Confidence-based Rephrasing, Refinement, and Selection

UncertaiNLP Workshop, EACL 2024

Elias Stengel-Eskin

03/22/2024

THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

UNC NLP

MURGE-LAB

Outline

Part I: Uncertainty in Human-Model Interactions

Calibrated Interpretation: Confidence Estimation in Semantic Parsing, Elias Stengel-Eskin and Benjamin Van Durme, TACL (2023)

Did You Mean...? Confidence-based Trade-offs in Semantic Parsing, Elias Stengel-Eskin and Benjamin Van Durme, EMNLP (2023)

Part II: Model-based Selection to Reduce Uncertainty

Rephrase, Augment, Reason: Visual Grounding of Questions for Vision-Language Models, Archiki Prasad, Elias Stengel-Eskin, Mohit Bansal, ICLR (2024)

Part III: Confidence for Model-Model Interactions

ReConcile: Round-Table Conference Improves Reasoning via Consensus among Diverse LLMs, Justin Chih-Yao Chen, Swarnadeep Saha, Mohit Bansal (2024)

MAGDi: Structured Distillation of Multi-Agent Interaction Graphs Improves Reasoning in Smaller Language Models, Justin Chih-Yao Chen*, Swarnadeep Saha*, Elias Stengel-Eskin Mohit Bansal (2024)

Executable parsing

Predicting executable programs

Do I have anything going on tonight?

```
(Yield (> (size
   (QueryEventResponse.results
      (… (EventOnDateWithTimeRange
         (EventOnDate (Today)
            (^(Event) EmptyStructConstraint))
            (Night)))) 0L))
```

Executable parsing

Predicting executable programs

Language as an API for interaction

Executable parsing

Predicting executable programs

Language as an API for interaction

Domains

- Querying: Zelle and Mooney (1996), Berant et al. (2013), Yu et al. (2018)
- Digital assistants: Semantic Machines (2020), Damonte et al. (2019)
- Robotics: Kate et al. (2005),Tellex et al. (2011), Artzi et al. (2013), Tellex et al. (2020)

Why semantic parsing?

Form vs. meaning

Focus on semantics

Restricted output space

Uncertainty in natural language generation: From theory to applications
Baan et al., 2023

Interpreting Predictive Probabilities: Model Confidence or Human Label Variation?
Baan et al., 2024

Why semantic parsing?

Form vs. meaning

Focus on semantics

Restricted output space

Measurability

When executable, accuracy is well-defined

Uncertainty in natural language generation: From theory to applications
Baan et al., 2023

Interpreting Predictive Probabilities: Model Confidence or Human Label Variation?
Baan et al., 2024

Why semantic parsing?

Form vs. meaning

Focus on semantics

Restricted output space

Measurability

When executable, accuracy is well-defined

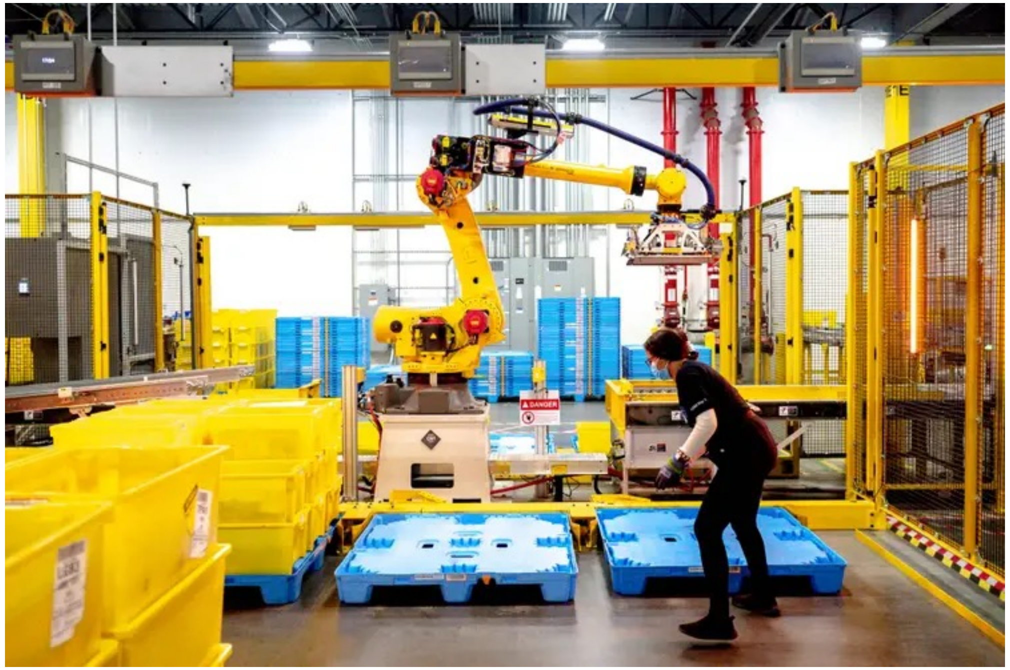
Safety concerns

Agents doing things in the real world

Uncertainty in natural language generation: From theory to applications
Baan et al., 2023

Interpreting Predictive Probabilities: Model Confidence or Human Label Variation?
Baan et al., 2024

Execution and safety



Execution and safety

A = `drop_item()` $\qquad \neg A$ = `do_nothing()`

Execution and safety

A = `drop_item()` $\neg A$ = `do_nothing()`

G = `drop_item` $\neg G$ = `do_nothing`

Execution and safety

A = `drop_item()` $\neg A$ = `do_nothing()`

G = `drop_item` $\neg G$ = `do_nothing`

I = *let's drop it*

Execution and safety

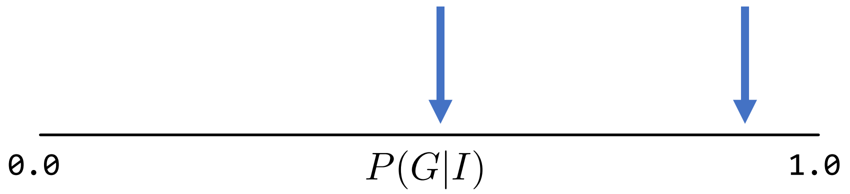
A = `drop_item()`

G = `drop_item`

I = *let's drop it*

$\neg A$ = `do_nothing()`

$\neg G$ = `do_nothing`

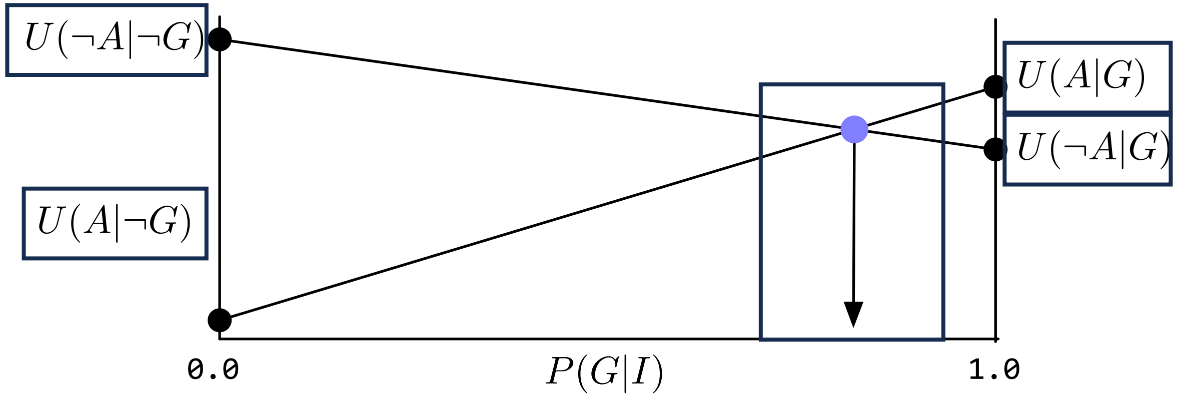


Execution and safety

A = `drop_item()` $\neg A$ = `do_nothing()`

G = `drop_item` $\neg G$ = `do_nothing`

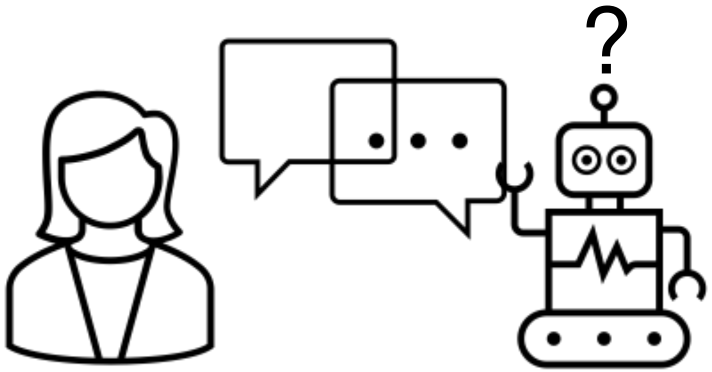
I = *let's drop it*



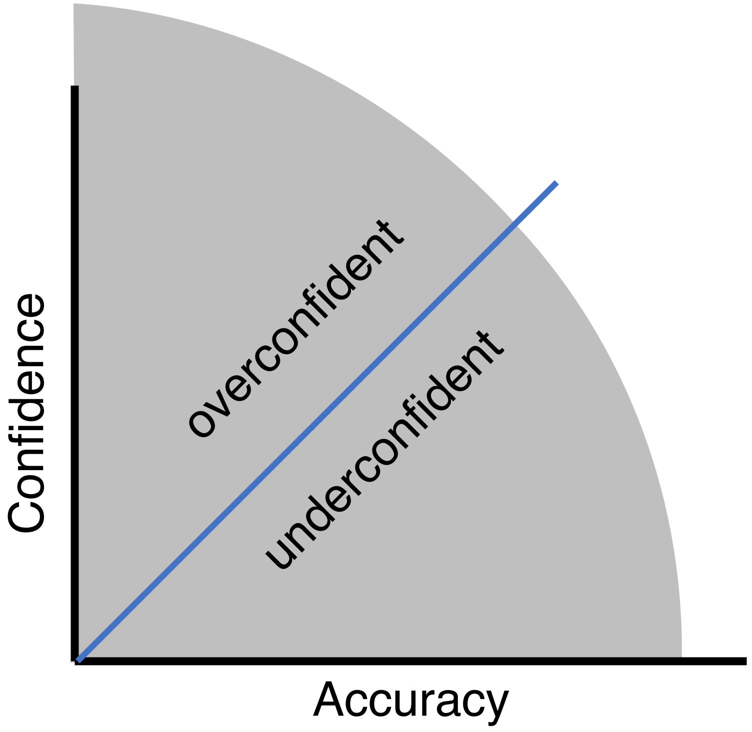

Principles of Mixed-Initiative User Interfaces

Eric Horvitz
Microsoft Research

Calibration



Available w/ out execution



Unavailable w/ out execution

Outline

Part I: Uncertainty in Human-Model Interactions

Calibrated Interpretation: Confidence Estimation in Semantic Parsing, Elias Stengel-Eskin and Benjamin Van Durme, TACL (2023)

Did You Mean...? Confidence-based Trade-offs in Semantic Parsing, Elias Stengel-Eskin and Benjamin Van Durme, EMNLP (2023)

Part II: Model-based Selection to Reduce Uncertainty

Rephrase, Augment, Reason: Visual Grounding of Questions for Vision-Language Models, Archiki Prasad, Elias Stengel-Eskin, Mohit Bansal, ICLR (2024)

Part III: Confidence for Model-Model Interactions

ReConcile: Round-Table Conference Improves Reasoning via Consensus among Diverse LLMs, Justin Chih-Yao Chen, Swarnadeep Saha, Mohit Bansal (2024)

MAGDi: Structured Distillation of Multi-Agent Interaction Graphs Improves Reasoning in Smaller Language Models, Justin Chih-Yao Chen*, Swarnadeep Saha*, Elias Stengel-Eskin Mohit Bansal (2024)

Goals

How to extract confidence from models?

How well-calibrated are semantic parsing models?

Datasets: digital assistant

SMCalFlow

Calendaring domain

Lisp-like programs

TreeDST

several domains

same format

🧑 *Do I have anything going on tonight?*

🤖

```
(Yield (> (size
   (QueryEventResponse.results
      (… (EventOnDateWithTimeRange
         (EventOnDate (Today)
            (^(Event) EmptyStructConstraint))
            (Night)))) 0L))
```

Conversational Semantic Parsing for Dialog State Tracking
Cheng et al., 2020

Task-oriented dialogue as dataflow synthesis
Semantic Machines et al., 2020

Measuring Calibration Error

$$ECE(\mathcal{B}) = \sum_{i=1}^{N} \frac{|\mathcal{B}_i|}{N} \left| \frac{\sum_{j \in \mathcal{B}_i} a_j}{|\mathcal{B}_i|} - \frac{\sum_{j \in \mathcal{B}_i} c_j}{|\mathcal{B}_i|} \right|$$

where $\mathcal{B}$ are the N bins,
a_i is the accuracy (0 or 1) and
c_j is the model confidence

