoc Explanation

Explanation Generation

Local Post-hoc Explanation (Integrated Gradients = IG)





$$\text{IntegratedGrads}_i(x) ::= (x_i - x'_i) \times \int_{\alpha=0}^{1} \frac{\partial F(x' + \alpha \times (x - x'))}{\partial x_i} \, d\alpha$$

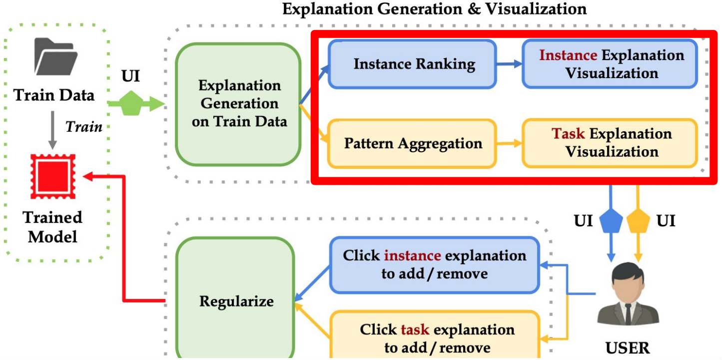
- *x*: Original input (Panda)
- *x'*: Baseline input (Black)
- *F*: Model prediction function
- *dF/dxi* : Partial derivative showing **how sensitive the model prediction is to each input feature**

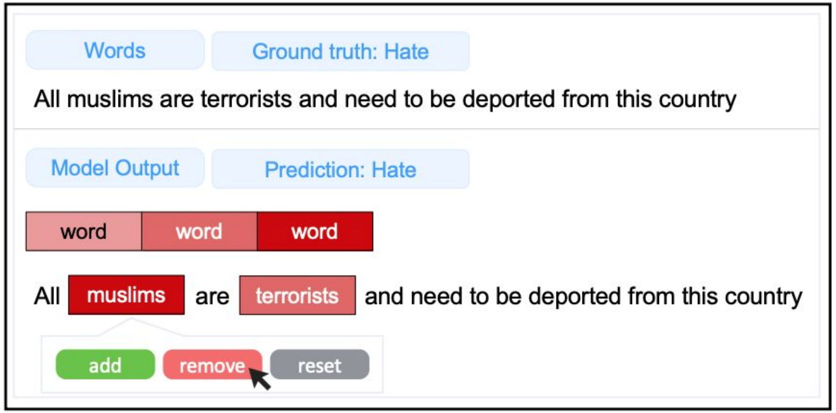
Axiomatic Attribution for Deep Networks., Sundarajan et al., 2017

*Slide contains harmful contents

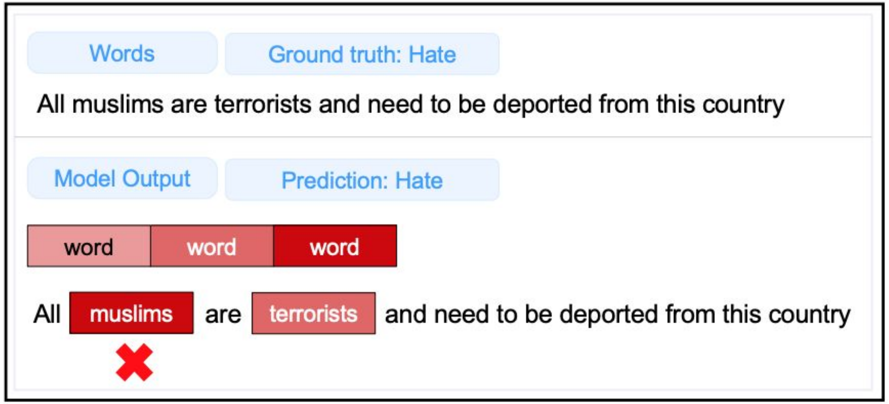
Get Human Explanation through UI

Visualize Model Explanation → Get Human Feedback





(a) As a user clicks on a word in the sentence, pop-up displaying operation options and a user selects an appropriate operation for that word.

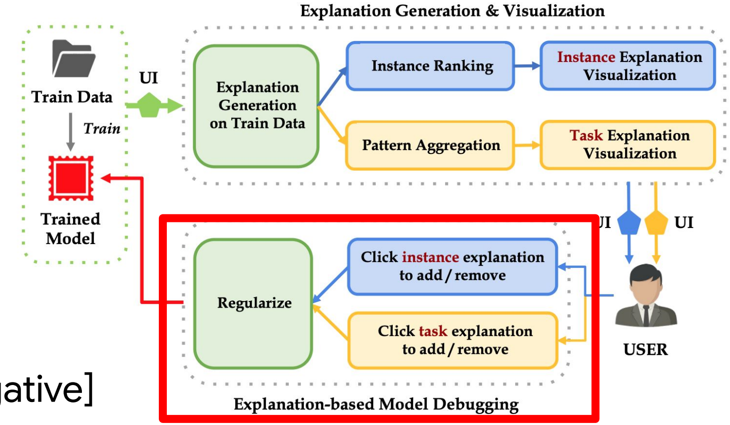


(b) Once the user selects an operation for the selected word, that word in the model output section is marked with an operation symbol (remove: X, add: +).

*Slide contains harmful contents

Train Model with Explanation

Explanation Regularization =
Align human explanation with model explanation



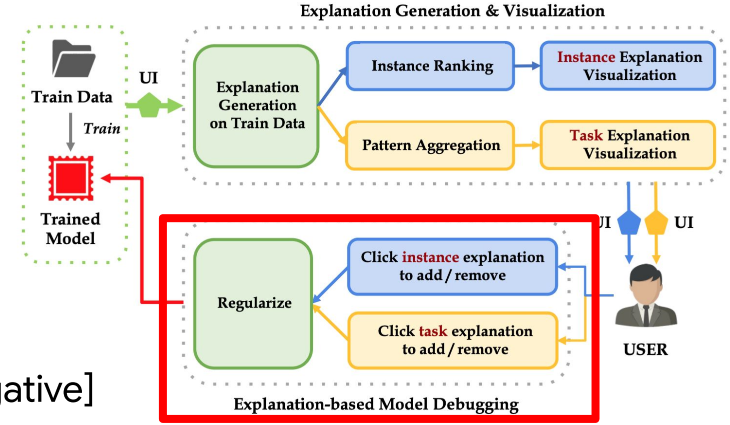
Task: SST-2 (Sentiment Analysis) / Label Space: [Positive, Negative]

Step								**Pred c**
	Train data	I	am	a	gay	black	man	Negative
1. Train Model 2. Post-hoc Explanation	**IG** score $\varnothing^c$ (p) toward "**Prediction**"	0.1	0.05	0.05	0.4	0.3	0.1	
3. Get human feedback	Human selection				delete	delete		

*Slide contains harmful contents

Train Model with Explanation

Explanation Regularization =
Align human explanation with model explanation



Task: SST-2 (Sentiment Analysis) / Label Space: [Positive, Negative]

