Reflex or Reflect

When Do Language Tasks Need Slow Reasoning?

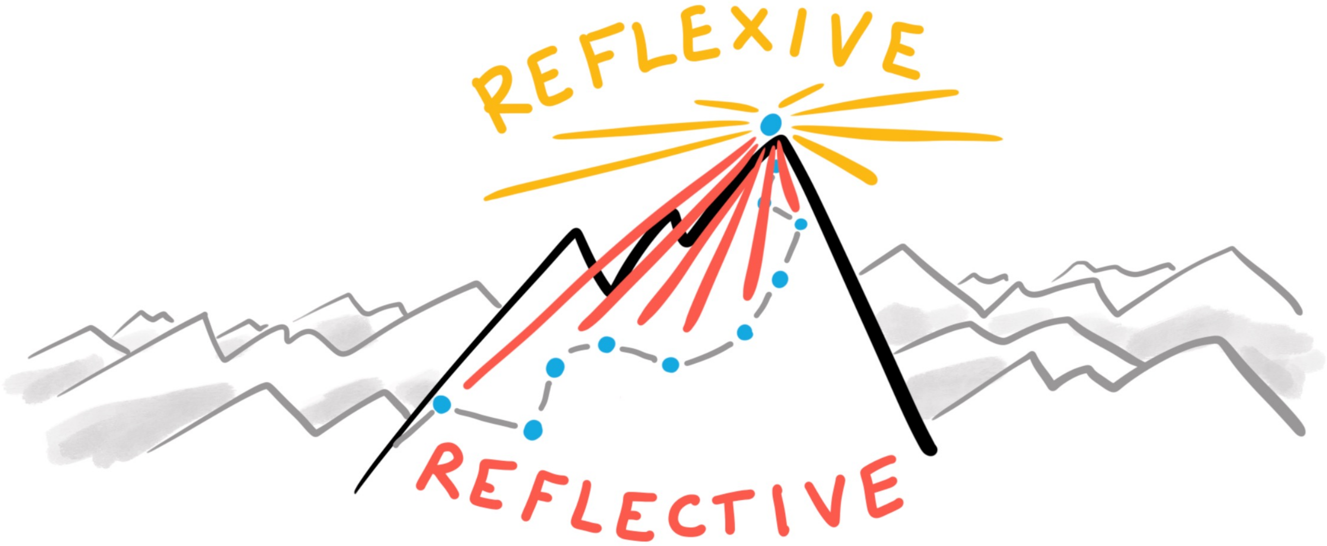
Xiang Ren

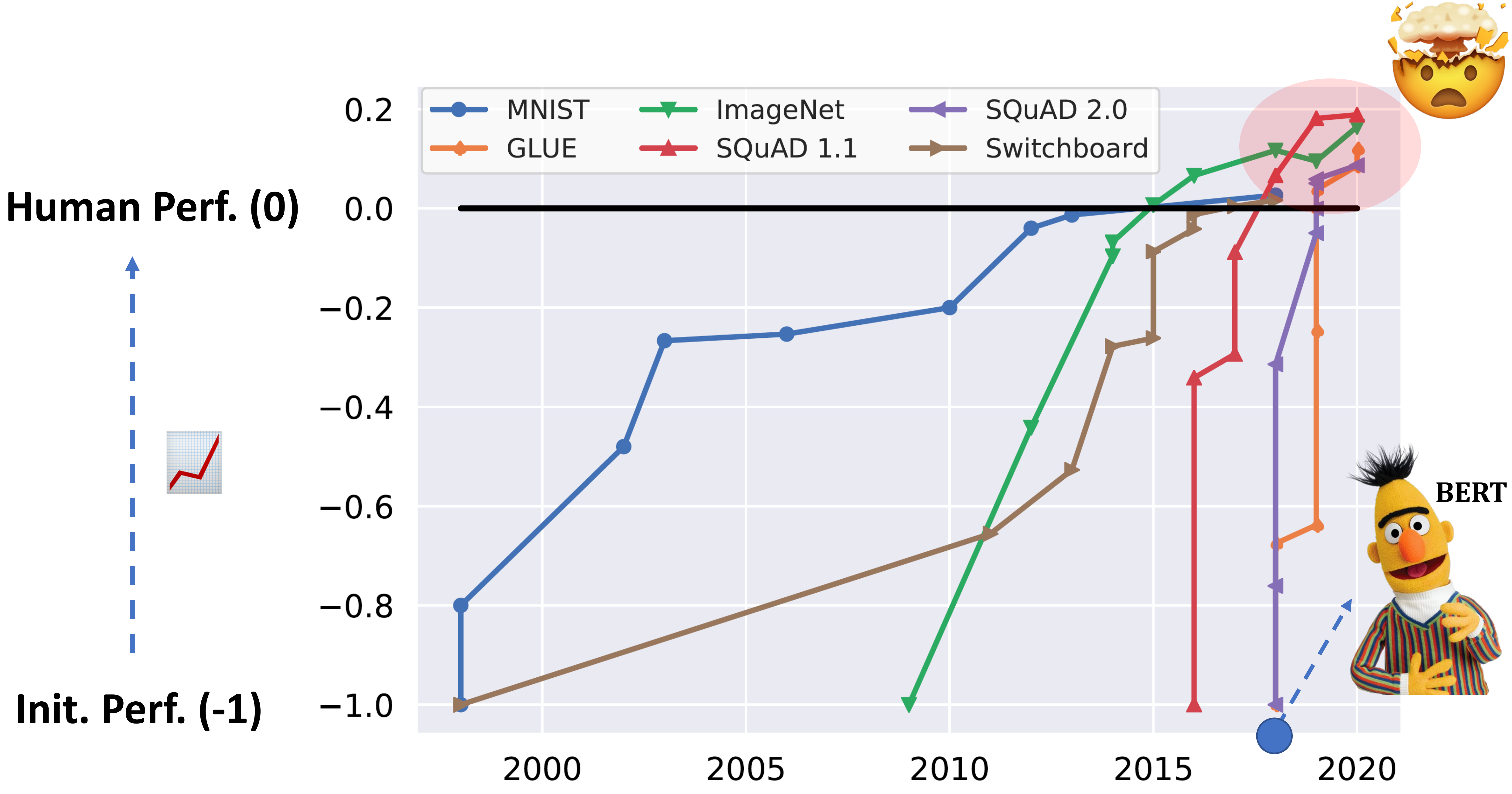
Associate Professor, CS & ISI

Viterbi Early Career Chair

University of Southern California

http://inklab.usc.edu





Initial baseline performance and human performance are normalized to **-1** and **0** respectively (Credit: Kiela et al., 2021).

Benchmark	GPT-4 Evaluated few-shot	GPT-3.5 Evaluated few-shot	LM SOTA Best external LM evaluated few-shot	SOTA Best external model (includes benchmark-specific training)
MMLU Multiple-choice questions in 57 subjects (professional & academic)	86.4% 5-shot	70.0% 5-shot	70.7% 5-shot U-PaLM	75.2% 5-shot Flan-PaLM
HellaSwag Commonsense reasoning around everyday events	95.3% 10-shot	85.5% 10-shot	84.2% LLAMA (validation set)	85.6% ALUM
AI2 Reasoning Challenge (ARC) Grade-school multiple choice science questions. Challenge-set.	96.3% 25-shot	85.2% 25-shot	84.2% 8-shot PaLM	85.6% ST-MOE
WinoGrande Commonsense reasoning around pronoun resolution	87.5% 5-shot	81.6% 5-shot	84.2% 5-shot PALM	85.6% 5-shot PALM
HumanEval Python coding tasks	67.0% 0-shot	48.1% 0-shot	26.2% 0-shot PaLM	65.8% CodeT + GPT-3.5
DROP (f1 score) Reading comprehension & arithmetic.	80.9 3-shot	64.1 3-shot	70.8 1-shot PaLM	88.4 QDGAT

https://openai.com/research/gpt-4


Your Magical AI-generated World

On My Wishlist: Reading the Air

Leslie and Jane are chatting at a coffee shop.

Leslie

Who was that man that I saw you with last night?

The latte is unbelievable here.