Measuring Calibration Error

1. Obtain word-level confidence (min over token probabilities)
2. Bin by confidence
3. Compute average accuracy per bin
4. Expected calibration error (ECE) is difference between confidence and accuracy

$$ECE(\mathcal{B}) = \sum_{i=1}^{N} \frac{|\mathcal{B}_i|}{N} \left| \frac{\sum_{j \in \mathcal{B}_i} a_j}{|\mathcal{B}_i|} - \frac{\sum_{j \in \mathcal{B}_i} c_j}{|\mathcal{B}_i|} \right|$$

```
 S      EL     ECT
0.92   0.31   0.85

SELECT → 0.31
```

where $\mathcal{B}$ are the N bins,

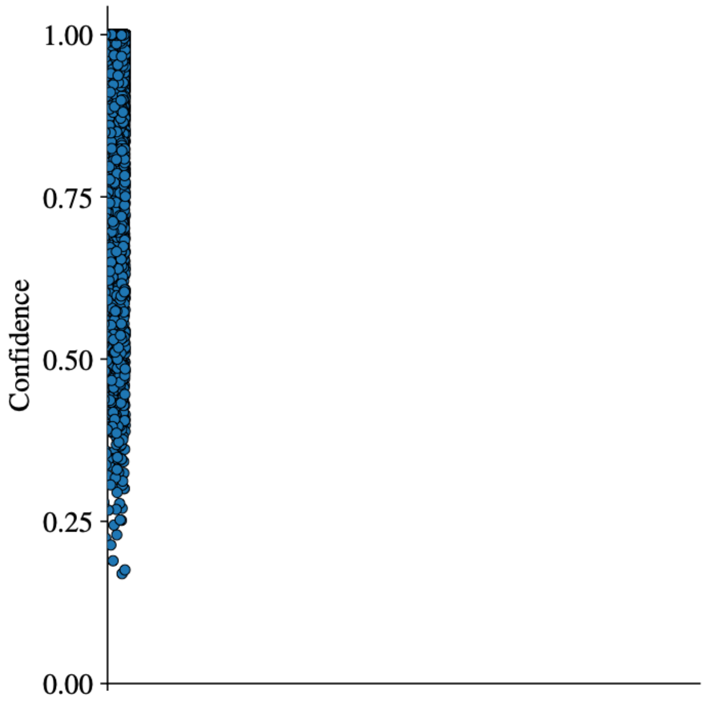
a_i is the accuracy (0 or 1) and

c_j is the model confidence

Measuring Calibration Error

1. **Obtain word-level confidence (min over token probabilities)**
2. Bin by confidence
3. Compute average accuracy per bin
4. Expected calibration error (ECE) is difference between confidence and accuracy

$$ECE(\mathcal{B}) = \sum_{i=1}^{N} \frac{|\mathcal{B}_i|}{N} \left| \frac{\sum_{j \in \mathcal{B}_i} a_j}{|\mathcal{B}_i|} - \frac{\sum_{j \in \mathcal{B}_i} c_j}{|\mathcal{B}_i|} \right|$$



where $\mathcal{B}$ are the N bins,

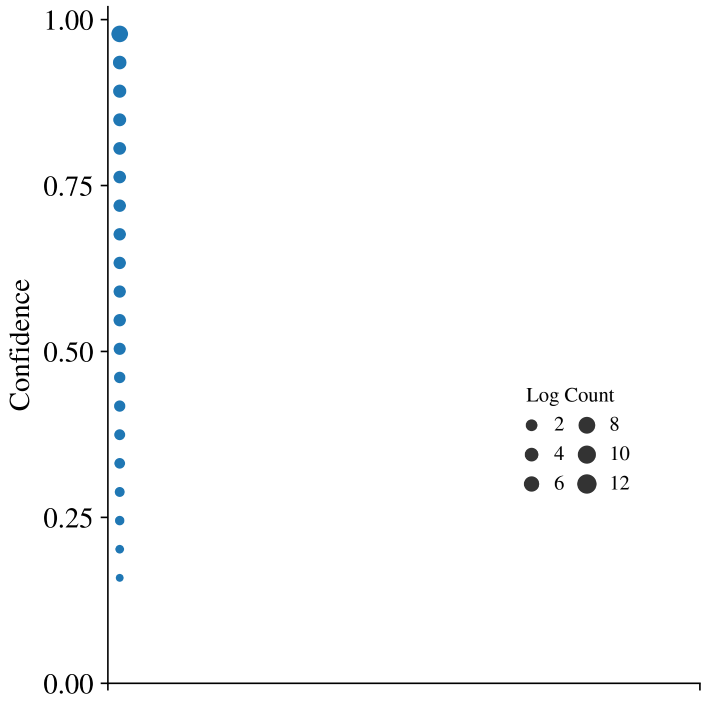
a_i is the accuracy (0 or 1) and

c_j is the model confidence

Measuring Calibration Error

1. Obtain word-level confidence (min over token probabilities)
2. **Bin by confidence**
3. Compute average accuracy per bin
4. Expected calibration error (ECE) is difference between confidence and accuracy

$$ECE(\mathcal{B}) = \sum_{i=1}^{N} \frac{|\mathcal{B}_i|}{N} \left| \frac{\sum_{j \in \mathcal{B}_i} a_j}{|\mathcal{B}_i|} - \frac{\sum_{j \in \mathcal{B}_i} c_j}{|\mathcal{B}_i|} \right|$$



where $\mathcal{B}$ are the N bins,

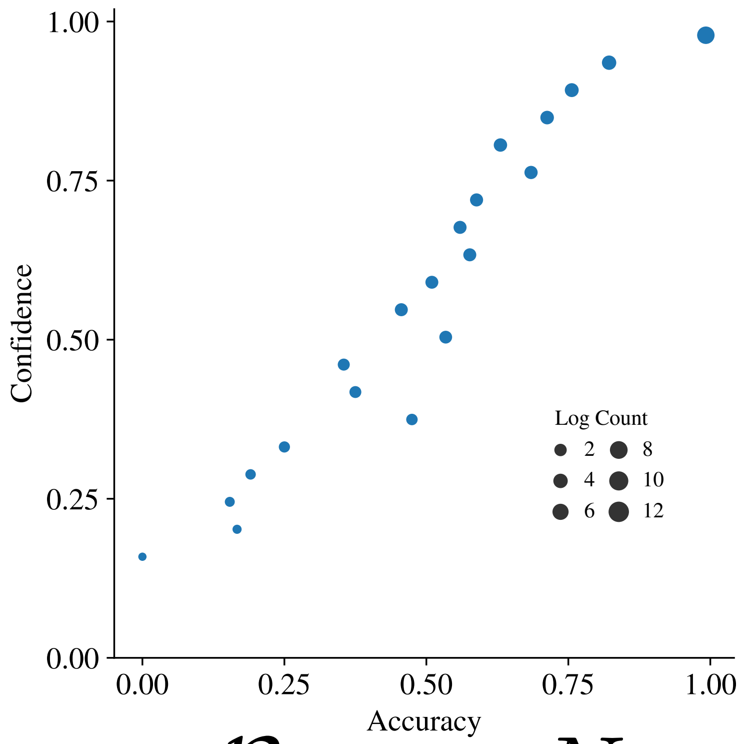
a_i is the accuracy (0 or 1) and

c_j is the model confidence

Measuring Calibration Error

1. Obtain word-level confidence (min over token probabilities)
2. Bin by confidence
3. **Compute average accuracy per bin**
4. Expected calibration error (ECE) is difference between confidence and accuracy

$$ECE(\mathcal{B}) = \sum_{i=1}^{N} \frac{|\mathcal{B}_i|}{N} \left| \frac{\sum_{j \in \mathcal{B}_i} a_j}{|\mathcal{B}_i|} - \frac{\sum_{j \in \mathcal{B}_i} c_j}{|\mathcal{B}_i|} \right|$$



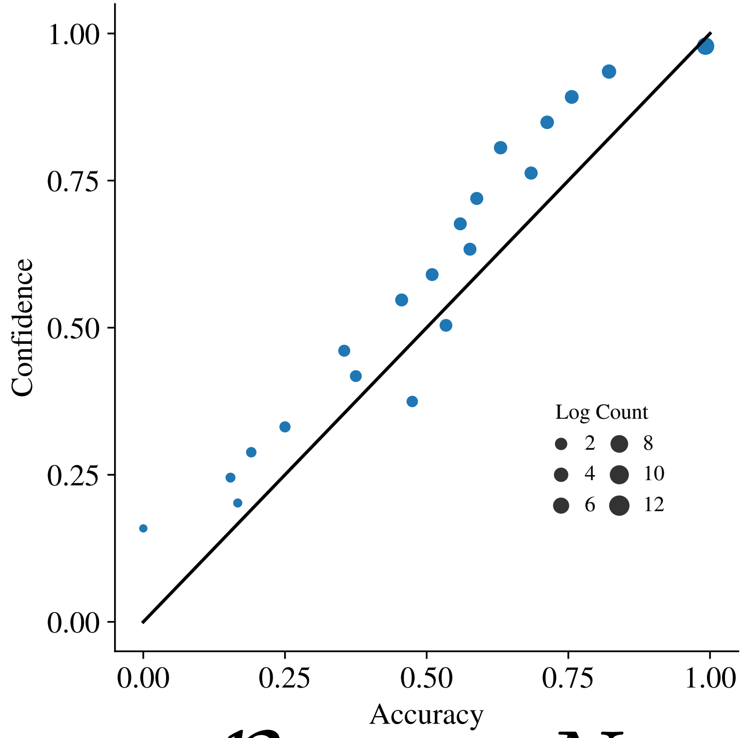
where $\mathcal{B}$ are the N bins,

a_i is the accuracy (0 or 1) and

c_j is the model confidence

Measuring Calibration Error

1. Obtain word-level confidence (min over token probabilities)
2. Bin by confidence
3. Compute average accuracy per bin
4. **Expected calibration error (ECE) is difference between confidence and accuracy**



$$ECE(\mathcal{B}) = \sum_{i=1}^{N} \frac{|\mathcal{B}_i|}{N} \left| \frac{\sum_{j \in \mathcal{B}_i} a_j}{|\mathcal{B}_i|} - \frac{\sum_{j \in \mathcal{B}_i} c_j}{|\mathcal{B}_i|} \right|$$

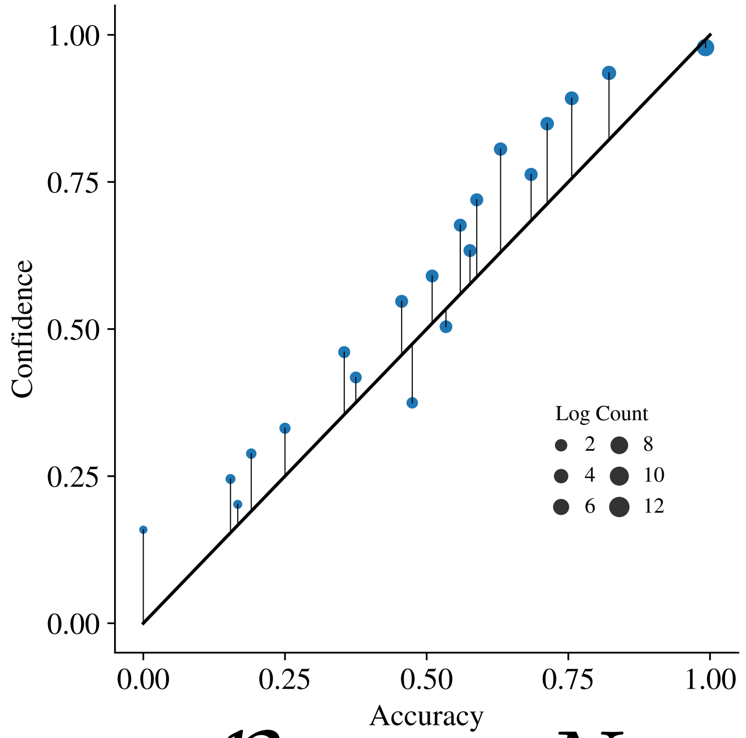
where $\mathcal{B}$ are the N bins,

a_i is the accuracy (0 or 1) and

c_j is the model confidence

Measuring Calibration Error

1. Obtain word-level confidence (min over token probabilities)
2. Bin by confidence
3. Compute average accuracy per bin
4. **Expected calibration error (ECE) is difference between confidence and accuracy**



$$ECE(\mathcal{B}) = \sum_{i=1}^{N} \frac{|\mathcal{B}_i|}{N} \left| \frac{\sum_{j \in \mathcal{B}_i} a_j}{|\mathcal{B}_i|} - \frac{\sum_{j \in \mathcal{B}_i} c_j}{|\mathcal{B}_i|} \right|$$

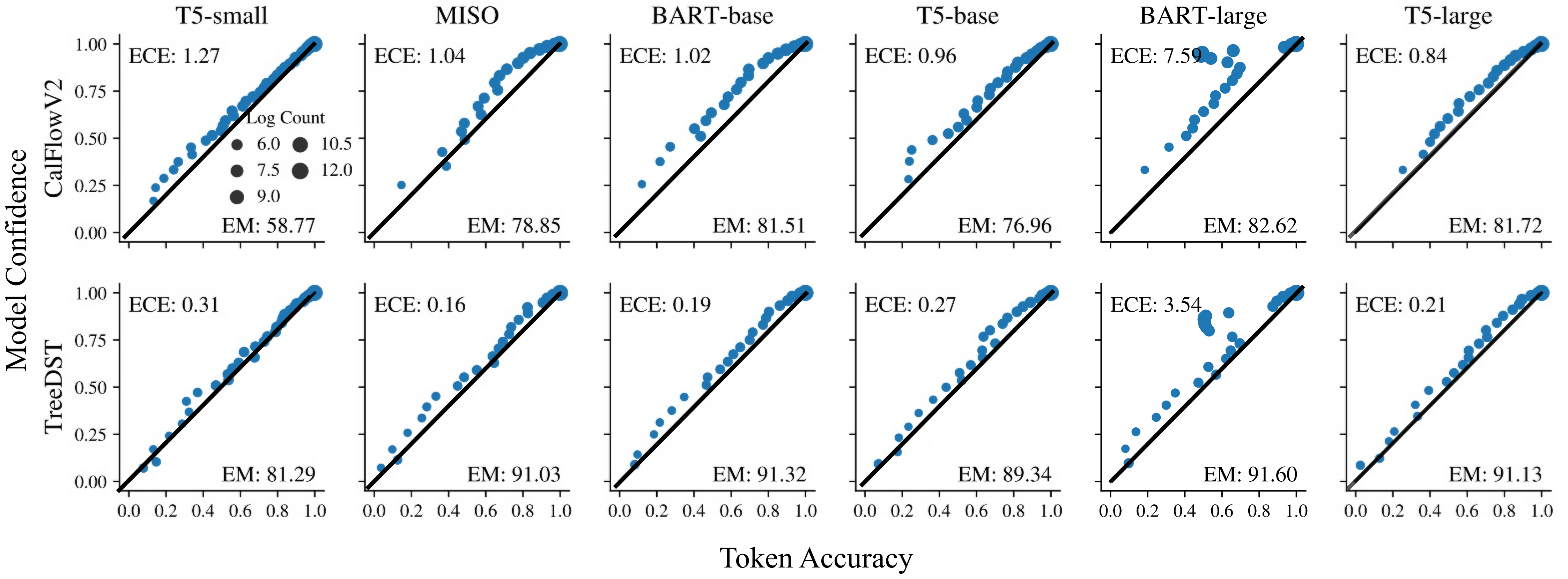
where $\mathcal{B}$ are the N bins,

a_i is the accuracy (0 or 1) and

c_j is the model confidence

How Calibrated Are Semantic Parsing Models?

How Calibrated Are Semantic Parsing Models?