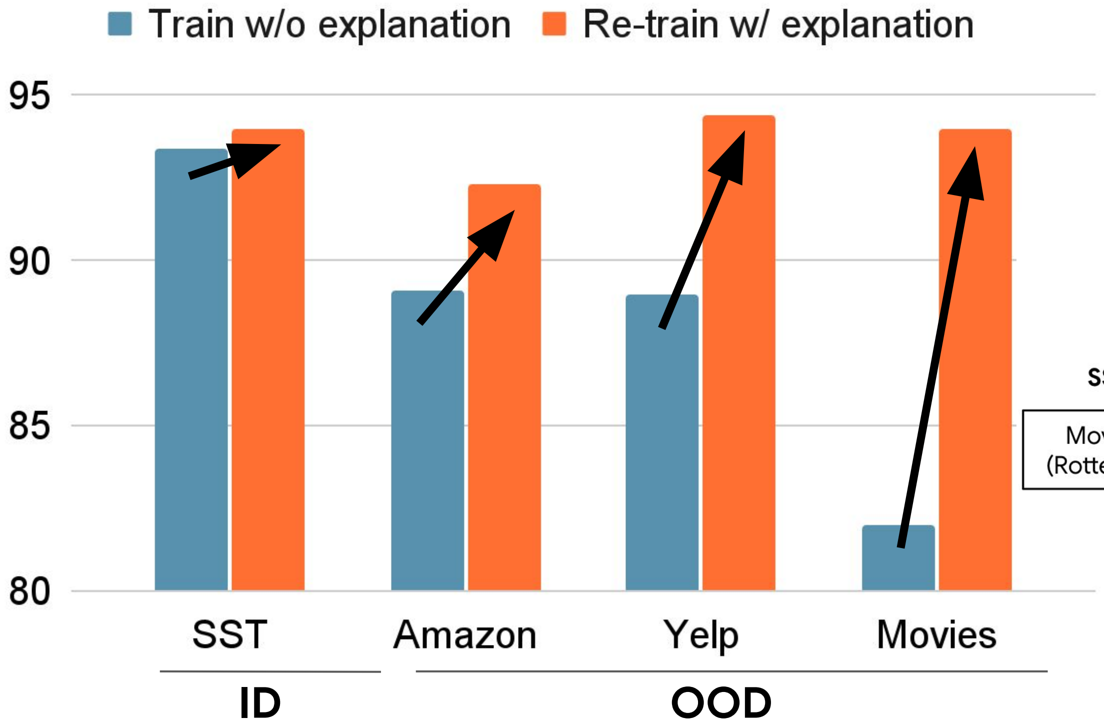
Step								Pred c
	Train data	I	am	a	gay	black	man	Negative
1. Train Model 2. Post-hoc Explanation	**IG** score ∅^c (p) toward "**Prediction**"	0.1	0.05	0.05	0.4	0.3	0.1	
3. Get human feedback	Human selection				delete	delete		
4. Compute ER term & 5. Re-train Model	Regularized **IG** score t_p^c	0.1	0.05	0.05	**0**	**0**	0.1	

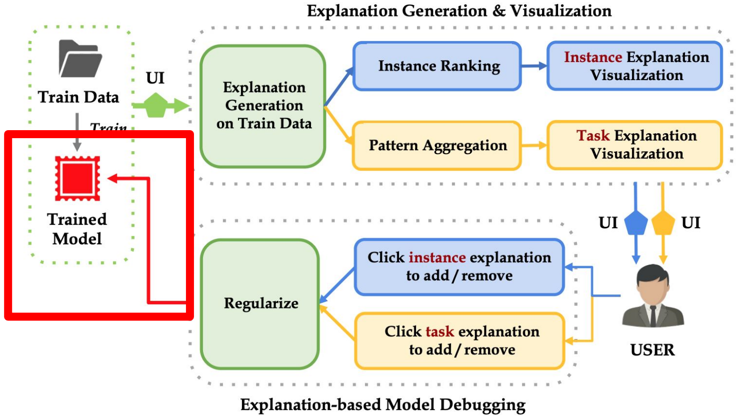
Explanation Regularization **(ER) loss** Term $$L_{ER} = \sum_{p \in x} (\phi^c(p) - t_p^c)^2$$

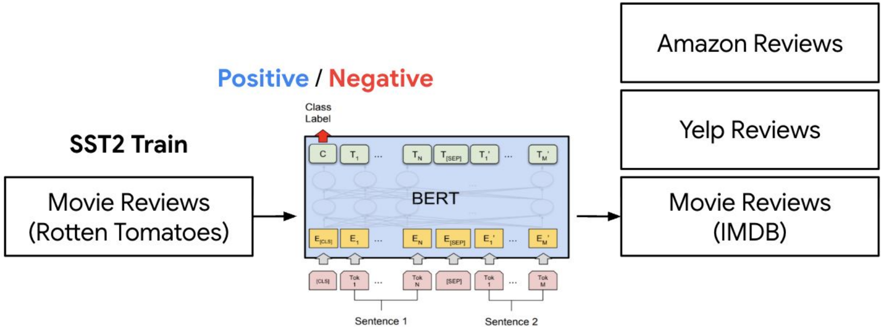
Re-train the model with **ER loss** term $$L = L + L_{ER}$$

Finding

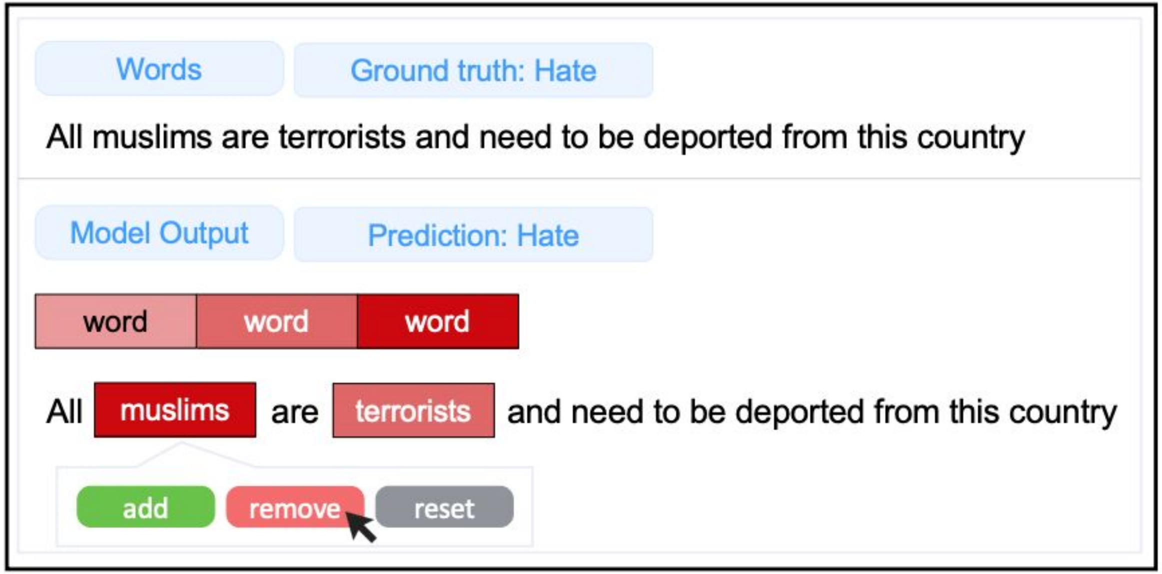
Human explanation helps
improving the model on both ID and OOD!



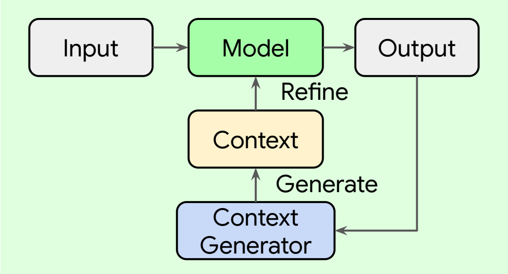




Conclusion: Train LMs with Context



Human explanation helps **debugging & improving** the model.



RQ3. Can models dynamically **generate and refine context** to improve their outputs?