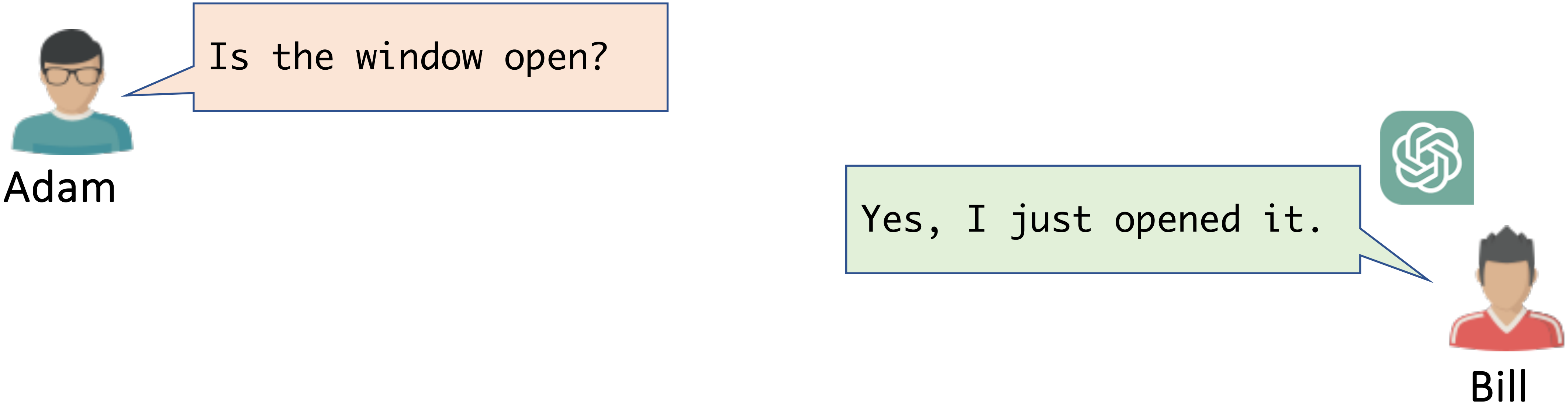
Jane

Leslie

I'm sorry, I didn't quite understand your response. You mentioned the latte, but didn't address my question about the man you were with last night.

On My Wishlist: Indirect Speech

Adam and Bill are working on a project in Bill's room. Bill opens the window to get some fresh air. A cold breeze blows in.



On My Wishlist: Indirect Speech

Adam and Bill are working on a project in Bill's room. Bill opens the window to get some fresh air. A cold breeze blows in.

Adam

Is the window open?

Bill

- Adam feels the breeze and would like to be warmer
- Adam probably wants to close the window
- (but Adam didn't want to seem rude)

On My Wishlist: Indirect Speech

Adam and Bill are working on a project in Bill's room. Bill opens the window to get some fresh air. A cold breeze blows in.

Is the window open?

Adam

Is it too cold? Do you want me to close it?

Bill

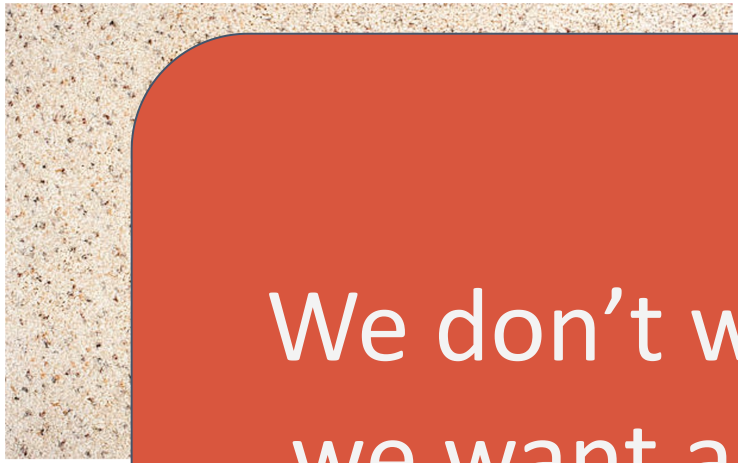
- Adam feels the breeze and would like to be warmer
- Adam probably wants to close the window
- (but Adam didn't want to seem rude)

Muscle-*Reflex* Style Language

Oh no, I spilled the food I prepared for dinner

Ah sorry to hear that!