Calibration in Semantic Parsing

Models surprisingly well-calibrated

Seq2seq + finetuning seems sufficient

Calibration in Semantic Parsing

Models surprisingly well-calibrated

More experiments/takeaways in the paper

- SQL vs digital assistant domains
- Execution accuracy
- Token frequency
- Perplexity and calibration
- Few-shot LLMs

Outline

Part I: Uncertainty in Human-Model Interactions

Calibrated Interpretation: Confidence Estimation in Semantic Parsing, Elias Stengel-Eskin and Benjamin Van Durme, TACL (2023)

Did You Mean...? Confidence-based Trade-offs in Semantic Parsing, Elias Stengel-Eskin and Benjamin Van Durme, EMNLP (2023)

Part II: Model-based Selection to Reduce Uncertainty

Rephrase, Augment, Reason: Visual Grounding of Questions for Vision-Language Models, Archiki Prasad, Elias Stengel-Eskin, Mohit Bansal, ICLR (2024)

Part III: Confidence for Model-Model Interactions

ReConcile: Round-Table Conference Improves Reasoning via Consensus among Diverse LLMs, Justin Chih-Yao Chen, Swarnadeep Saha, Mohit Bansal (2024)

MAGDi: Structured Distillation of Multi-Agent Interaction Graphs Improves Reasoning in Smaller Language Models, Justin Chih-Yao Chen*, Swarnadeep Saha*, Elias Stengel-Eskin Mohit Bansal (2024)

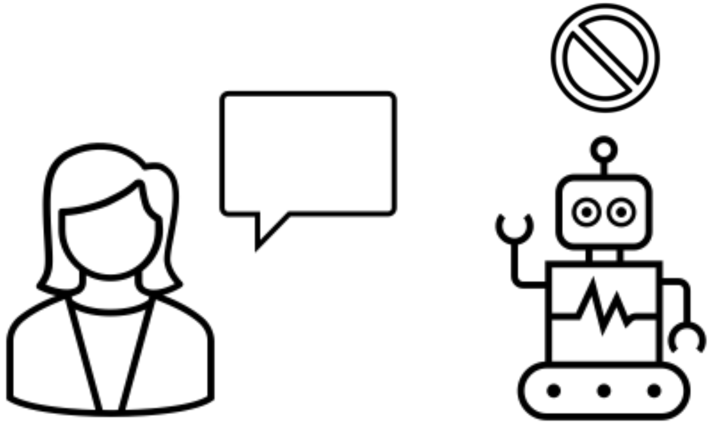
Selective Prediction

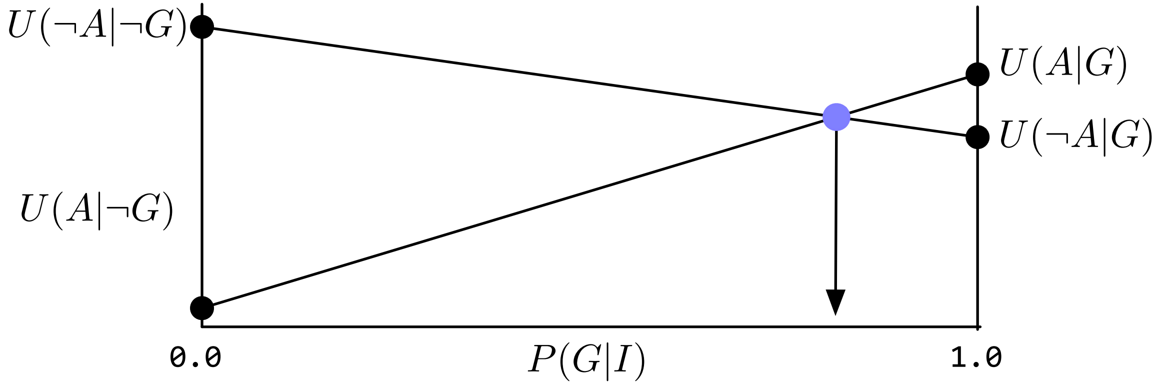
When do we accept or reject a prediction?

A = `drop_item()` $\neg A$ = `do_nothing()`

G = `drop_item` $\neg G$ = `do_nothing`

I = *let's drop it*





Selective classification for deep neural networks
Y. Geifman and R. El-Yaniv, 2017

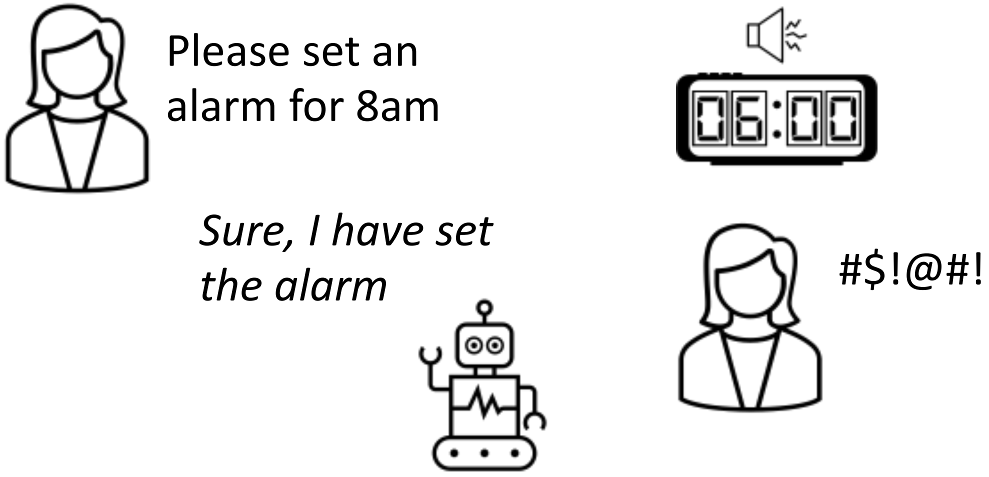
Problem: False rejection

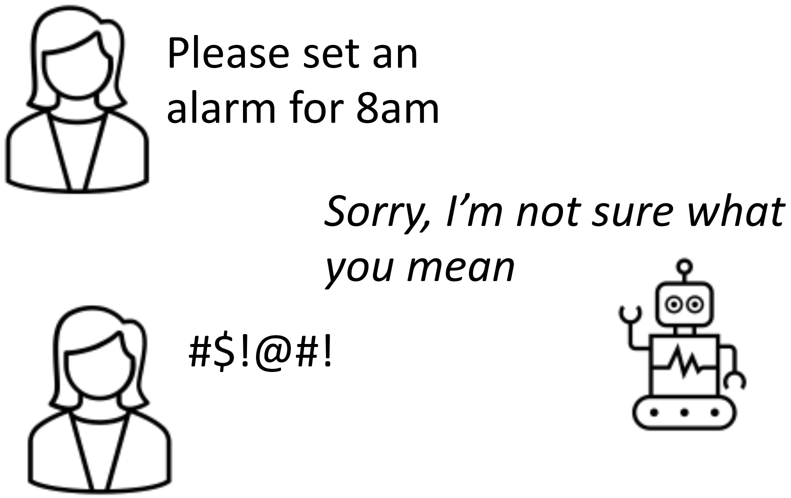
False positives harm safety

Unintended consequences

False negatives harm usability

Lack of action means increased frustration





Usability vs. Safety

Usability: executing instructions

Safety: not making mistakes

Full usability

Do I have anything going on today?

`(Yield (< (…`

Execute program

Full safety

Do I have anything going on today?

`(Yield (< (…`

Reject program

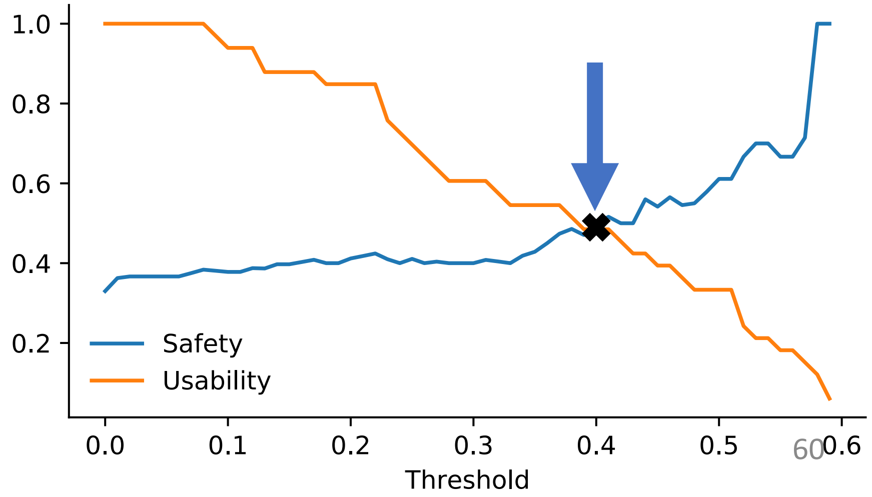
Threshold

Confidence < x

Reject program

Confidence >= x

Execute program



Usability vs. Safety

Threshold

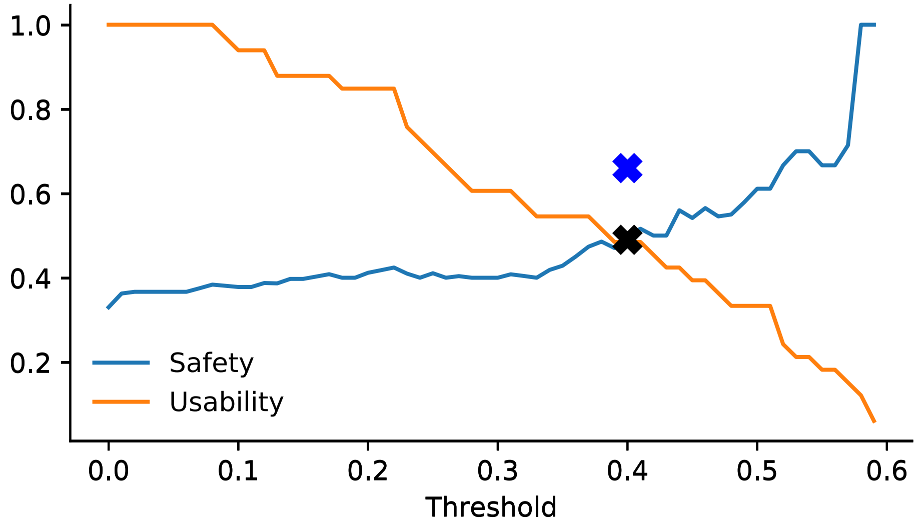
Confidence < x

Reject program

Confidence >= x

Execute program

Human-in-the-loop

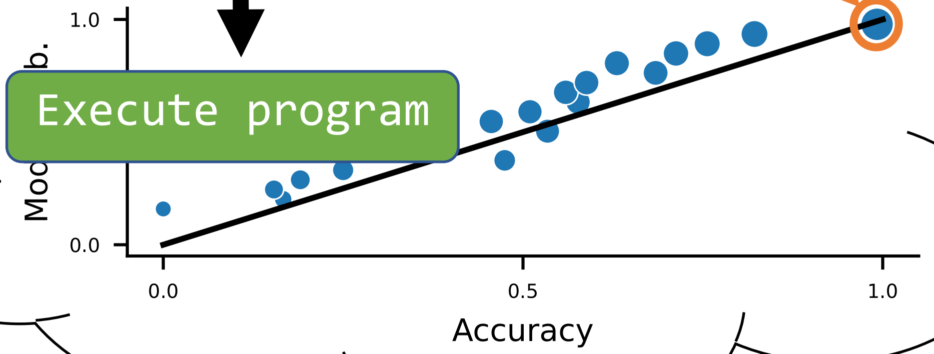


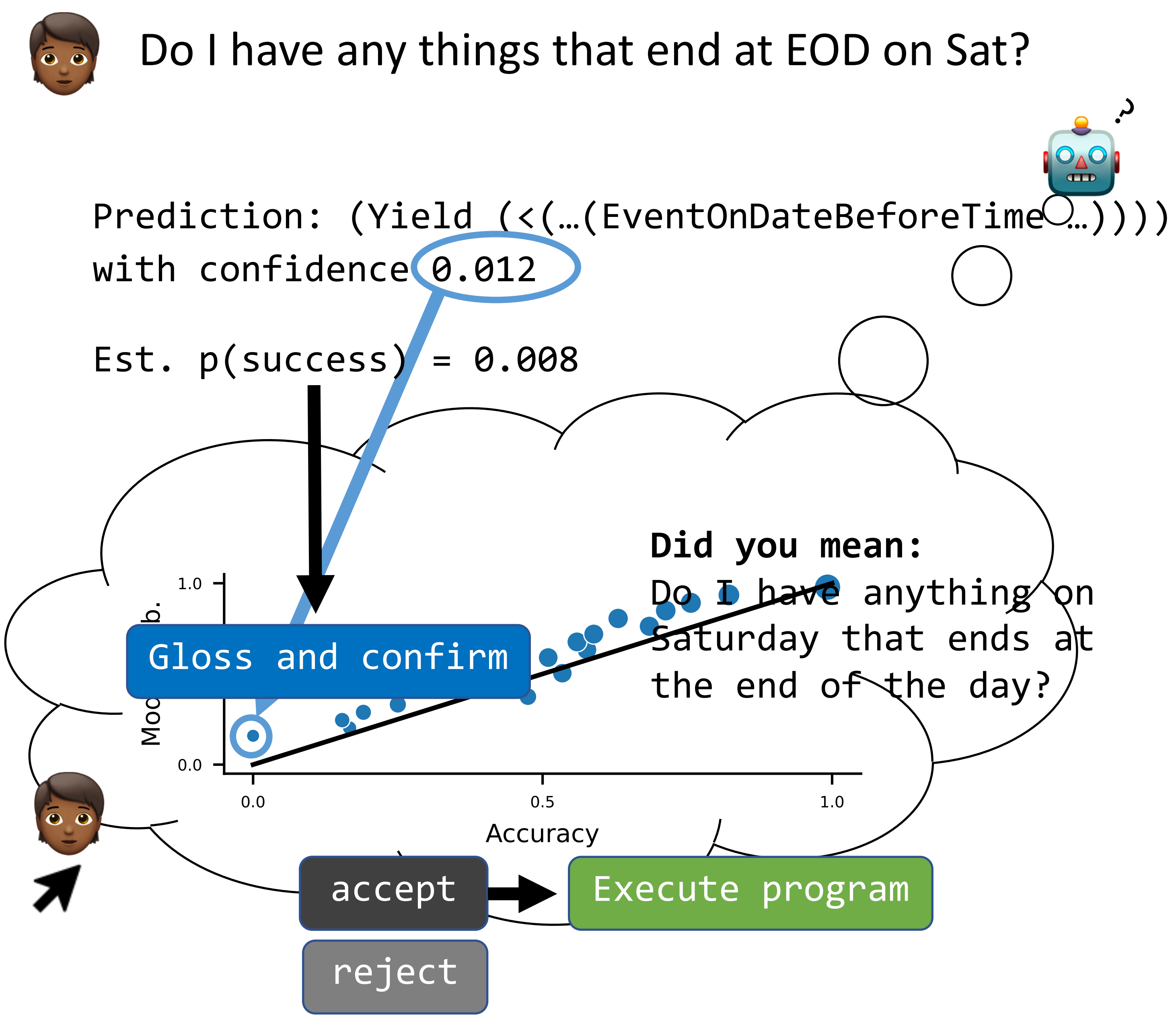
DidYouMean

Do I have anything going on tonight?

```
Prediction: (Yield (< (…(EventOnDateWithTimeRange…))))
with confidence 0.98

Est. p(success) = 0.983
```




Do I have any things that end at EOD on Sat?
Prediction: (Yield (<(…(EventOnDateBeforeTime(…))))
with confidence 0.012
Est. p(success) = 0.008
Did you mean:
Do I have anything on Saturday that ends at the end of the day?
Gloss and confirm
1.0
0.0
0.0
0.5
1.0
Accuracy
accept
Execute program
reject

DidYouMean details

Parsing model

Translates queries to programs

MISO (Stengel-Eskin et al., 2022)