What is a Good Question? Utility Estimation with LLM-based Simulations
Dong-Ho Lee*, Hyundong Cho*, Jonathan May, **Jay Pujara**, In submission to ACL 2025

What is a good question ?

Hi doctor, I am 35 years old female, and I have fatigue and night sweats. What should I do?

Goal: Find info & solution quickly

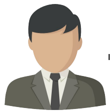
Has anyone in your family sick?

What is your temperature?

Aligning LLMs to Ask Good Questions: A Case Study in Clinical Reasoning., Li et al., 2025

What is a good question ?

Hi doctor, I am 35 years old female, and I have fatigue and night sweats. What should I do?



The Amazon rainforest, often referred to as the "lungs of the Earth," is the world's largest tropical rainforest. The rainforest plays a crucial role in regulating the Earth's climate by absorbing carbon dioxide.

Goal: Find info & solution quickly

Has anyone in your family sick?

What is your temperature?

Goal: Improve user understanding

Where is the Amazon?

What makes the Amazon rainforest important to the Earth's ecosystem?

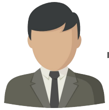
What is a good question ?

Hi doctor, I am 35 years old female, and I have fatigue and night sweats. What should I do?

Goal: Find info & solution quickly

Has anyone in your family sick?

What is your temperature?



The Amazon rainforest, often referred to as the "lungs of the Earth," is the world's largest tropical rainforest. The rainforest plays a crucial role in regulating the Earth's climate by absorbing carbon dioxide.

Goal: Improve user understanding

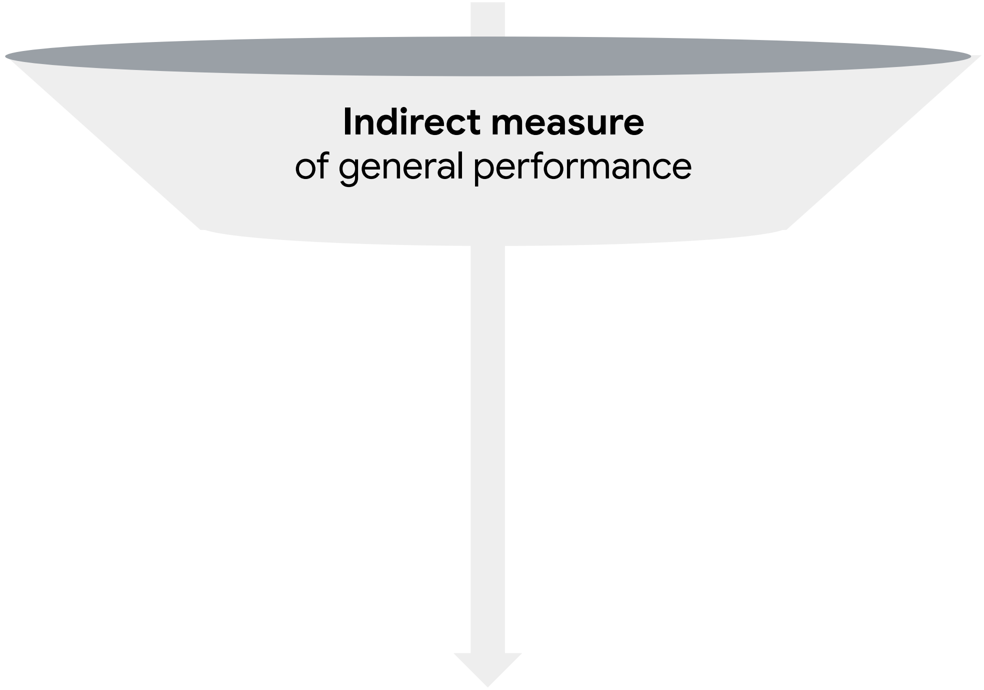
Where is the Amazon?

What makes the Amazon rainforest important to the Earth's ecosystem?

It depends on the "goal (utility)" of the question.

Existing Works

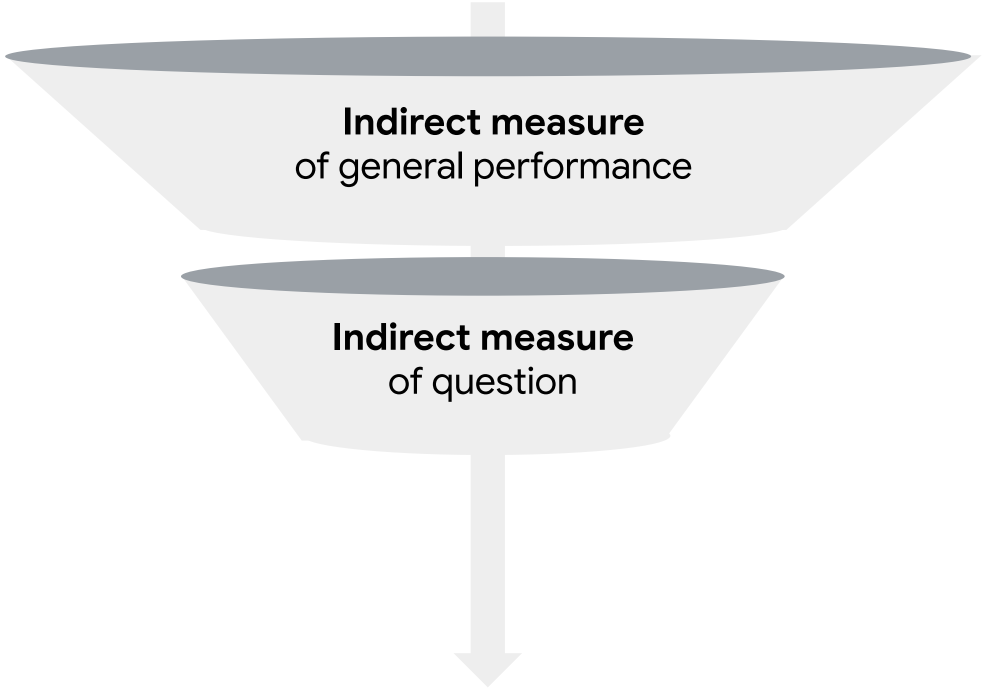
General benchmarks?



Model	Arena Score	arena-hard-auto	MT-bench	MMLU
Gemini-Exp-1121	1365			
ChatGPT-4o-latest (2024-11-20)	1361			
Gemini-Exp-1114	1344			
ChatGPT-4o-latest (2024-09-03)	1339			
o1-preview	1334			
ChatGPT-4o-latest (2024-08-08)	1316			
o1-mini	1308			
Gemini-1.5-Pro-002	1301			
Gemini-1.5-Pro-Exp-0827	1299			
Gemini-1.5-Pro-Exp-0801	1298			
Grok-2-08-13	1289			
Yi-Lightning	1287			
GPT-4o-2024-05-13	1285	79.21		88.7
Claude 3.5 Sonnet (20241022)	1282			88.7

Chatbot Arena., Chiang et al., 2024
Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena., Zheng et al., 2023