"*Reflect*" Style Language

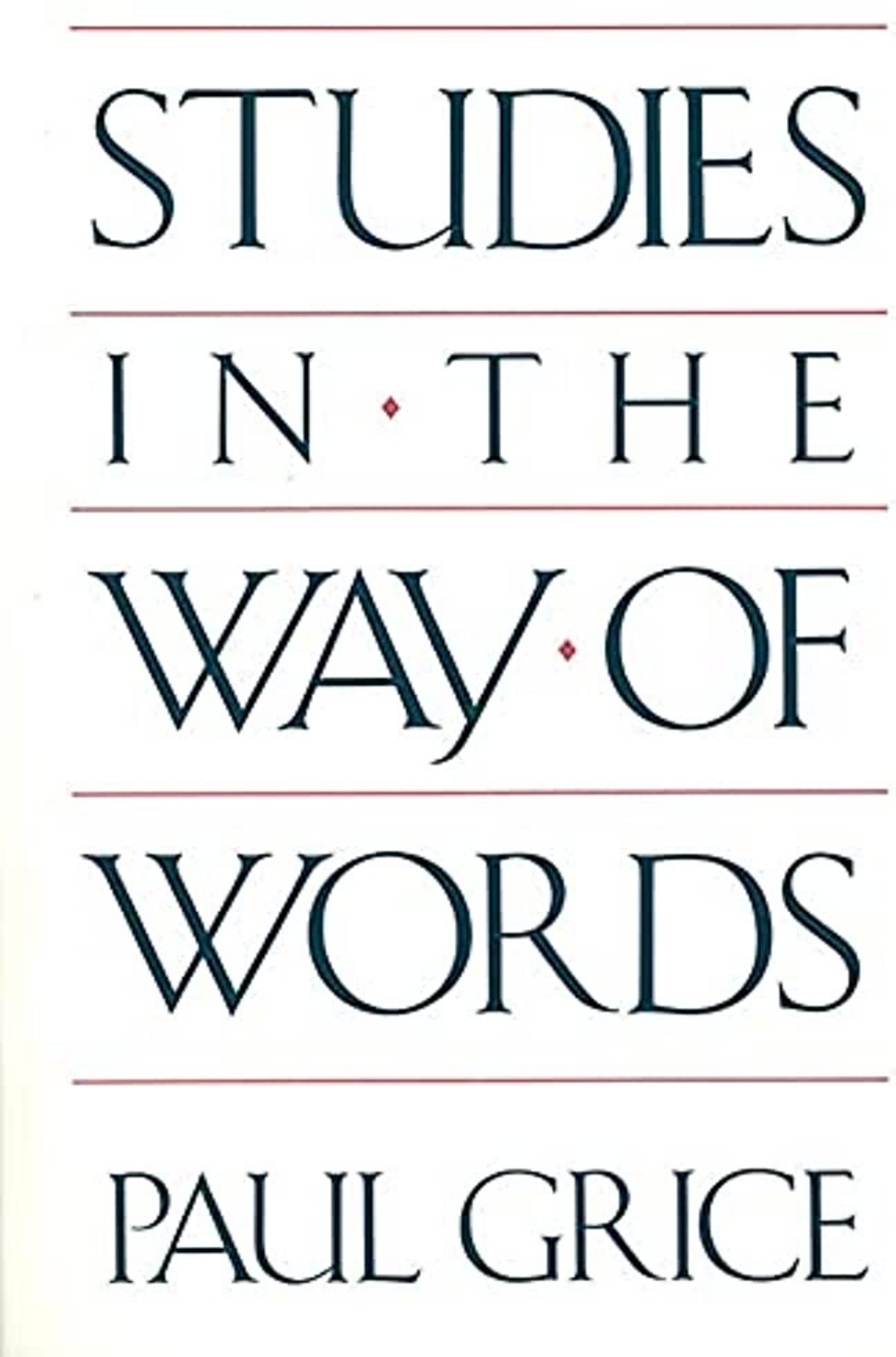


They might be

We don't want a ***knee-jerk machine***,
we want a ***thinking communicator!***

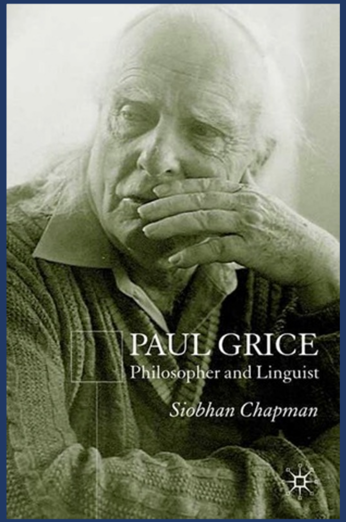
for dinner

from your favorite
pasta place?



Paul Grice's Maxims on *cooperative principles*

Communication is a **collaborative** effort with **intents** and people tend to "*minimize the total effort spent*". [Least collaborative effort]



Due to least collaborative effort, we need to **make inferences** to draw conclusions about the speaker's **intentions, emotion states, and experiences.**[Build Common Ground]

“*Reflect*” Style Language

Deep communication abilities

- ➢ Pragmatics
- ➢ Understanding Intent
- ➢ Commonsense Inferences
- ➢ Theory-of-Mind

be
and
up

Oh no
spill
food
for d

orry! How
et’s clean
nd order
ur favorite
lace?

"*Reflect*" Style Language

Why Challenging?

- Often implicit in training corpora → more prone to generate *shallow* replies

- Appropriate answers require ***slow reasoning*** about others' true intents and common sense

pasta place?

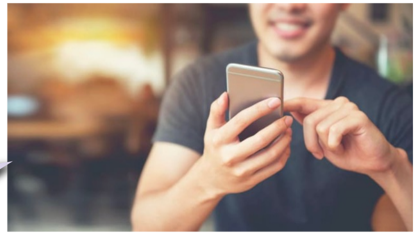
How do we reply in conversations?