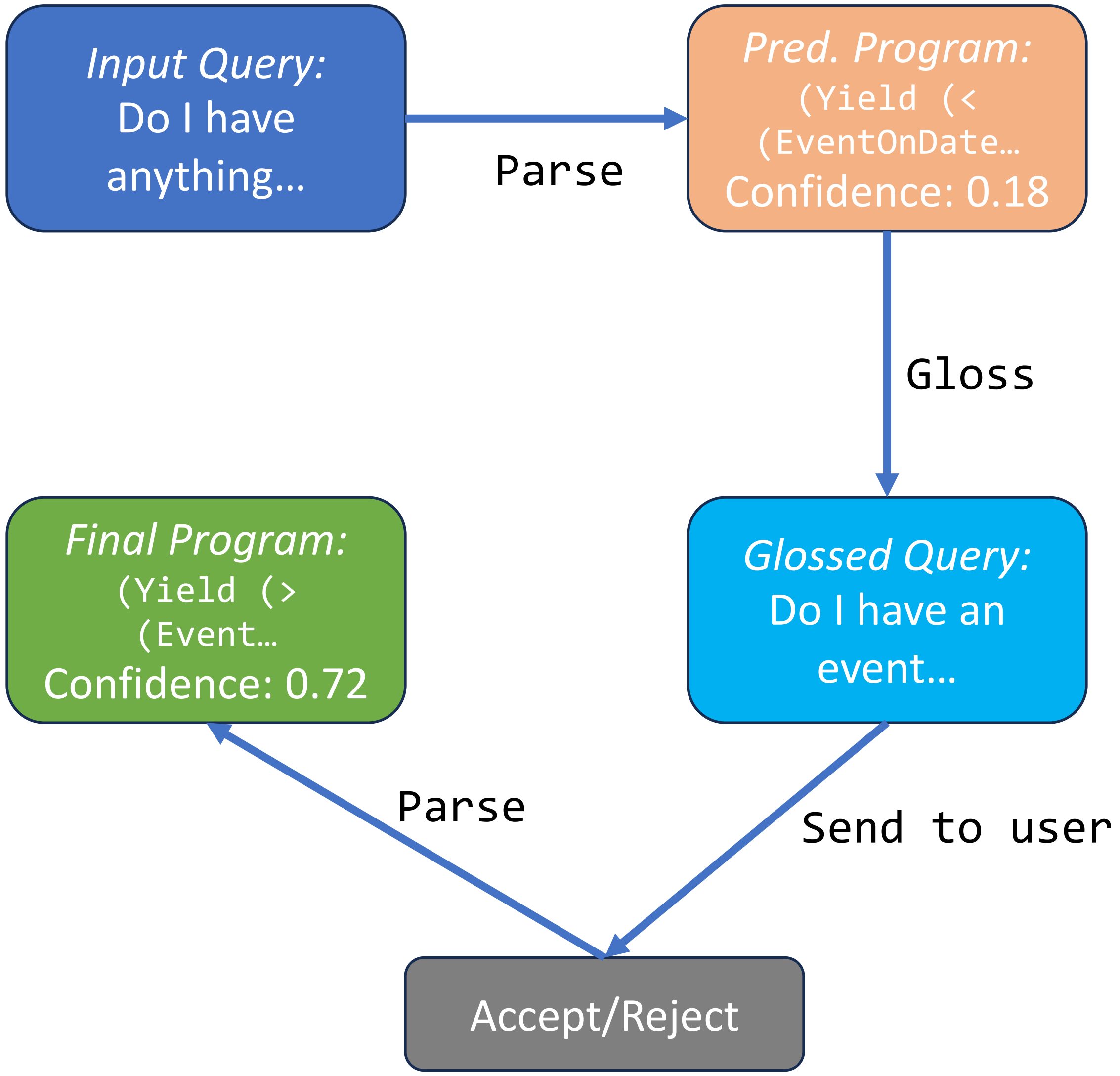
Finetuned seq2graph model

Glossing model

Translates programs to queries

BART-large seq2seq

Finetuned on reverse data



When More Data Hurts: A Troubling Quirk in Developing Broad-Coverage Natural Language Understanding Systems
Stengel-Eskin et al., 2022

DidYouMean Results

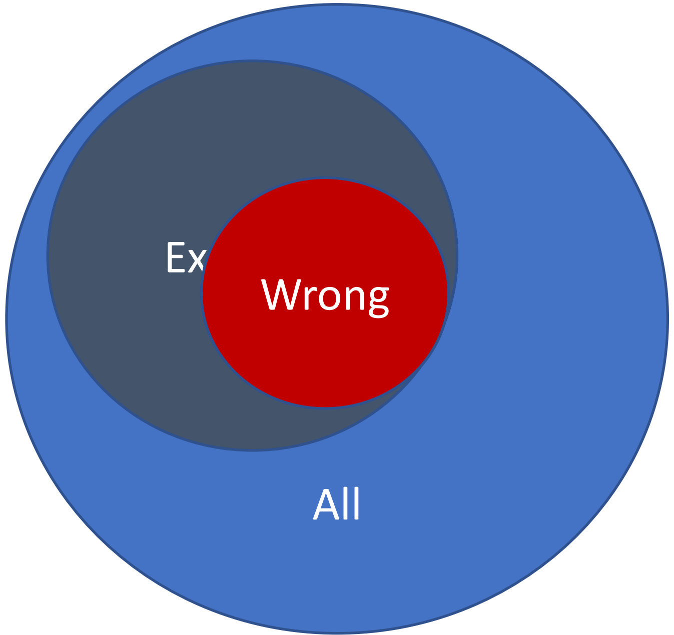
User study

Mturk annotators, 100 examples (below 0.6 confidence)

Coverage and Risk

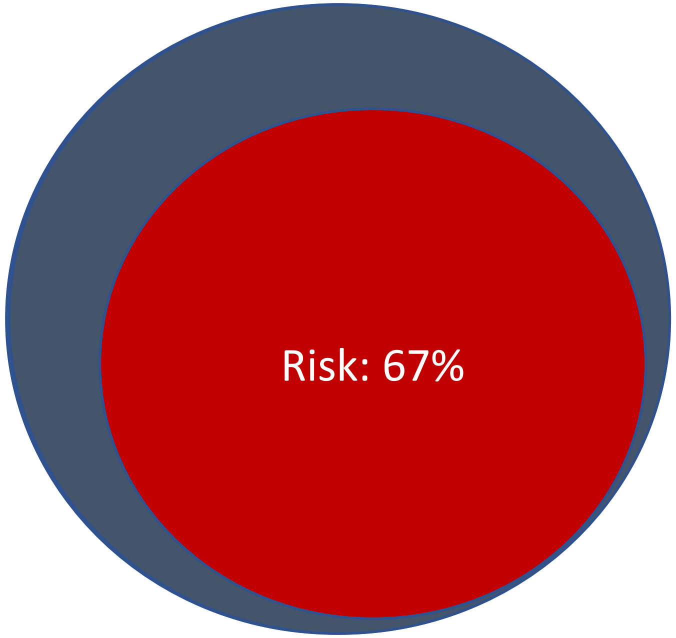
Coverage (usability): % of programs executed

Risk (safety): % of executed programs that were incorrect

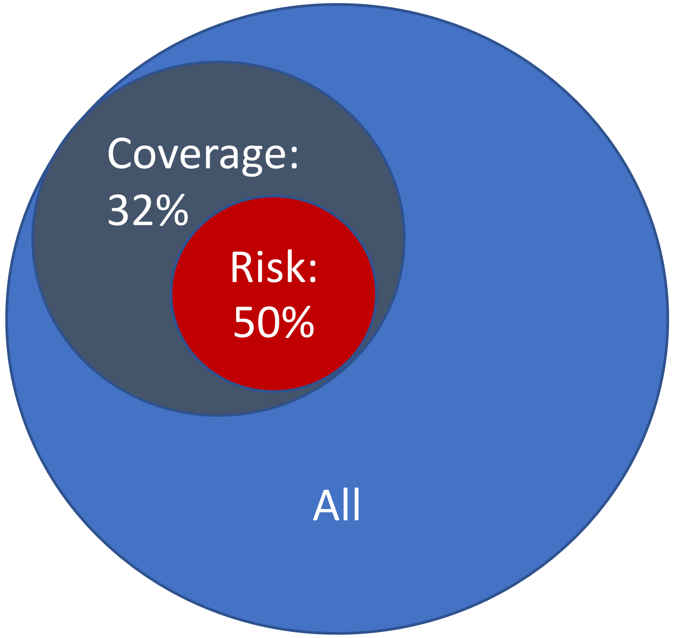


DidYouMean Results

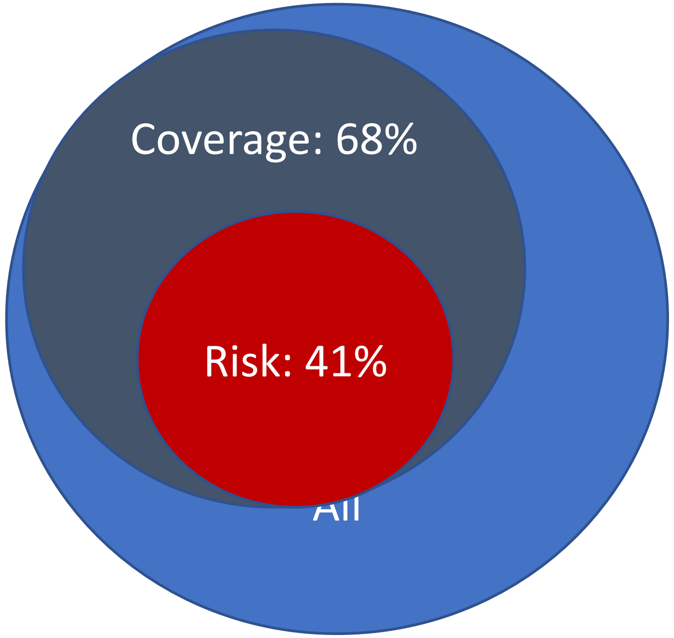
Baseline 1:
Accept everything



Baseline 2:
Threshold



DidYouMean



Takeaways

Calibration enables thresholding

Thresholding can improve safety

Human interaction improves balance

When uncertainty is high, defer to human judgment

Can we use the model to reduce uncertainty?

Outline

Part I: Uncertainty in Human-Model Interactions

Calibrated Interpretation: Confidence Estimation in Semantic Parsing, Elias Stengel-Eskin and Benjamin Van Durme, TACL (2023)

Did You Mean...? Confidence-based Trade-offs in Semantic Parsing, Elias Stengel-Eskin and Benjamin Van Durme, EMNLP (2023)

Part II: Model-based Selection to Reduce Uncertainty

Rephrase, Augment, Reason: Visual Grounding of Questions for Vision-Language Models, Archiki Prasad, Elias Stengel-Eskin, Mohit Bansal, ICLR (2024)

Part III: Confidence for Model-Model Interactions

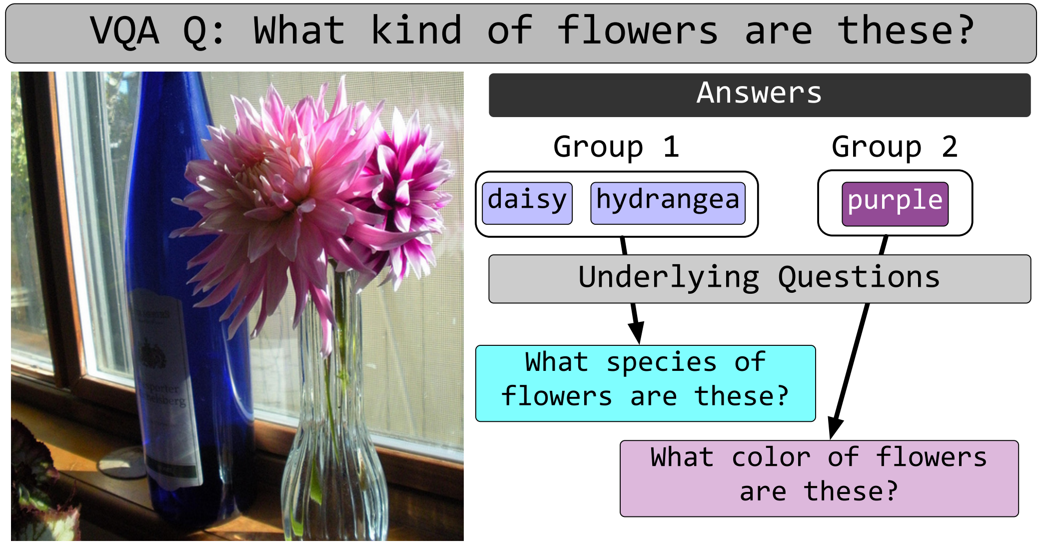
ReConcile: Round-Table Conference Improves Reasoning via Consensus among Diverse LLMs, Justin Chih-Yao Chen, Swarnadeep Saha, Mohit Bansal (2024)

MAGDi: Structured Distillation of Multi-Agent Interaction Graphs Improves Reasoning in Smaller Language Models, Justin Chih-Yao Chen*, Swarnadeep Saha*, Elias Stengel-Eskin Mohit Bansal (2024)

Uncertainty and Underspecification

VQA questions

May not provide sufficient information to answer correctly



Question: Where are we at?