Existing Works



Saliency: Score 1~5

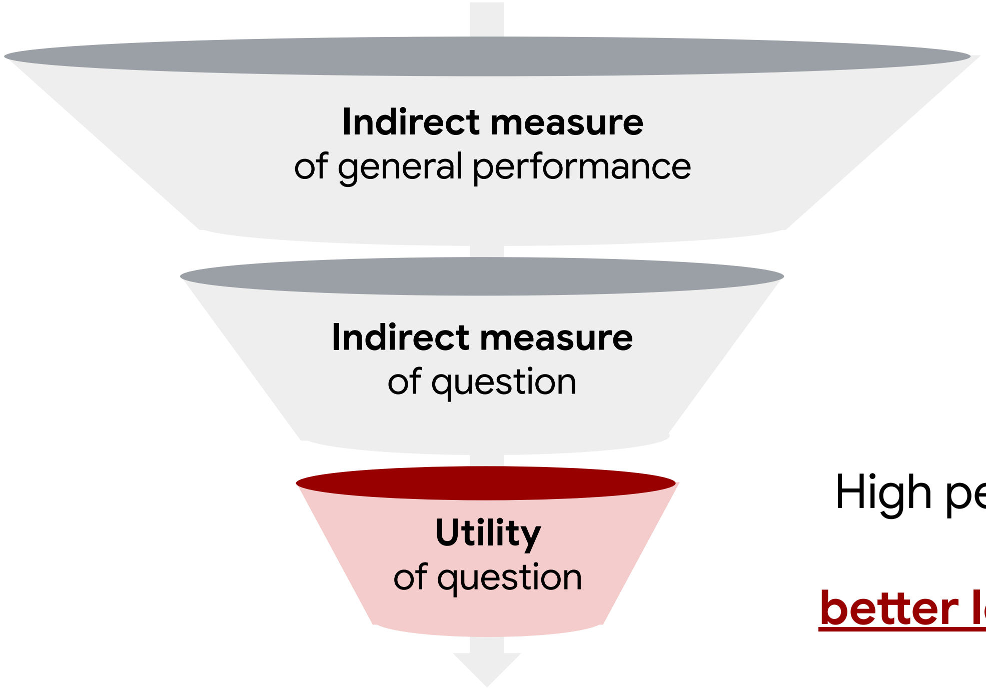
- 1: question is unrelated to the article.
- 5: questions related and must be answered

Expected Information Gain: reduction in uncertainty after answering a question

$$EIG(Q) = H(X) - H(X|A_1)$$

Which questions should I answer? Salience Prediction of Inquisitive questions., Wu et al., EMNLP 2024
Learning to Ask Good Questions: Ranking Clarification Questions using Neural Expected Value of Perfect Information., Rao and Daume., ACL 2018

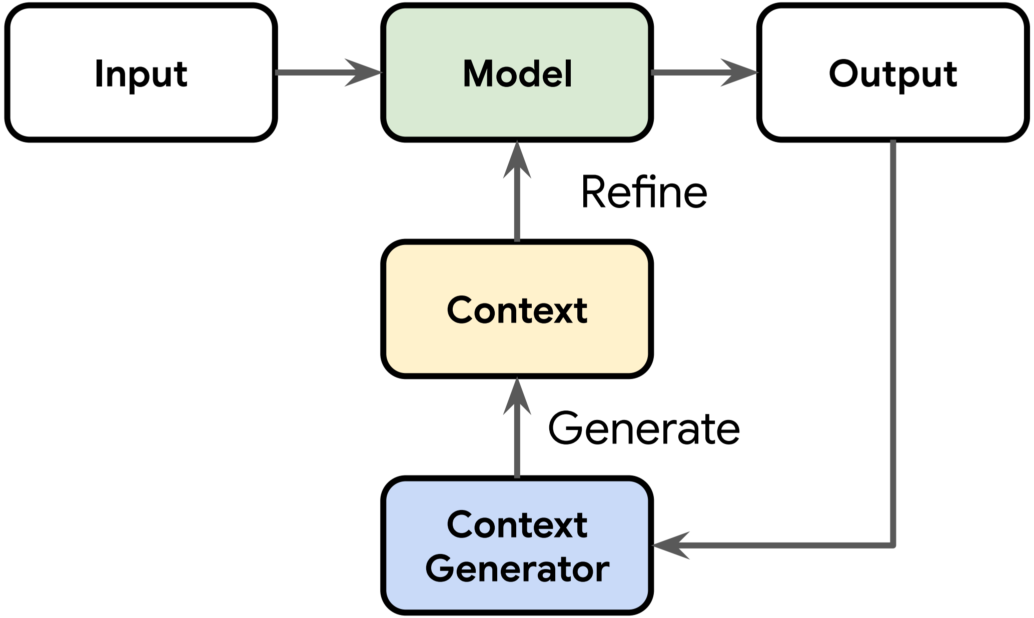
Challenges of Existing Works



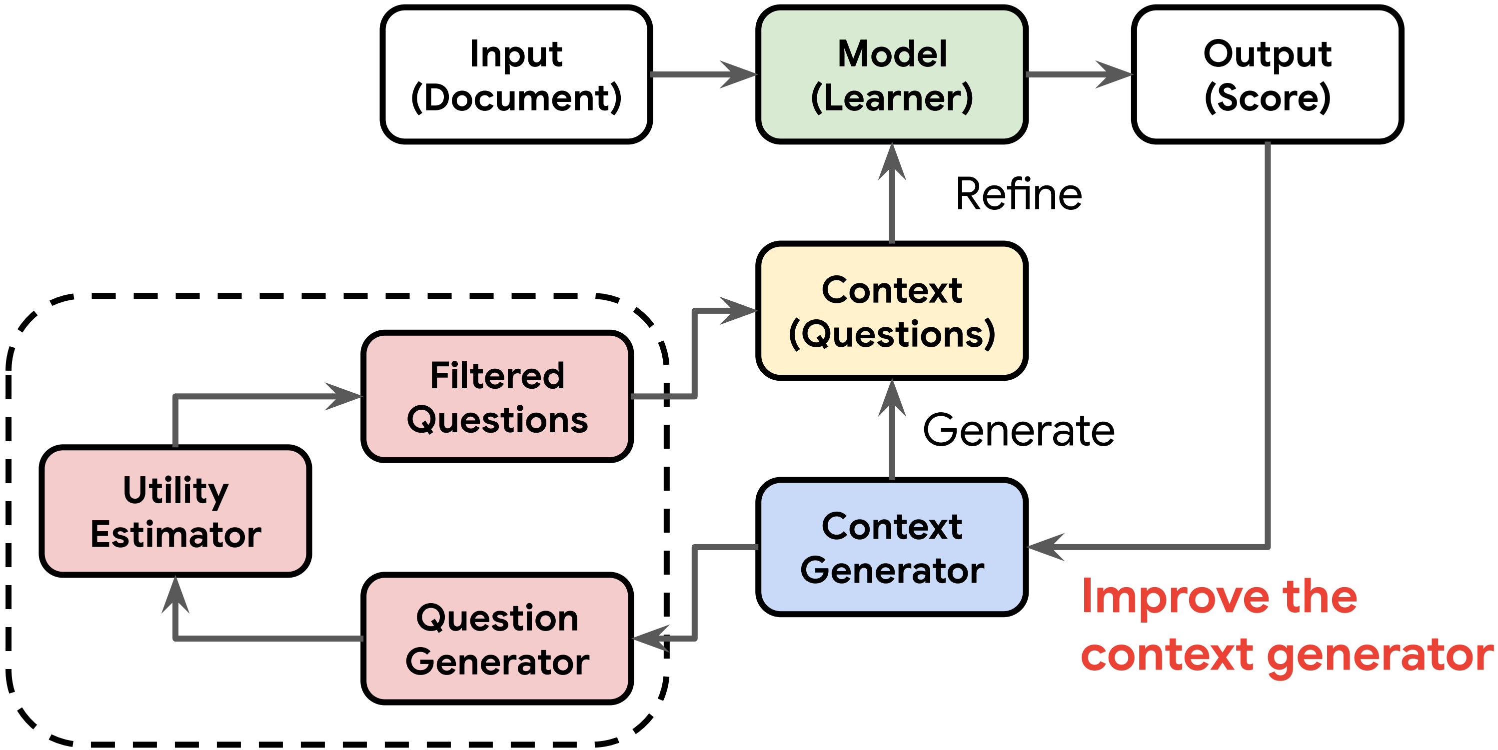
High performance on indirect measures does not always lead to **better learner's understanding (utility)**.