I'm going to sing in front of hundreds tomorrow...

How do we reply in conversations?

I'm going to perform in a piano recital tomorrow...

Performing in front of audience can cause ***anxiety***



Deep breaths, you'll do great!

Recalling & Combining common sense with information expressed in NL to ***make inferences***

Producing ***consistent*** inferences amidst ***logically-equivalent yet linguistically-varied*** paraphrases

Clark, H. H., & Brennan, S. E. (1991). *Grounding in communication.*

RICA: Robust Inference on Commonsense Axioms

- Test model's robustness against linguistic variations
- Focus on implicit commonsense inferences
- Scalable probe set construction process

in Proc. of EMNLP 2021

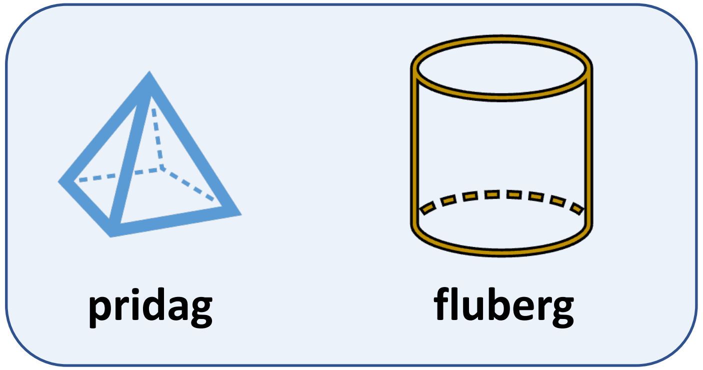
RICA: Evaluating Robust Inference Capabilities Based on Commonsense Axioms

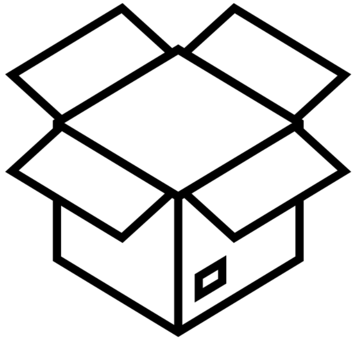
Pei Zhou Rahul Khanna Seyeon Lee Bill Yuchen Lin Daniel Ho
Jay Pujara Xiang Ren
Department of Computer Science and Information Sciences Institute
University of Southern California
{peiz,rahulkha,seyeonle,yuchen.lin,hsiaotuh,jpujara,xiangren}@usc.edu

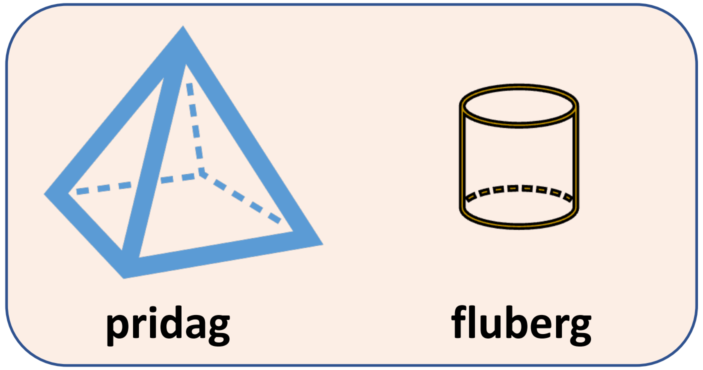
RICA
EMNLP’21

Commonsense Logic to Probe:

A.Size < B.Size → P(A in Container) > P(B in Container)







<u>A pridag is smaller than a fluberg,</u> so it is [MASK] to put a pridag into a box than a fluberg.

easier (86.6%)
harder (1.1%)

<u>A fluberg is smaller than a pridag,</u> so it is [MASK] to put a pridag into a box than a fluberg.

easier (87.2%)
harder (1.3%)

RICA: Robust Inference on Commonsense Axioms

- Examples:
  - **Original**: “A is heavier than B, so A is <better> at sinking than B.”
  - **Negation**: “A is heavier than B, so A is **not** <worse> at sinking than B.”
  - **Entity Swap**: “**B** is heavier than **A**, so A is <worse> at sinking than B.”
  - **Antonym**: “A is heavier than B, so A is <worse> at floating than B.”
  - …

(Zhou et al., EMNLP’21)