Why did the chicken cross the road? Rephrasing and analyzing ambiguous questions in VQA
Stengel-Eskin et al., 2023

Why does a visual question have different answers?
Bhattacharya et al., 2019

Dealing with semantic underspecification in multimodal NLP
Pezzelle, 2023

RepARe: Rephrase, Augment, Reason

Rephrase questions to make them easier to answer

Augment questions with visual information

Reason about example and visual world

RepARe: Rephrase, Augment, Reason

Rephrase questions to make them easier to answer

Augment questions with visual information

Reason about example and visual world

Focus: zero-shot VQA with a large vision-language model

RepARe: Rephrase, Augment, Reason

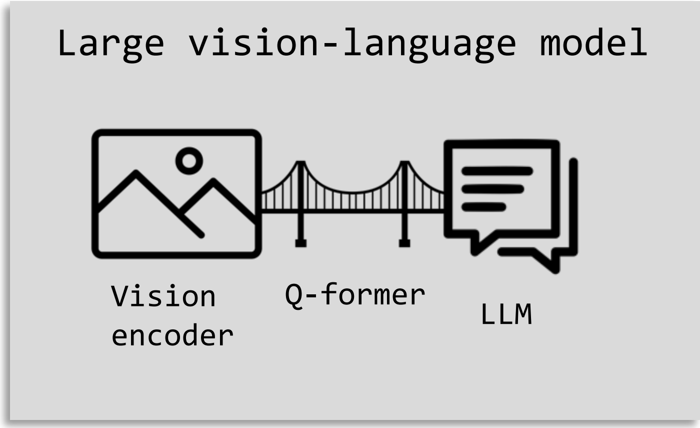
Rephrase questions to make them easier to answer

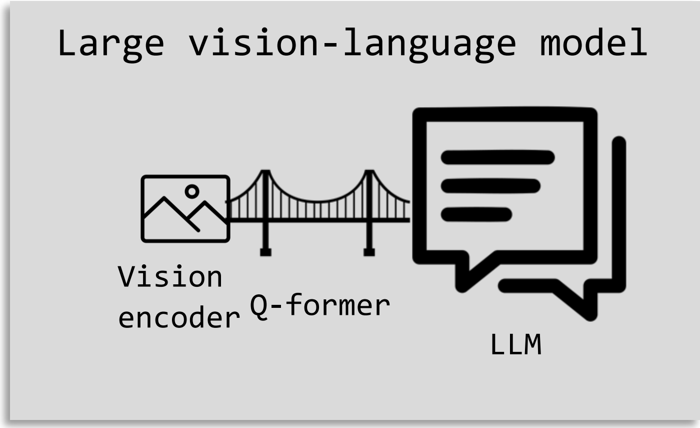
Augment questions with visual information

Reason about example and visual world

Focus: zero-shot VQA with a large vision-language model

Key insight: Asymmetric strength and asymmetric ability





QA ability:

Captioning ability:

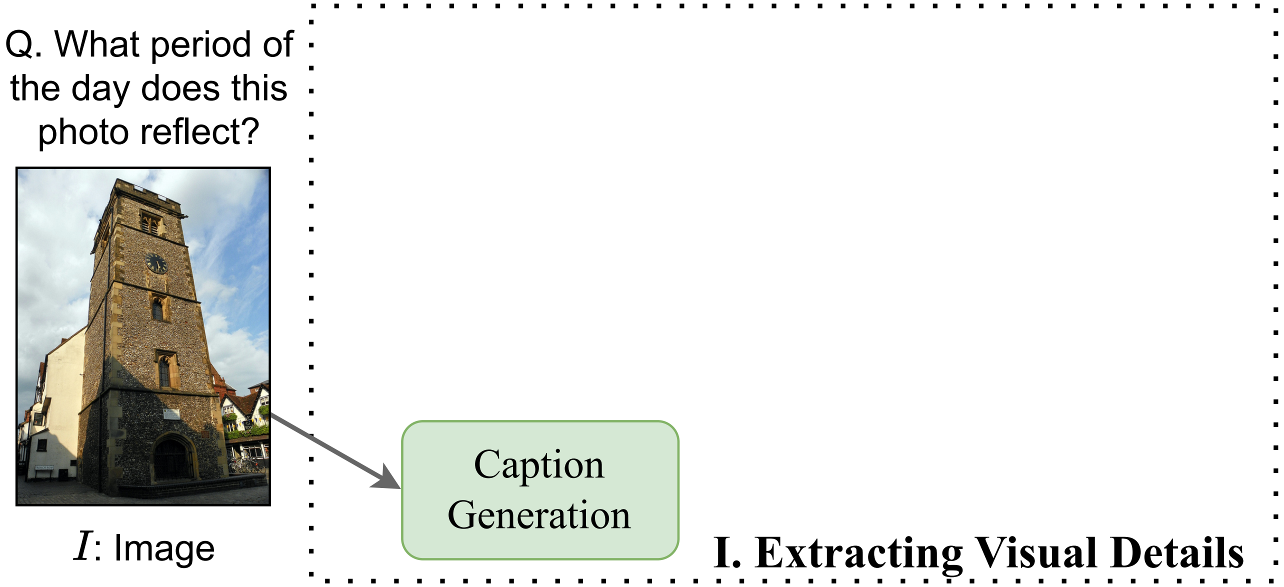
RepARe phases

Q. What period of the day does this photo reflect?

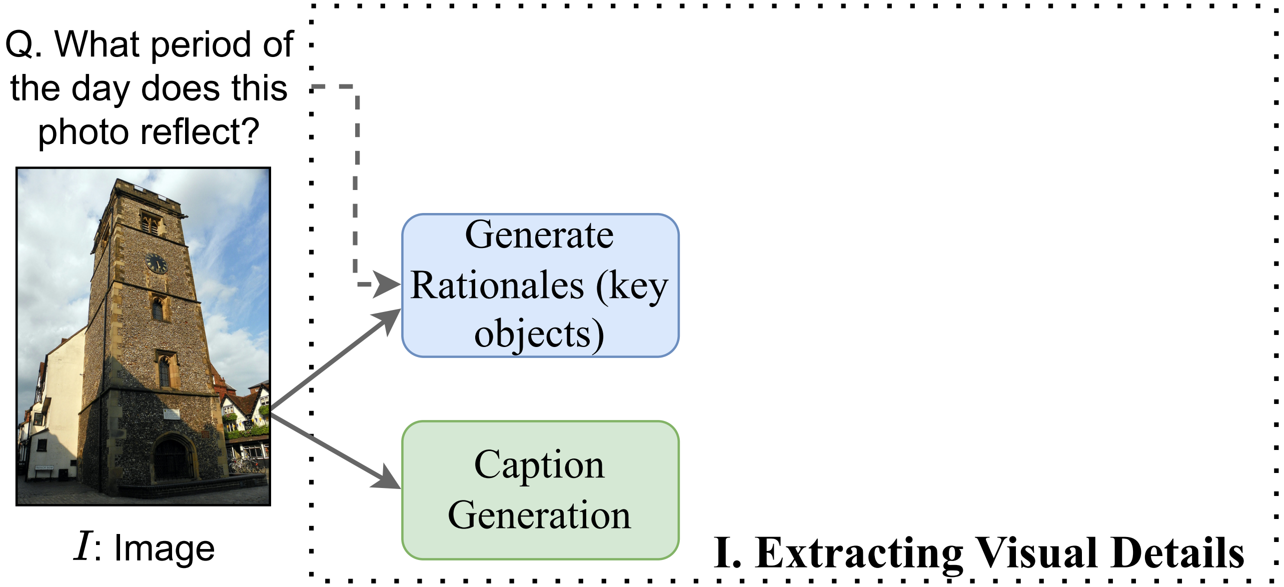
I: Image

I. Extracting Visual Details

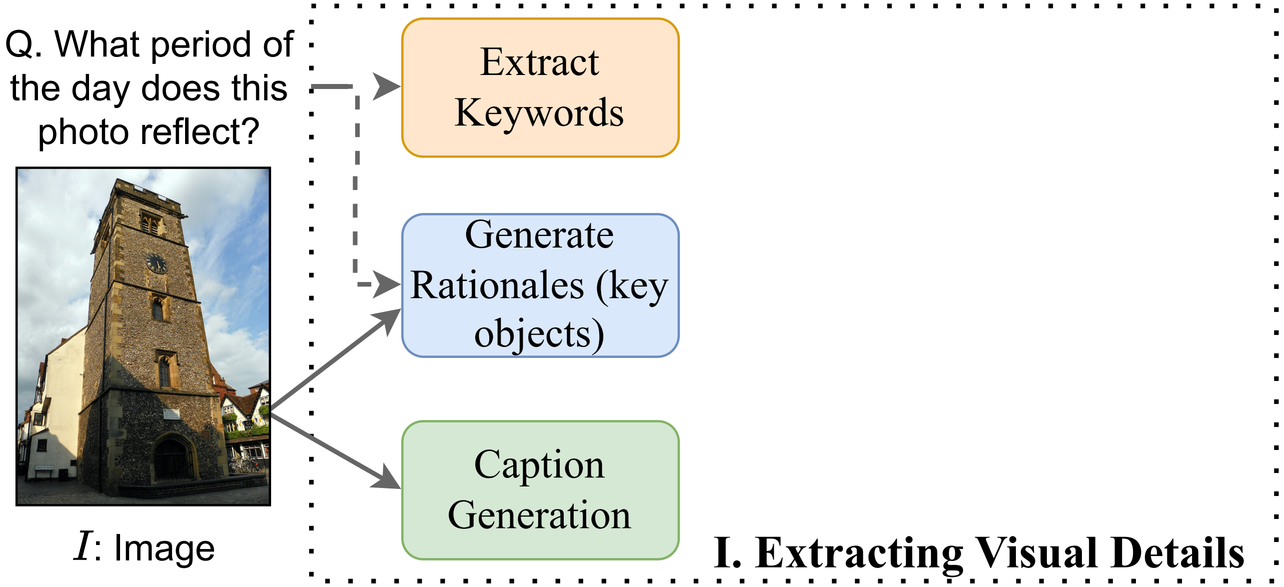
RepARe phases



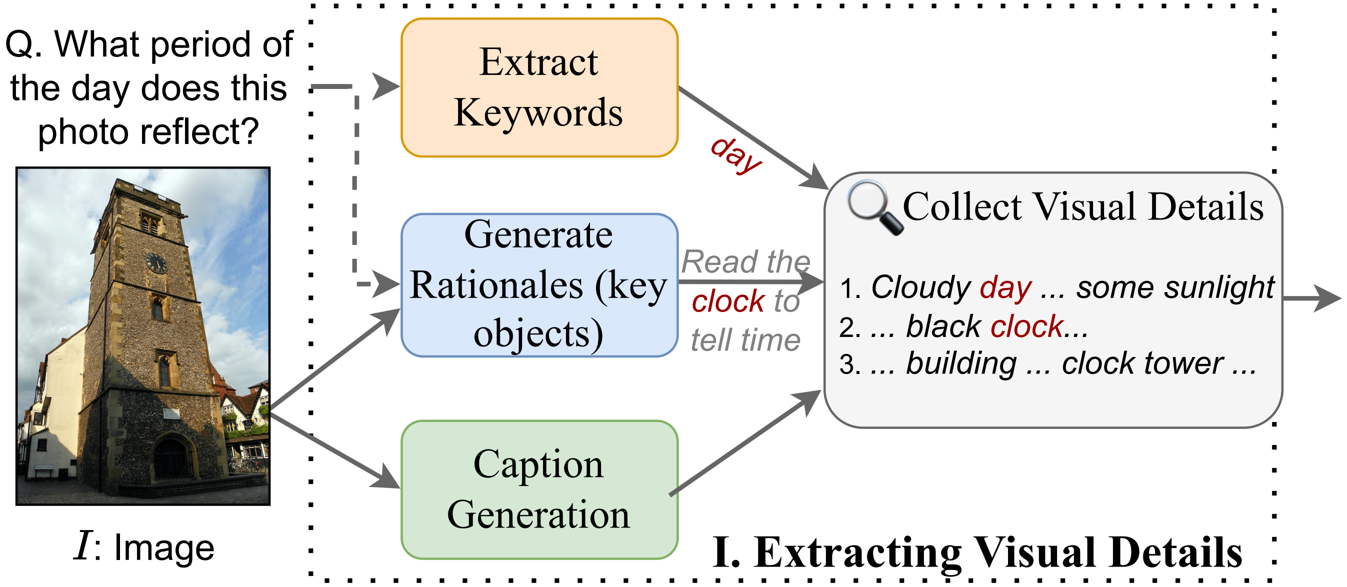
RepARe phases



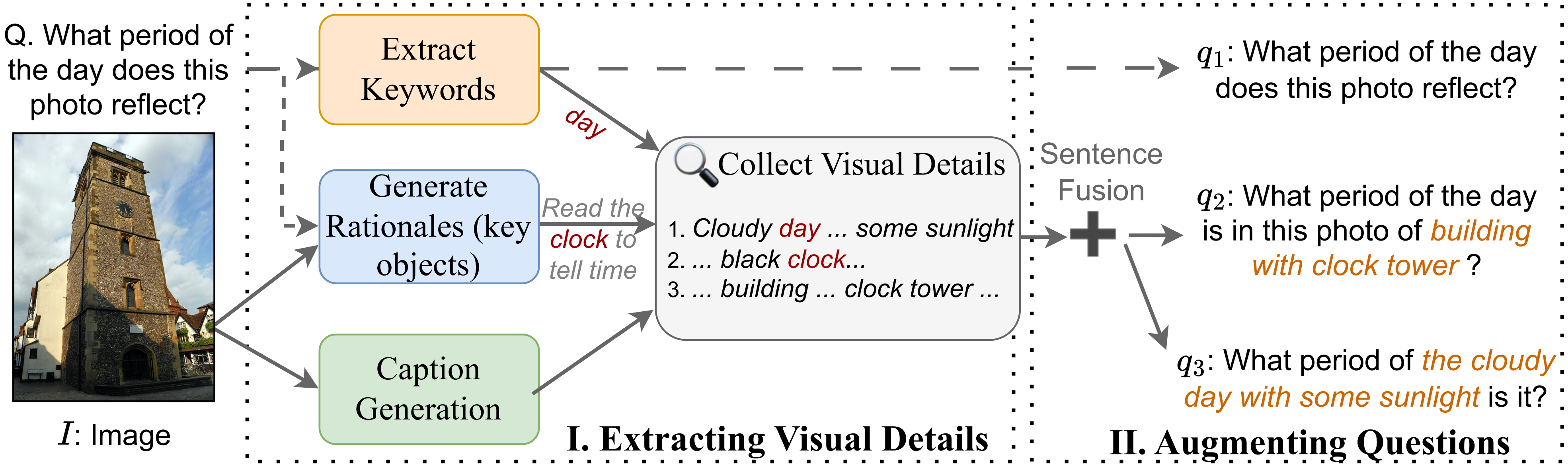
RepARe phases



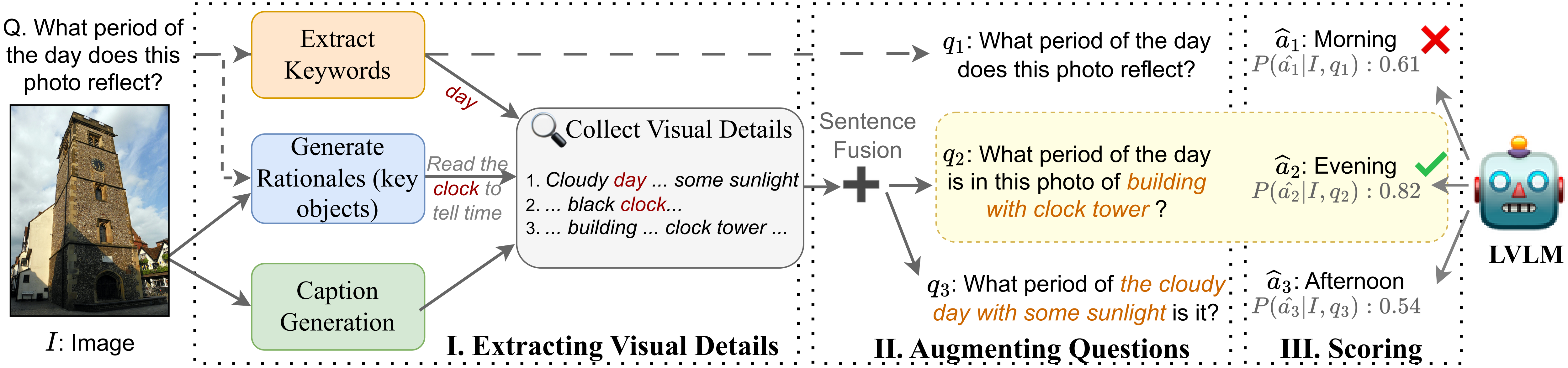
RepARe phases



RepARe phases



RepARe phases

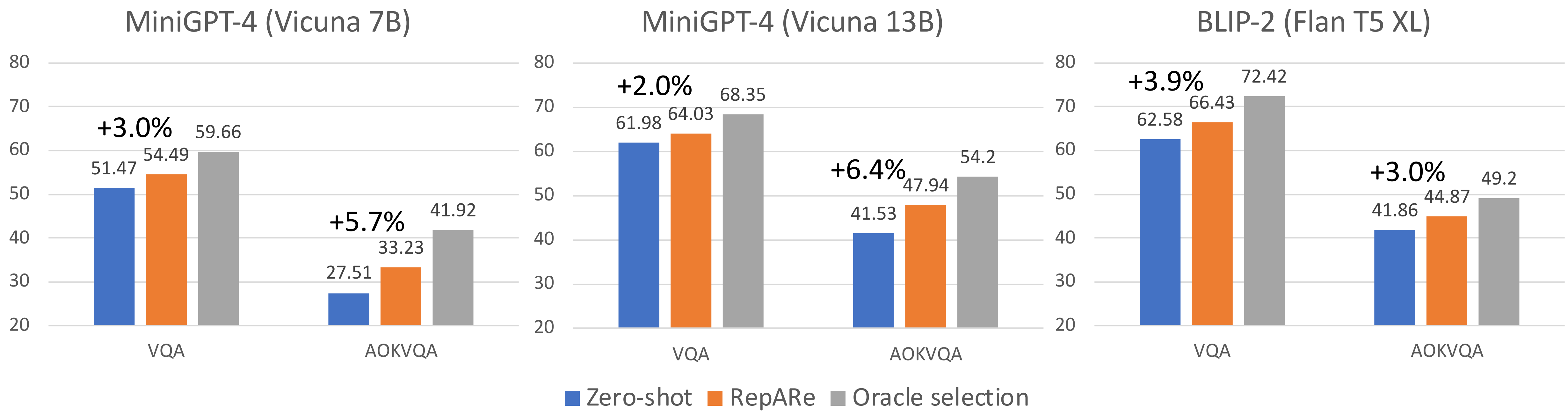


Chosen via ARGMAX over confidence

Choose the question that increases answer confidence the most

<u>i.e. question that reduces uncertainty</u>

Results



MiniGPT-4 (Vicuna 7B)
80
70
60
50
40
30
20
+3.0%
51.47
54.49
59.66
+5.7%
27.51
33.23
41.92
VQA
AOKVQA
MiniGPT-4 (Vicuna 13B)
80
70
60
50
40
30
20
+2.0%
61.98
64.03
68.35
+6.4%
41.53
47.94
54.2
VQA
AOKVQA
BLIP-2 (Flan T5 XL)
80
70
60
50
40
30
20
+3.9%
62.58
66.43
72.42
+3.0%
41.86
44.87
49.2
VQA
AOKVQA
Zero-shot
RepARe
Oracle selection

Qualitative Examples

Localization

Q: *Does the water have ripples?*

RepARe Q: Does the water have the small ripples around the boats?

Qualitative Examples

RepAR can help with reasoning

Q: *What will be built here one day?*

RepARe Q: What will be built at this construction site?

Outline

Part I: Uncertainty in Human-Model Interactions

Calibrated Interpretation: Confidence Estimation in Semantic Parsing, Elias Stengel-Eskin and Benjamin Van Durme, TACL (2023)

Did You Mean...? Confidence-based Trade-offs in Semantic Parsing, Elias Stengel-Eskin and Benjamin Van Durme, EMNLP (2023)

Part II: Model-based Selection to Reduce Uncertainty

Rephrase, Augment, Reason: Visual Grounding of Questions for Vision-Language Models, Archiki Prasad, Elias Stengel-Eskin, Mohit Bansal, ICLR (2024)

Part III: Confidence for Model-Model Interactions

ReConcile: Round-Table Conference Improves Reasoning via Consensus among Diverse LLMs, Justin Chih-Yao Chen, Swarnadeep Saha, Mohit Bansal (2024)

MAGDi: Structured Distillation of Multi-Agent Interaction Graphs Improves Reasoning in Smaller Language Models, Justin Chih-Yao Chen*, Swarnadeep Saha*, Elias Stengel-Eskin Mohit Bansal (2024)

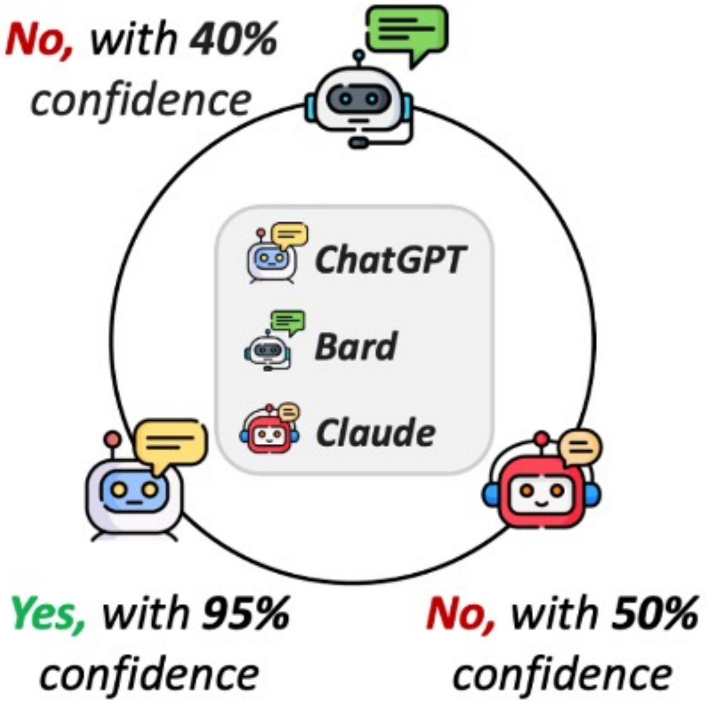
Overview

LLMs struggle with complex reasoning

ReConcile: A multi-model and multi-agent framework

Key components:

- Multi-LLM discussion
- Multiple discussion rounds
- Learning to convince other agents
- Confidence-weighted voting



Three stages of discussion

1

Initial response generation

Obtain initial response from each model (ChatGPT, Claude and Bard)

Three stages of discussion

1

Initial response generation

Obtain initial response from each model (ChatGPT, Claude and Bard)

2

multi-round discussion

Aggregate the information by grouping + **confidence estimation**

Example of grouping: `There are` **2** `agents think the answer is` **no**`, and` **1** `agent thinks the answer is` **yes**`.`

Three stages of discussion

1

Initial response generation

Obtain initial response from each model (ChatGPT, Claude and Bard)

2

multi-round discussion

Aggregate the information by grouping + **confidence estimation**

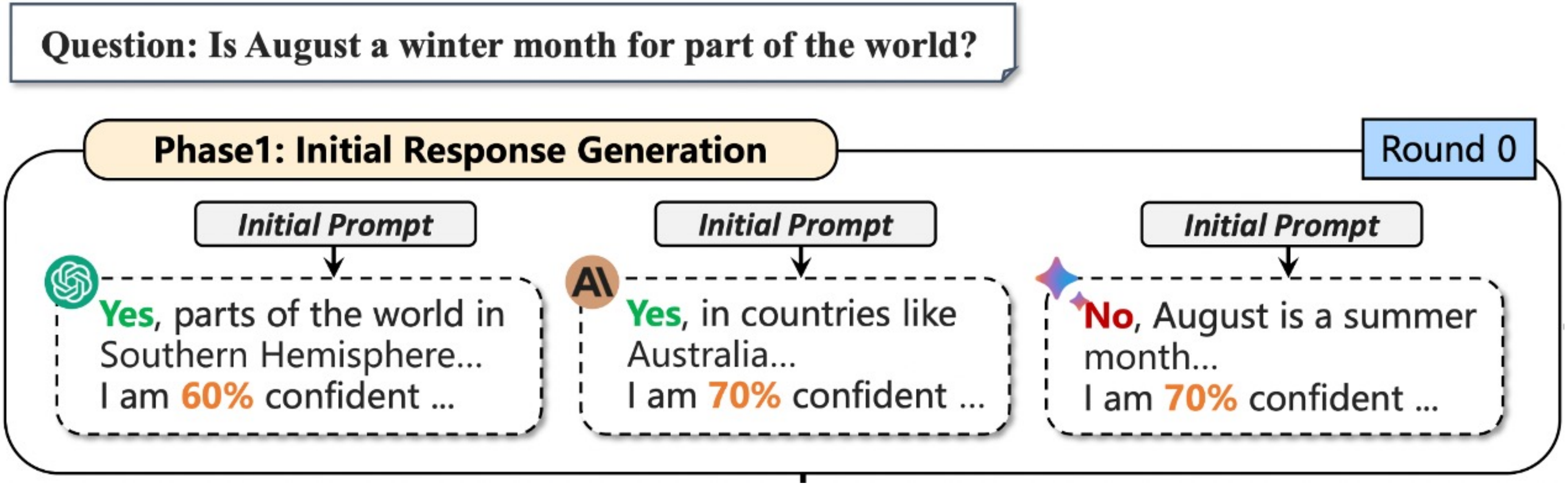
3

final answer generation

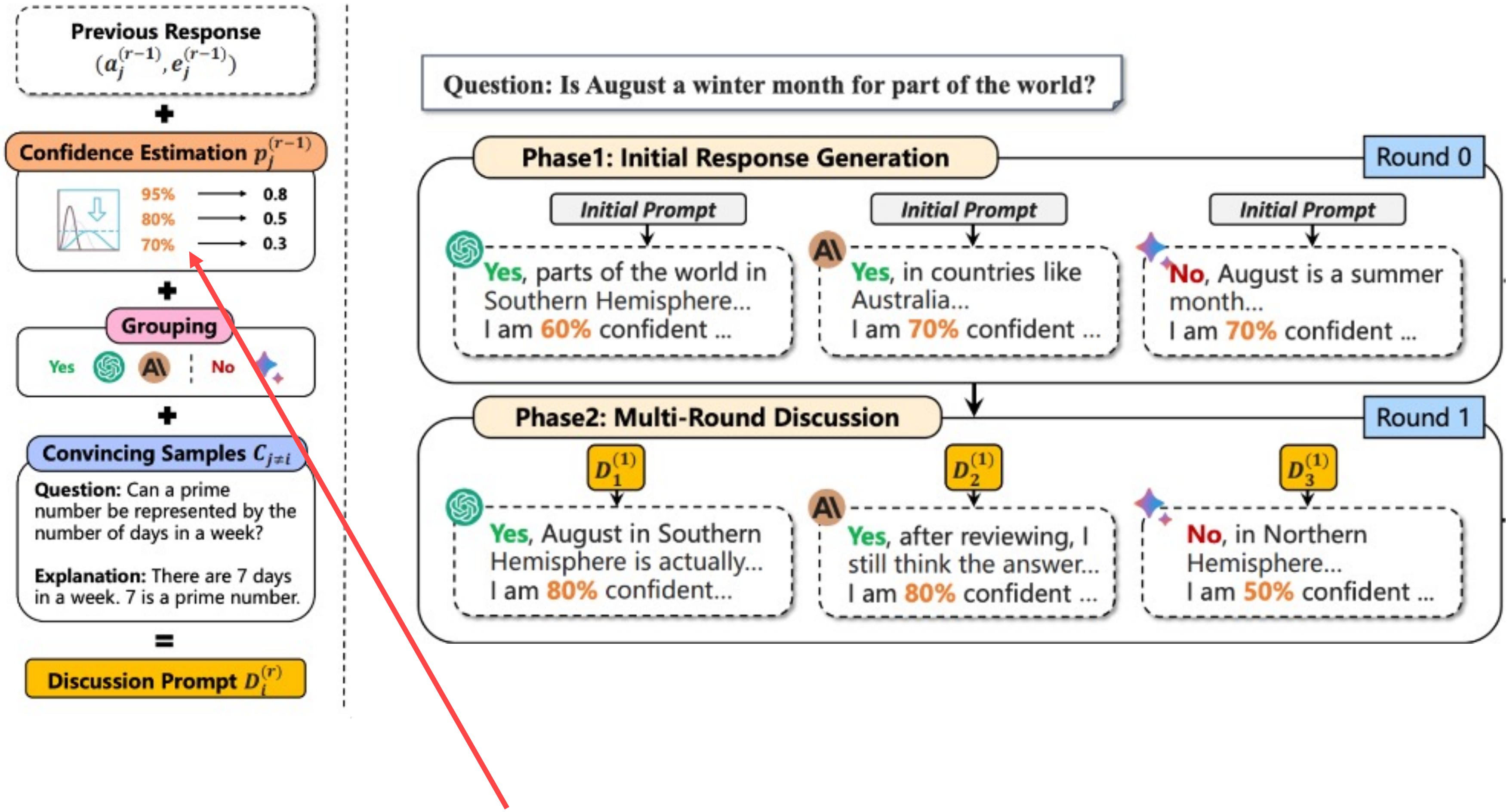
Confidence-weighted vote to derive the final answer

Example of grouping: `There are` **2** `agents think the answer is` **no**`, and` **1** `agent thinks the answer is` **yes**`.`