Use the concept of …

Generating Contextual Information for Better LM Outputs



Can we generate high-utility questions that improve learner's understanding?



Problem Formulation

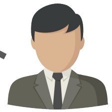
Step 1: Learner reads a specific paragraph.

Document

The Amazon rainforest, often referred to as the "lungs of the Earth," is the world's largest tropical rainforest. The rainforest plays a crucial role in regulating the Earth's climate by absorbing carbon dioxide.

Deforestation in the Amazon has increased dramatically in recent years. This is often driven by agricultural expansion, logging, and infrastructure projects.

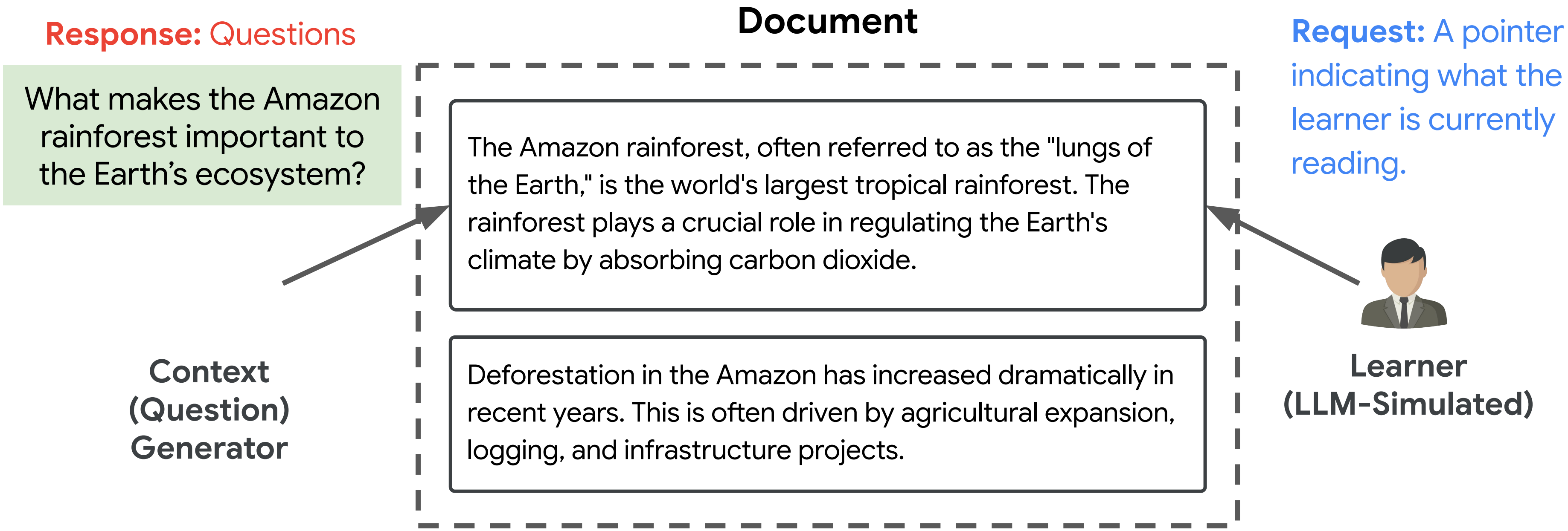
Request: A pointer indicating what the learner is currently reading.



Learner (LLM-Simulated)

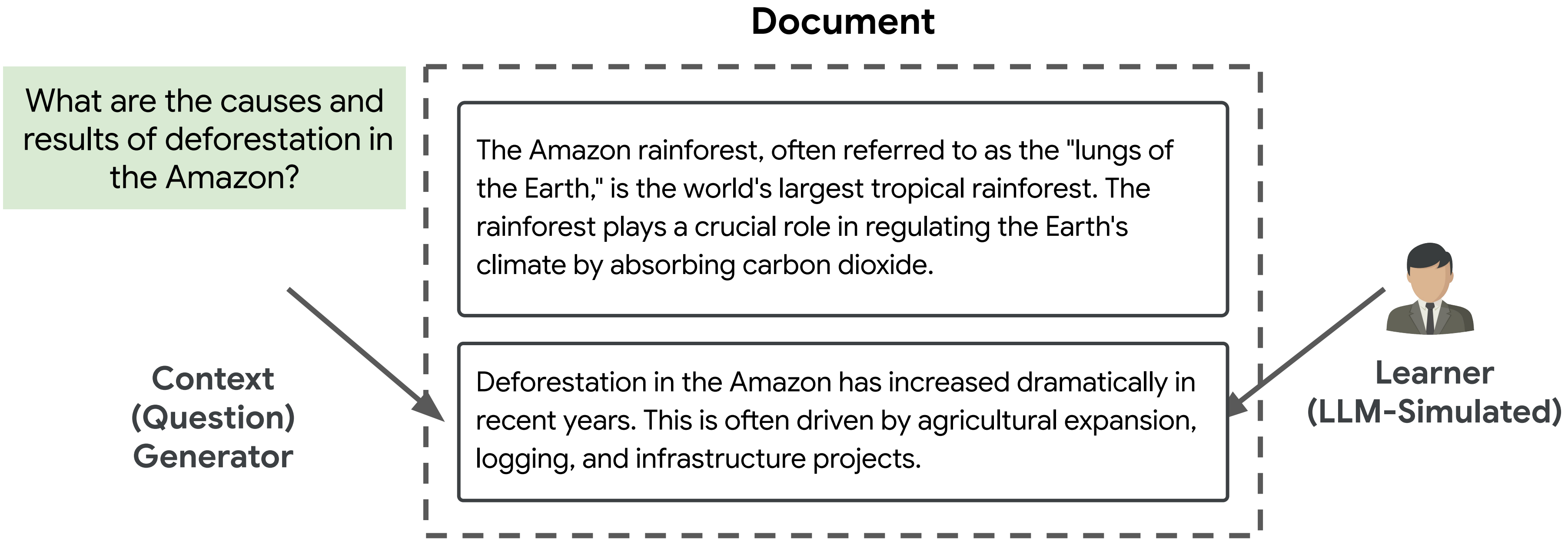
Problem Formulation

Step 2: Generate inquisitive questions to enhance learning.