RICA: Robust Inference on Commonsense Axioms

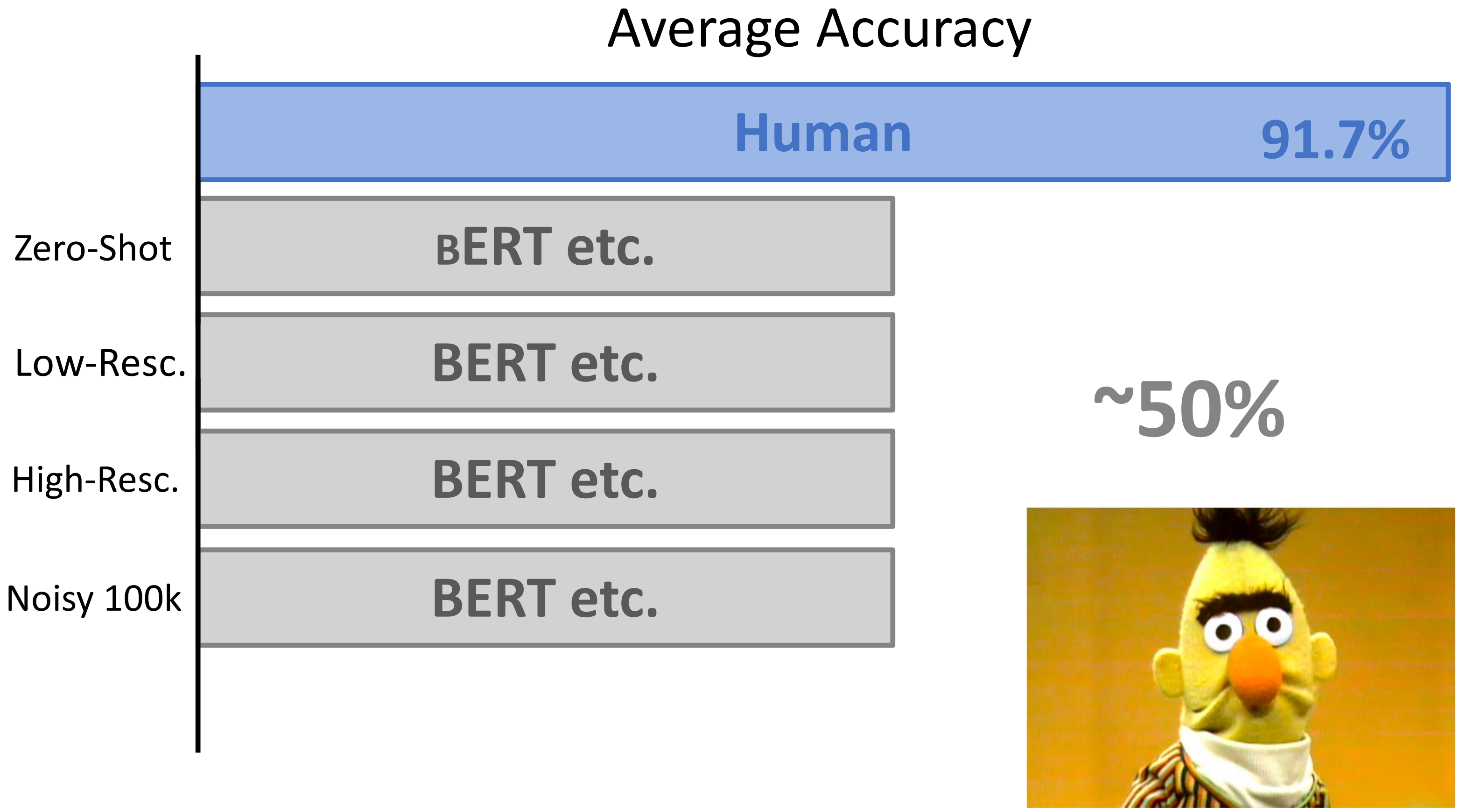
- Masked word prediction task: **Choose <better> or <worse>**:
  - **Original**: "A is heavier than B, so A is <MASK> at sinking than B."
  - **Perturb1**: "A is heavier than B, so A is **not** <MASK> at sinking than B."
  - **Perturb2**: "**B** is heavier than **A**, so A is <MASK> at sinking than B."
  - **Perturb3**: "A is heavier than B, so A is <MASK> at **floating** than B."
  - …

(Zhou et al., EMNLP'21)

Results: Human-Curated Set

- Random-guessing like performance on *all settings* for all models.