Phase 1: Initial response

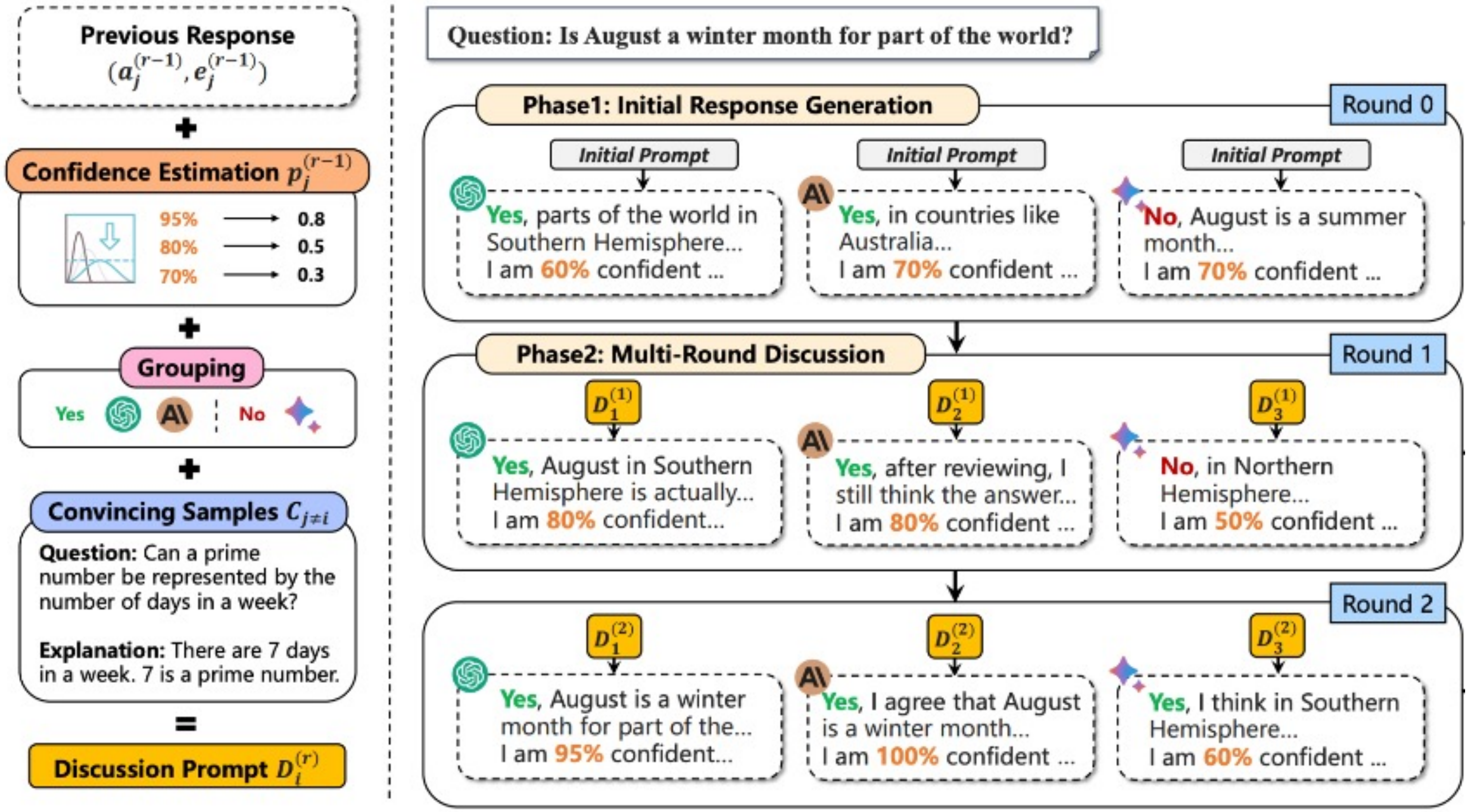


Phase 2: Multi-round discussion

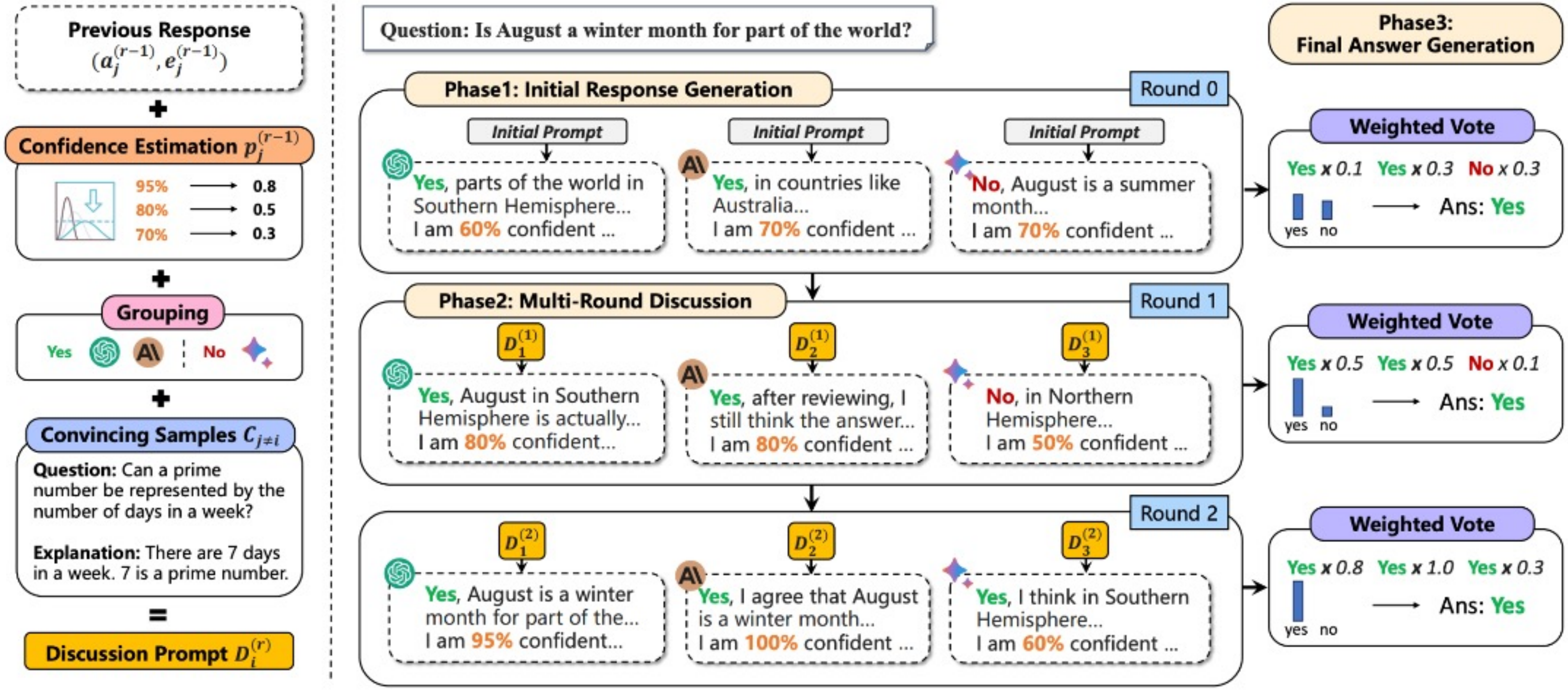


Corrected confidence scores

Phase 2: Multi-round discussion



Phase 3: Answer generation

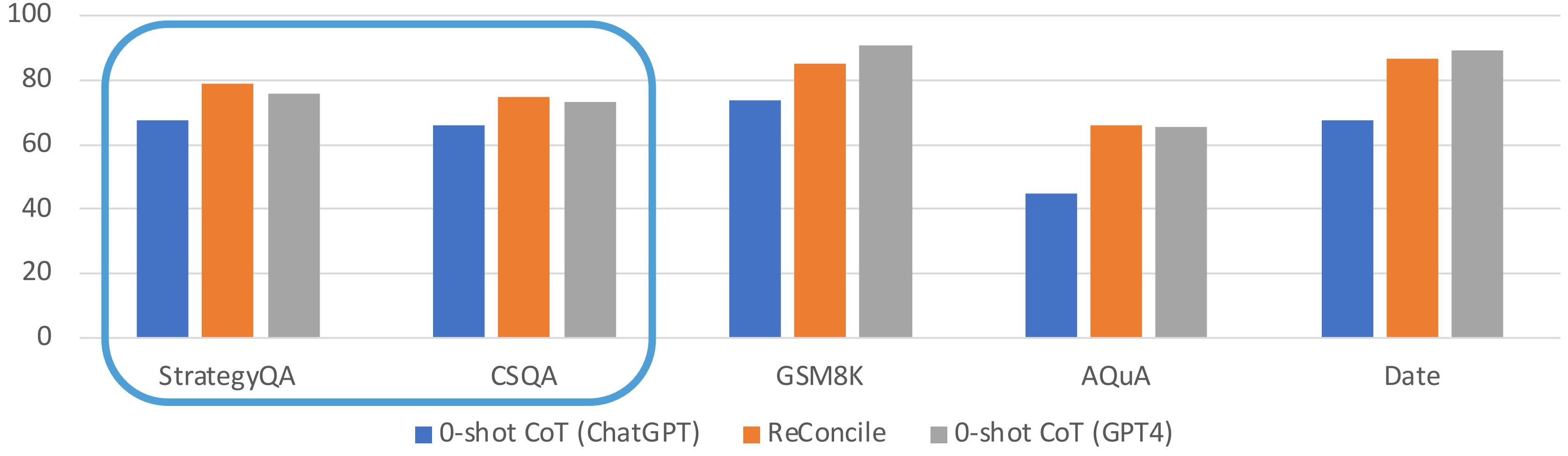


Results

Most powerful / expensive model considered

Method Category	Method	Agent	StrategyQA	CSQA	GSM8K	AQuA	Date
Vanilla Single-agent	Zero-shot CoT	GPT-4	*75.6±4.7*	*73.3±0.4*	*90.7±1.7*	*65.7±4.6*	*89.0±2.2*
	Zero-shot CoT	ChatGPT	67.3±3.6	66.0±1.8	73.7±3.1	44.7±0.5	67.7±1.2
	Zero-shot CoT	Bard	69.3±4.4	56.8±2.7	58.7±2.6	33.7±1.2	50.2±2.2
	Zero-shot CoT	Claude2	73.7±3.1	66.7±2.1	79.3±3.6	60.3±1.2	78.7±2.1
Advanced Single-agent	Self-Refine (SR)	ChatGPT	66.7±2.7	68.1±1.8	74.3±2.5	45.3±2.2	66.3±2.1
	Self-Consistency (SC)	ChatGPT	73.3±2.1	70.9±1.3	80.7±1.2	54.0±2.9	69.0±0.8
	SR + SC	ChatGPT	72.2±1.9	71.9±2.1	81.3±1.7	58.3±3.7	68.7±1.2
Single-model Multi-agent	Debate	×3	66.7±3.1	62.7±1.2	83.0±2.2	65.3±3.1	68.0±1.6
	Debate	×3	65.3±2.5	66.3±2.1	56.3±1.2	29.3±4.2	46.0±2.2
	Debate	×3	71.3±2.2	68.3±1.7	70.7±4.8	62.7±2.6	75.3±3.3
	Debate+Judge	×3	69.7±2.1	63.7±2.5	74.3±2.9	57.3±2.1	67.7±0.5
Multi-model Multi-agent	RECONCILE	, ,	**79.0**±1.6	**74.7**±0.4	**85.3**±2.2	**66.0**±0.8	**86.7**±1.2

ReConcile w/o GPT4 outperforms it!



Outline

Part I: Uncertainty in Human-Model Interactions

Calibrated Interpretation: Confidence Estimation in Semantic Parsing, Elias Stengel-Eskin and Benjamin Van Durme, TACL (2023)

Did You Mean...? Confidence-based Trade-offs in Semantic Parsing, Elias Stengel-Eskin and Benjamin Van Durme, EMNLP (2023)

Part II: Model-based Selection to Reduce Uncertainty

Rephrase, Augment, Reason: Visual Grounding of Questions for Vision-Language Models, Archiki Prasad, Elias Stengel-Eskin, Mohit Bansal, ICLR (2024)

Part III: Confidence for Model-Model Interactions

ReConcile: Round-Table Conference Improves Reasoning via Consensus among Diverse LLMs, Justin Chih-Yao Chen, Swarnadeep Saha, Mohit Bansal (2024)

MAGDi: Structured Distillation of Multi-Agent Interaction Graphs Improves Reasoning in Smaller Language Models, Justin Chih-Yao Chen*, Swarnadeep Saha*, Elias Stengel-Eskin Mohit Bansal (2024)

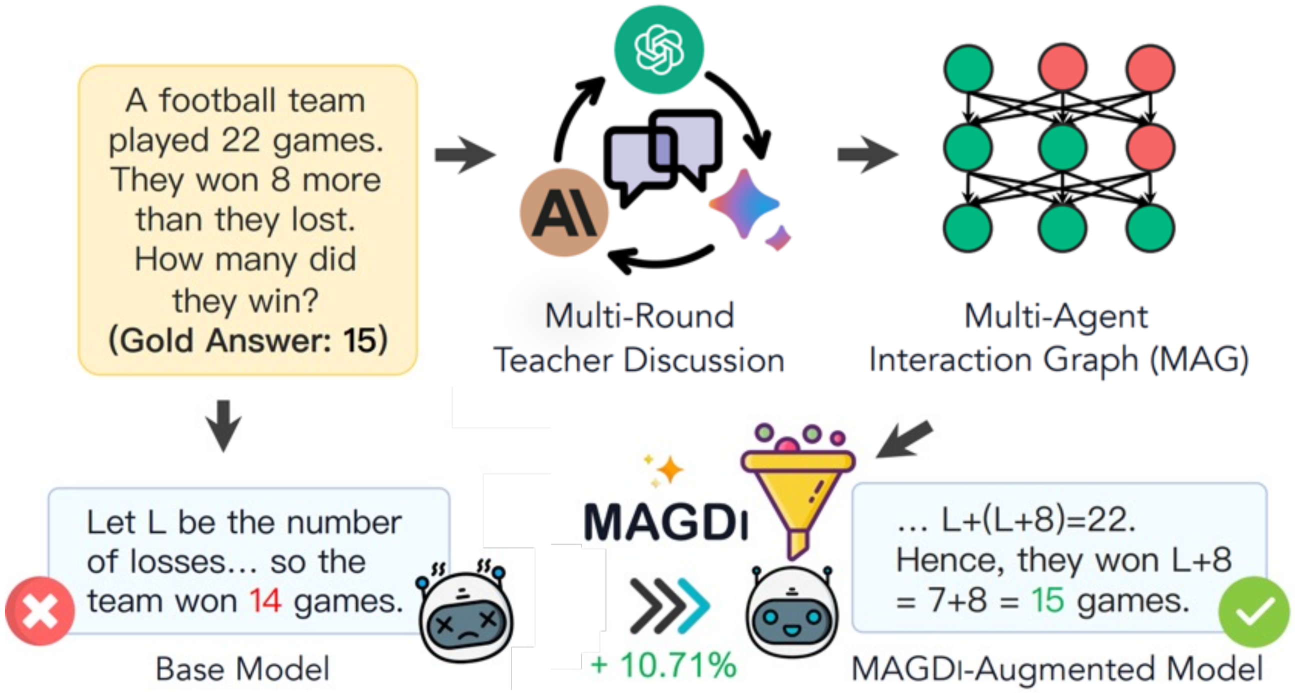
Can we distill ReConcile into a single model?