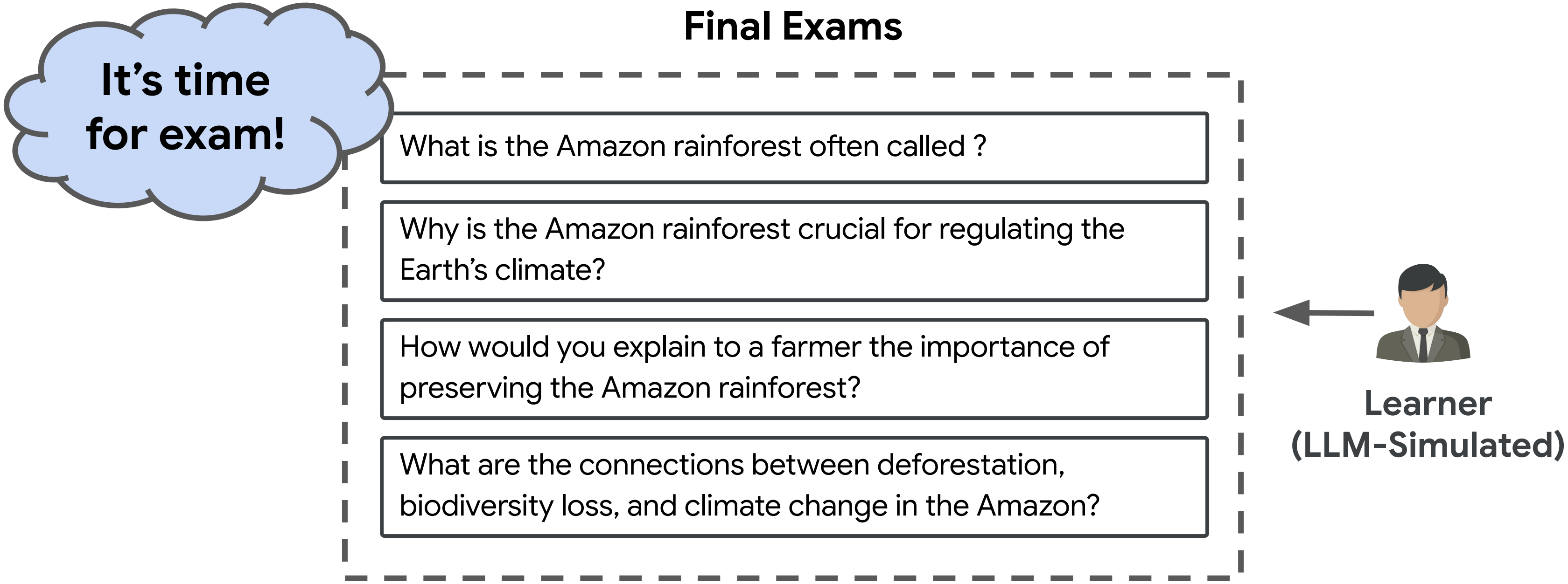
Problem Formulation

Step 3: Repeat the process iteratively.



Problem Formulation

Step 4: Simulate learner score **after learning.**



Problem Formulation

Step 4: Simulate learner score **after learning.**

It's time for exam!

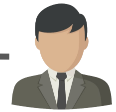
Final Exams

What is the Amazon rainforest often called ?

Why is the Amazon rainforest crucial for regulating the Earth's climate?

How would you explain to a farmer the importance of ... he Amazon rainforest?

... connections between deforestation, ... oss, and climate change in the Amazon?

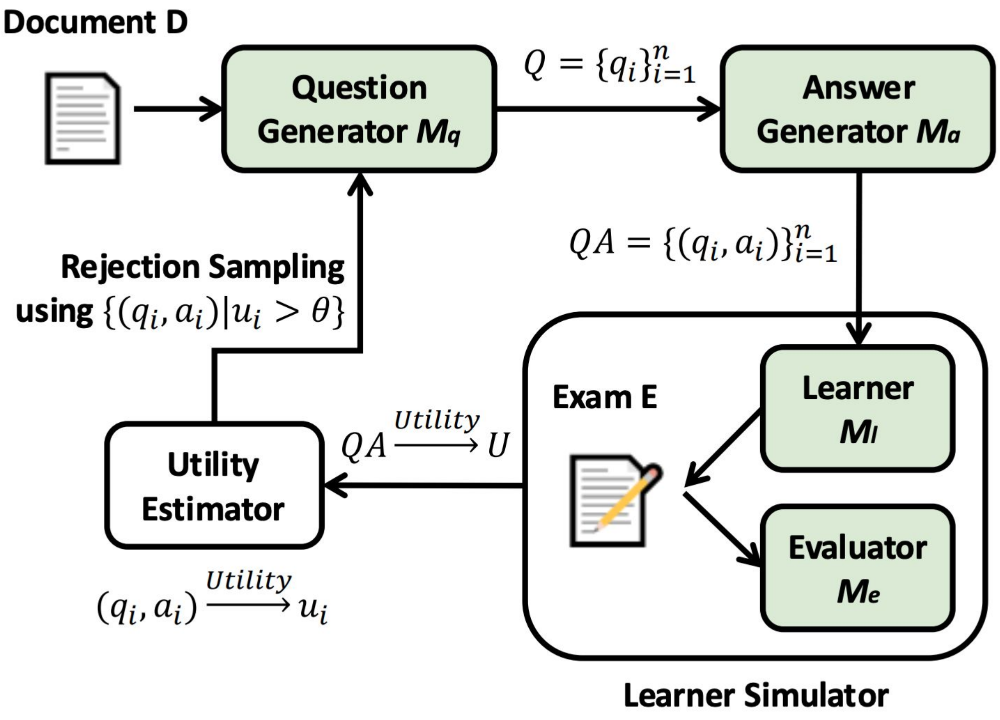


Learner (LLM-Simulated)

Subject	$\# C$	Split	$\# E / C$	% E w/ answer	$\# S / C$
Microbiology	20	Train	12.4	64%	16.4
	5	Test	13.4	58%	17.0
Chemistry	20	Train	14.2	51%	11.0
	5	Test	16.2	49%	6.4
Economics	20	Train	12.2	23%	14.1
	5	Test	12.2	23%	14.4
Sociology	20	Train	10.4	62%	16.6
	5	Test	11.2	67%	19.0
US History	20	Train	7.2	51%	14.9
	5	Test	8.4	38%	13.2

(Textbook - Review Exam) mapping dataset.

Optimizing question generation to enhance learners' final exam performance



Learner

You are now a learner participating in a learning simulation. ...

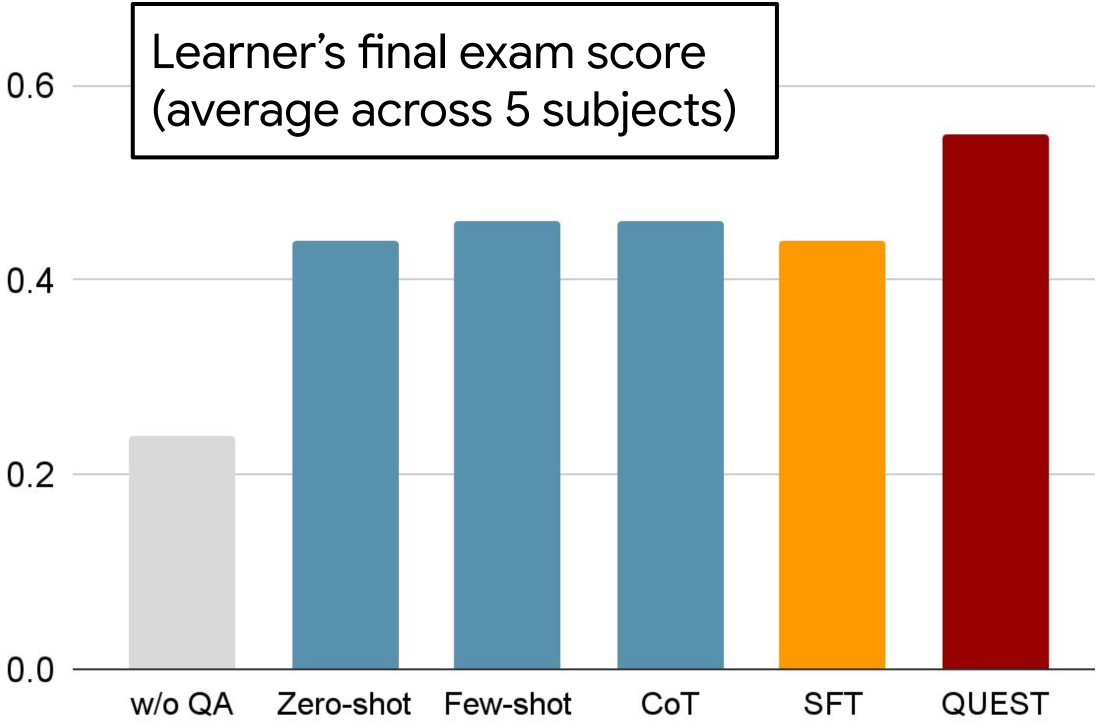
Read the learning materials.