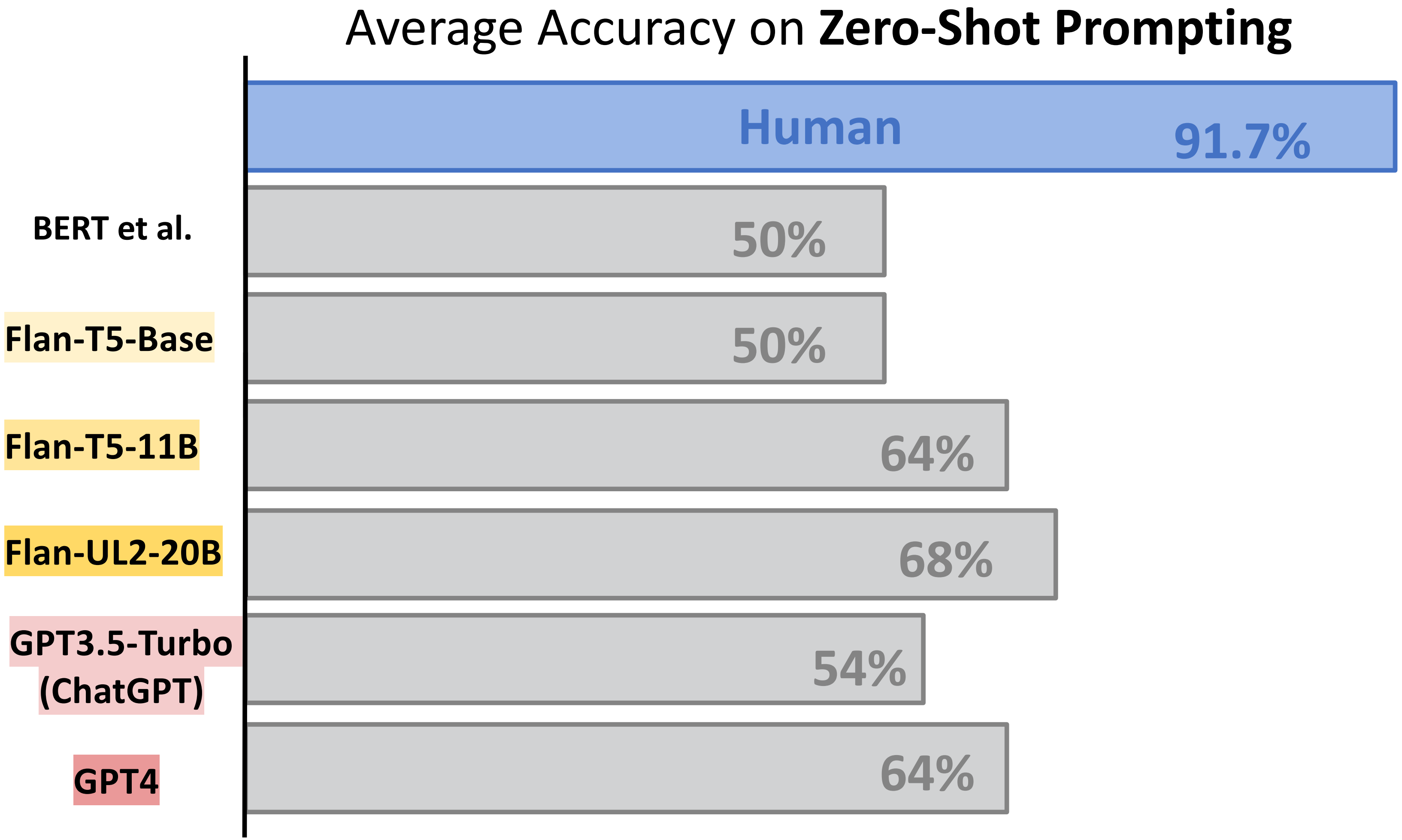
- Training on similar data does not help achieve real robustness



Results: How About Fancy New LLMs?