Strong performance boost but…

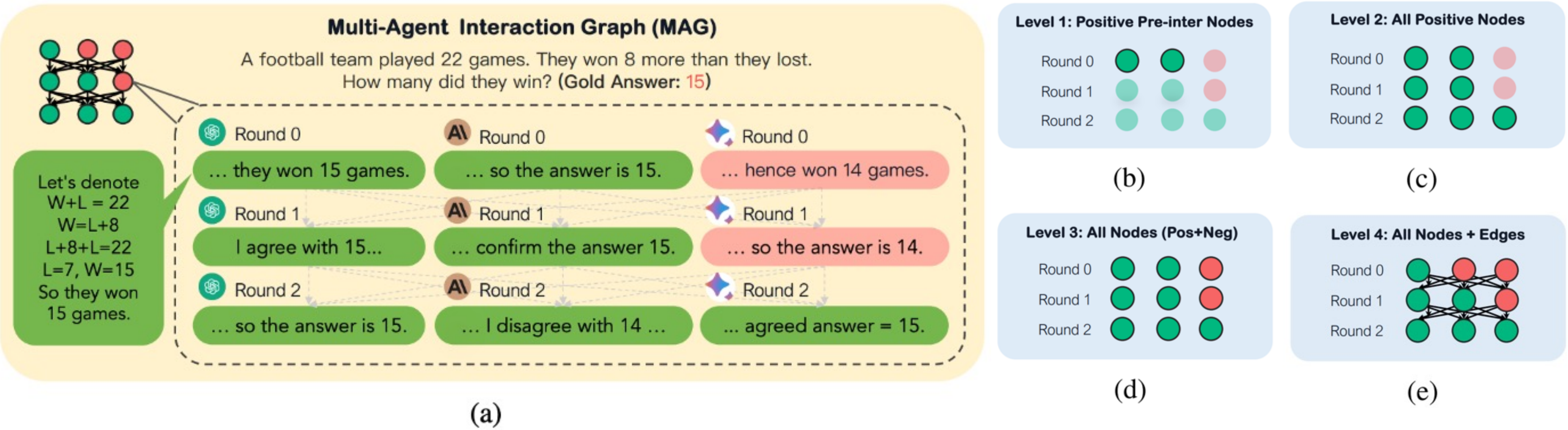
Multiple LLMs across multiple rounds

Expensive!

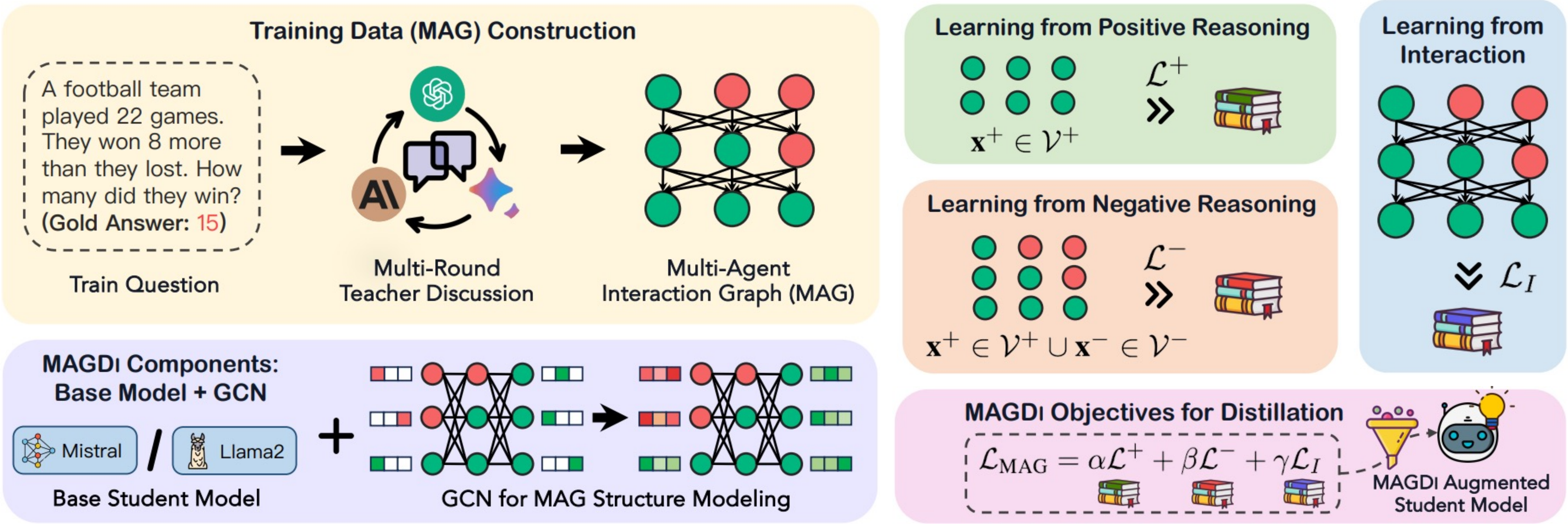
Open-source LLMs



Levels of distillation



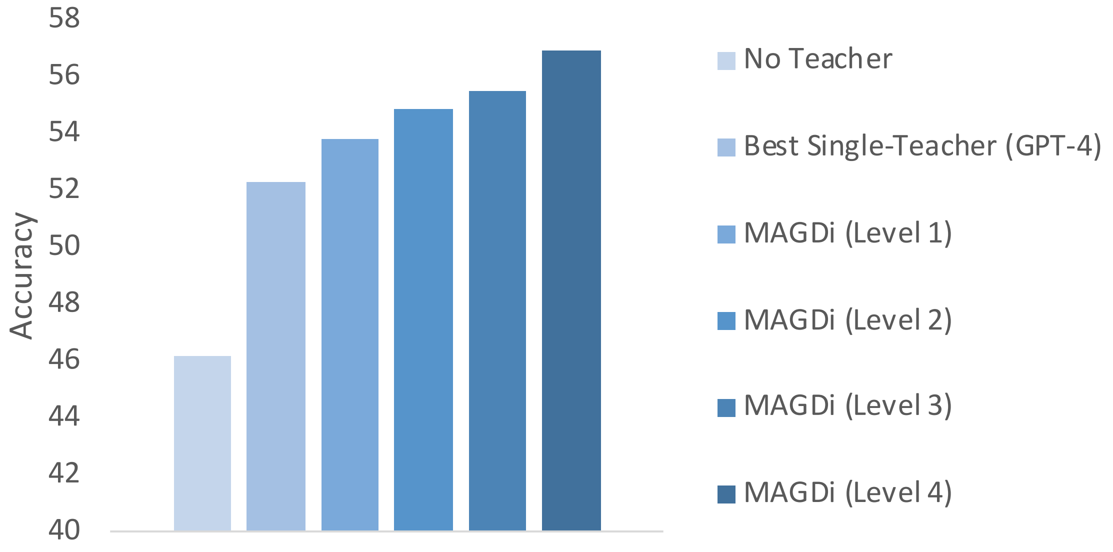
Details



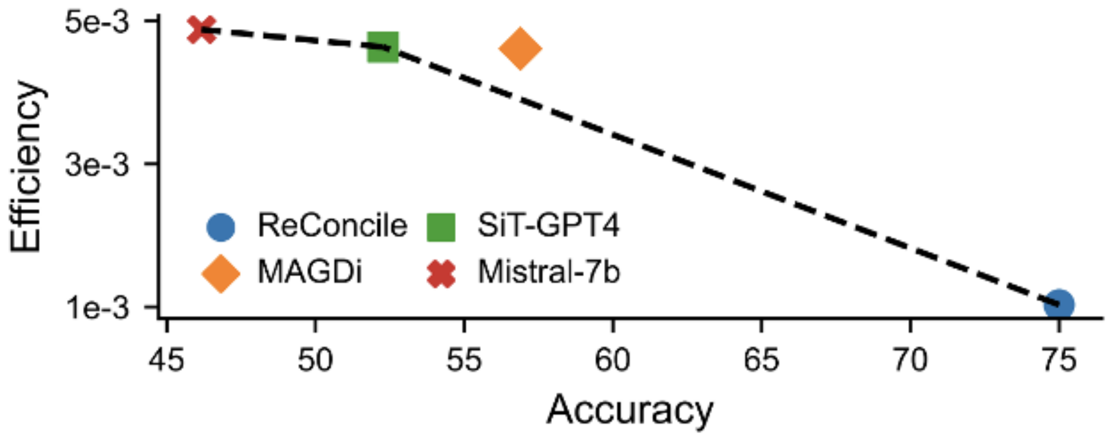
Training Data (MAG) Construction
A football team played 22 games. They won 8 more than they lost. How many did they win? (Gold Answer: 15)
Train Question
Multi-Round Teacher Discussion
Multi-Agent Interaction Graph (MAG)
Learning from Positive Reasoning
$\mathcal{L}^+$
$\mathbf{x}^+ \in \mathcal{V}^+$
Learning from Negative Reasoning
$\mathcal{L}^-$
$\mathbf{x}^+ \in \mathcal{V}^+ \cup \mathbf{x}^- \in \mathcal{V}^-$
Learning from Interaction
$\mathcal{L}_I$
MAGDi Components: Base Model + GCN
Mistral
Llama2
Base Student Model
GCN for MAG Structure Modeling
MAGDi Objectives for Distillation
$\mathcal{L}_{\text{MAG}} = \alpha\mathcal{L}^+ + \beta\mathcal{L}^- + \gamma\mathcal{L}_I$
MAGDi Augmented Student Model

Results

Mistral-7B average across StrategyQA, CSQA, ARC, GSM8K, MATH



Best tradeoff between efficiency and performance



Conclusions

Part I: Confidence and human-model interaction

Calibration in semantic parsing

What can we do when interacting with calibrated models?

Conclusions

Part I: Confidence and human-model interaction

Calibration in semantic parsing

What can we do when interacting with calibrated models?

Part II: Confidence for selection

Improving zero-shot VQA by reducing uncertainty

Conclusions

Part I: Confidence and human-model interaction

Calibration in semantic parsing

What can we do when interacting with calibrated models?

Part II: Confidence for selection

Improving zero-shot VQA by reducing uncertainty

Part III: Confidence for model-model interaction

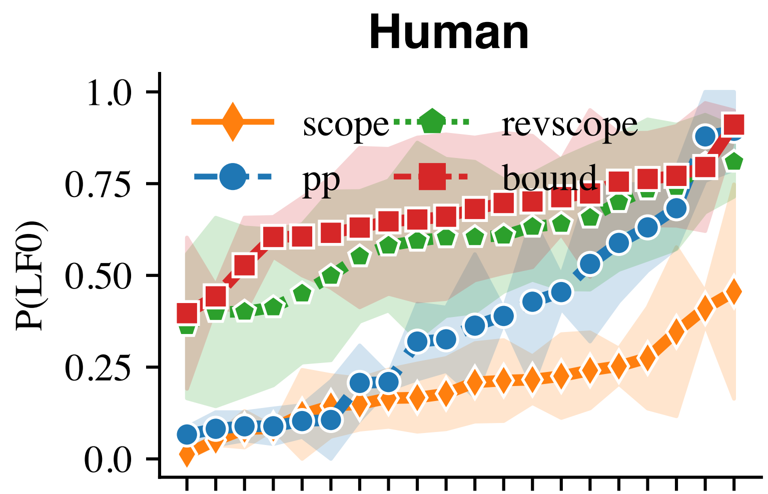
Improving reasoning via LLM collaboration

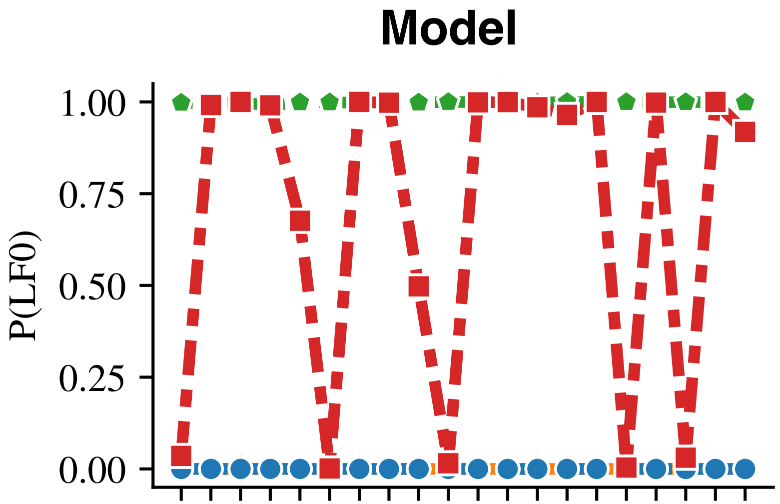
Future directions

Dealing with ambiguity

Semantic vs. form uncertainty

Models do not capture human uncertainty





Zero and few-shot semantic parsing with ambiguous inputs
Stengel-Eskin et al., 2023

Future directions

Dealing with ambiguity

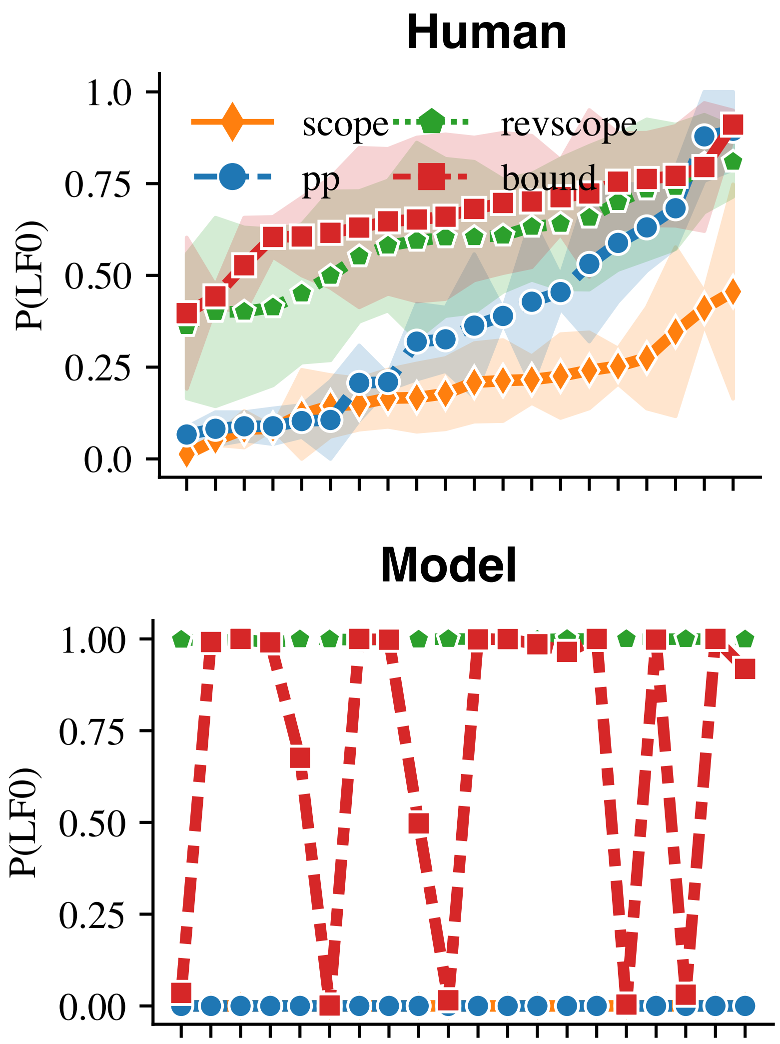
Semantic vs. form uncertainty

Models do not capture human uncertainty

Confidence as a proxy for accuracy

Safety-usability tradeoffs

Also: improving accuracy via confidence



Zero and few-shot semantic parsing with ambiguous inputs
Stengel-Eskin et al., 2023

Thank you!