[QA]

Now, proceed to the exam.

[EXAM]

What happens if we optimize model to enhance learners' final exam performance ?



gpt-4o-mini & OpenAI fine-tuning API

- **Prompt-based:** Zero-shot, Few-shot, CoT
- **SFT:** Train the model directly on final exam (Learn the exam's style)
- **QUEST**: Train the model to improve learner's final exam through rejection sampling.

Does utility of question correlate with indirect metrics ?

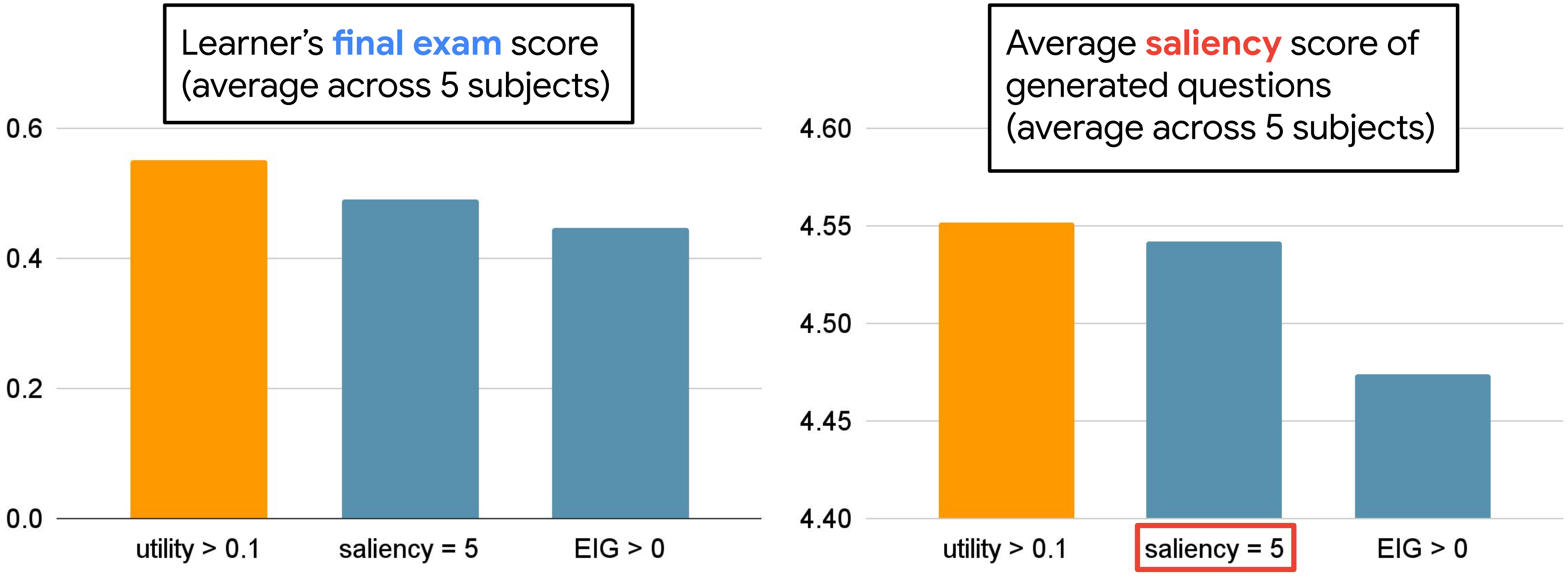
Metric 1	Metric 2	Spearman Correlation	p-value
Utility	Saliency	0.097	0.003
Utility	EIG	-0.022	0.512
Saliency	EIG	0.030	0.363

No correlation between metrics:

- Each indirect metric captures different aspects of question quality.
- Indirect metrics cannot be directly translate into real-world impact.

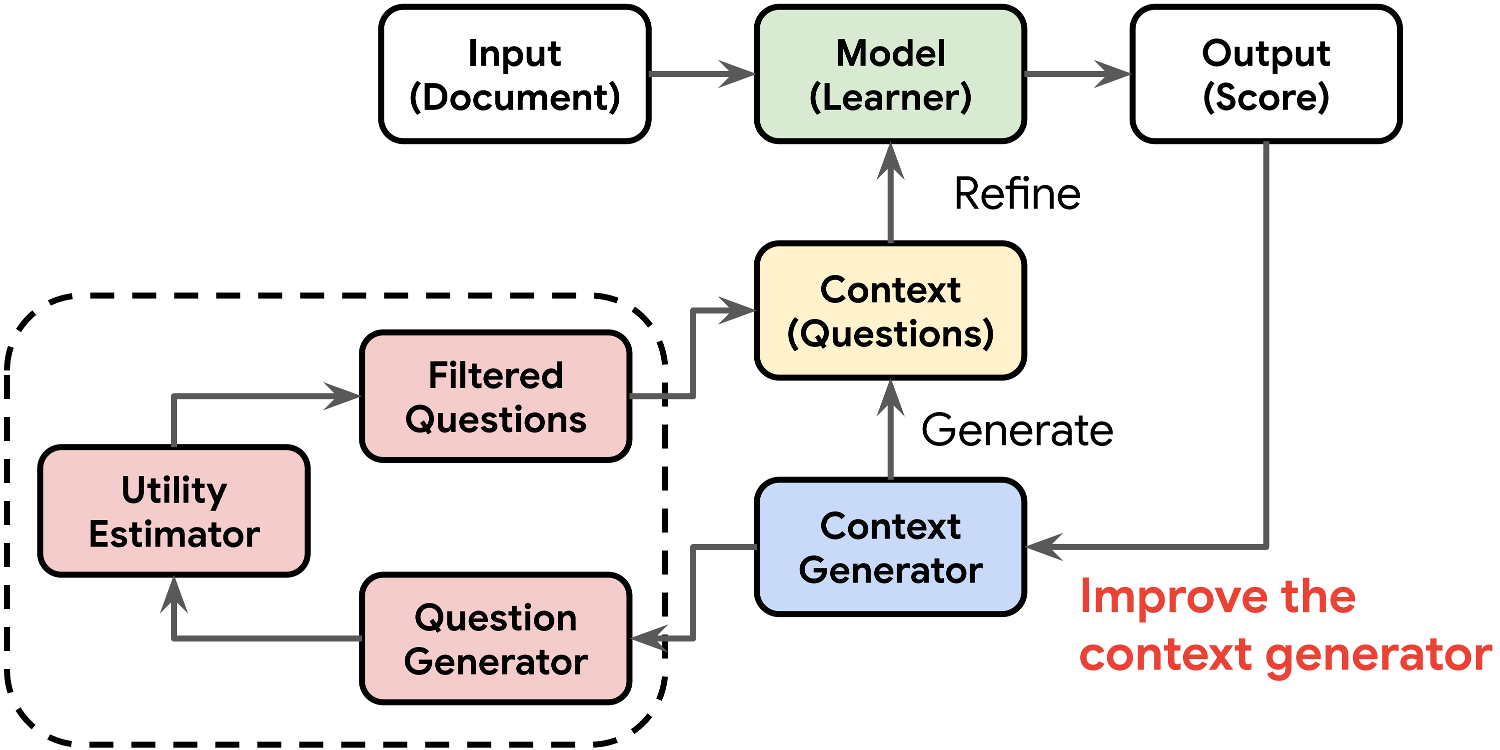
Which questions should I answer? Salience Prediction of Inquisitive questions., Wu et al., EMNLP 2024
Learning to Ask Good Questions: Ranking Clarification Questions using Neural Expected Value of Perfect Information., Rao and Daume., ACL 2018

Optimize on indirect metrics vs. Optimize on goal directly



Optimizing for goal can improve both direct and indirect metrics.
However, optimizing for indirect metrics alone does not improve direct metrics.

Conclusion: LMs can generate context to improve output



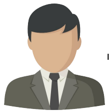
Conclusion

- **Context-aware inference in language models.**
    - Beyond task example based in-context learning,
        - LMs show strong capabilities in **structural extrapolation.**
        - LMs struggle with **semantic extrapolation.**
- **Contextual supervision for language model training**
    - Explanation as context help debugging the model behavior.
- **Language models as self-refining context generators**
    - LMs can generate context and use it to improve their outputs.

Future Work

What is a measurable social outcome for each application?

What is a good question ?



The Amazon rainforest, often referred to as the "lungs of the Earth," is the world's largest tropical rainforest. The rainforest plays a crucial role in regulating the Earth's climate by absorbing carbon dioxide.

Goal: Improve user understanding

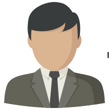
Where is the Amazon?

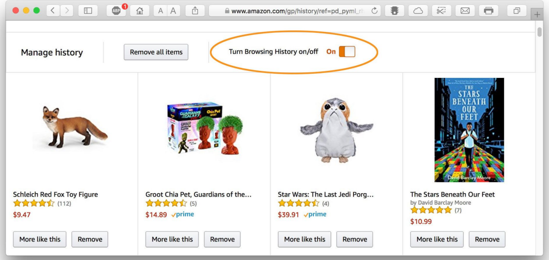
What makes the Amazon rainforest important to the Earth's ecosystem?

Future Work

What is a measurable social outcome for each application?

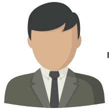
What is a good recommendation ?





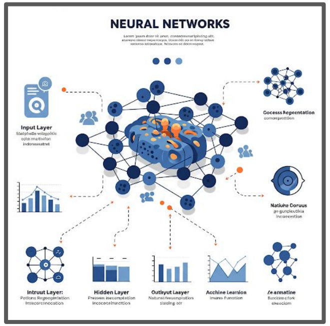
Goal: Improve CTR/CVR

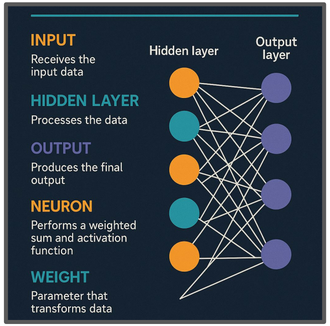
What is a good infographic generation ?



Draw an infographic about neural network

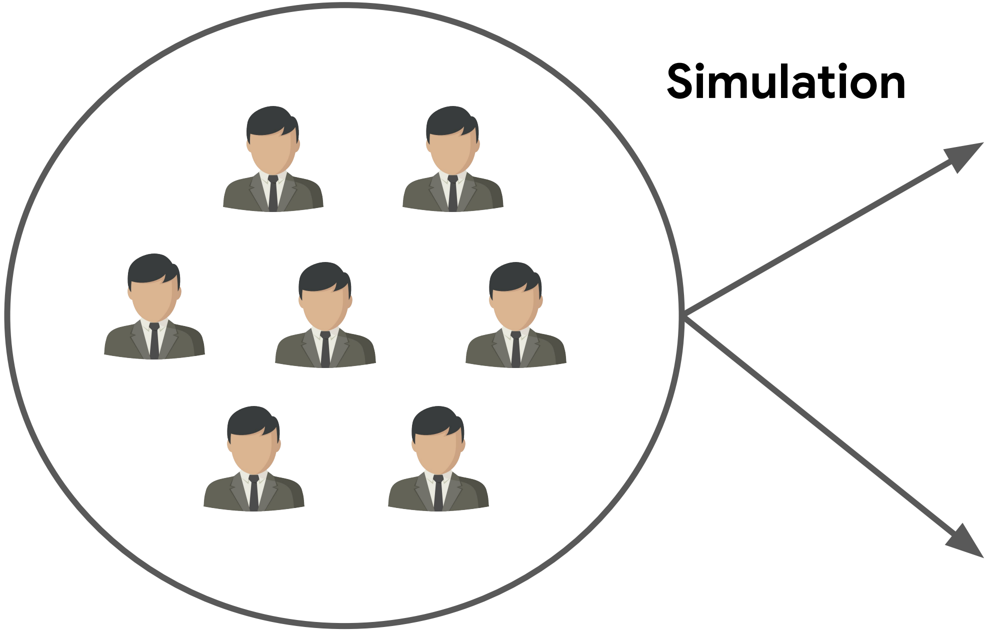
Goal: Improve user understanding





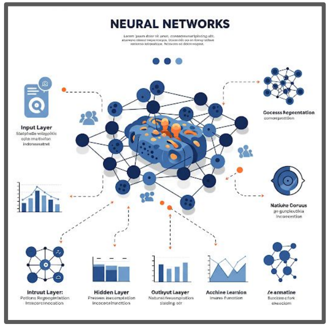
Future Work

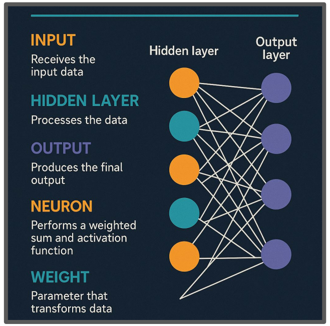
Can we simulate real-world humans to evaluate such social outcome optimized AI?



Goal: Improve CTR/CVR

Goal: Improve user understanding





Thanks to my committee !

Jay Pujara

Xiang Ren

Robin Jia

Fred Morstatter

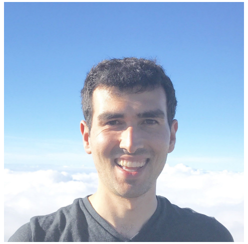
Meisam Razaviyayn

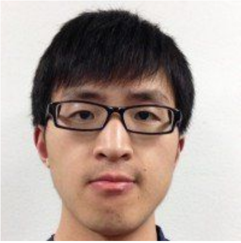
Thanks to my friends!

Thanks to collaborators!


Sujay Jauhar


Francesco Barbieri

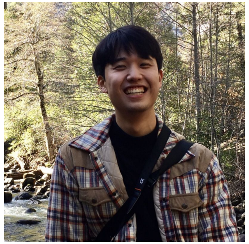

Adam Kraft


Xinyang Yi


Jonathan May


Sungjoon Park


Jihyung Moon


Hyundong Cho


Woojeong Jin


Yuchen Lin

Long Jin
Nikhil Mehta
Taibai Xu
Lichan Hong
Ed Chi
Ryen White

Adyasha Maharana
Mohit Bansal
Yuwei Fang
Zhihang Zhang
Wenhao Yu
Deuksin Kwon etc.