RICA still remains challenging to LLMs

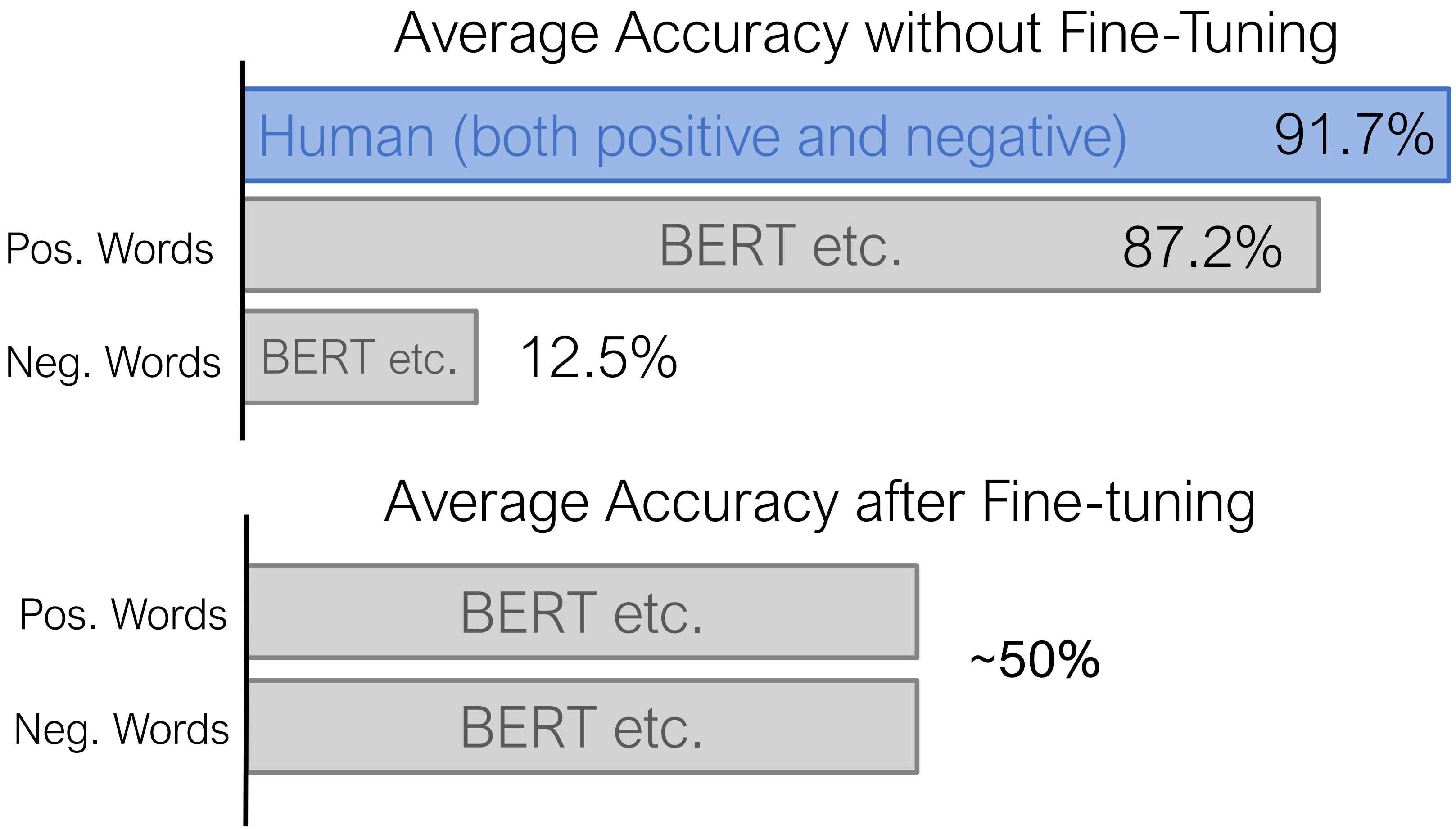
- Larger models tend to perform better for T5-family models
- GPT-family models seem less magical
  - Bidirectional attention better captures logic with perturbations?



Analysis: Positivity Bias

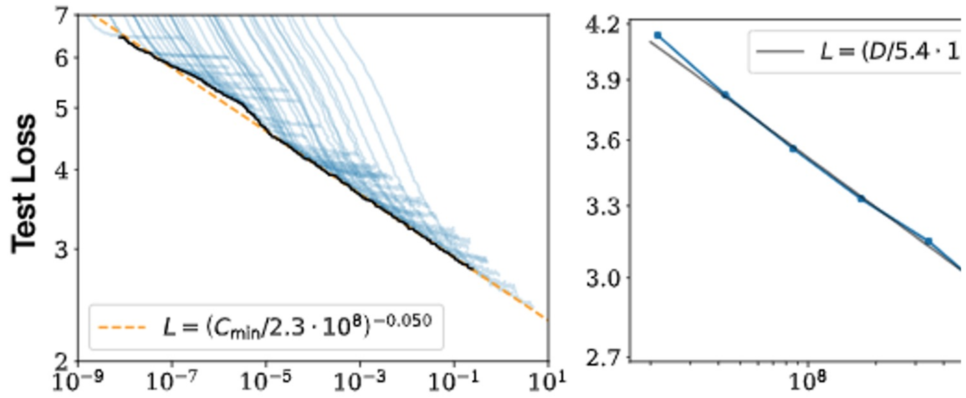
- Heavy bias towards positive-valence words such as "*more*", "*better*".

- Fine-tuning on RICA mitigates the imbalance issue (but still fails)

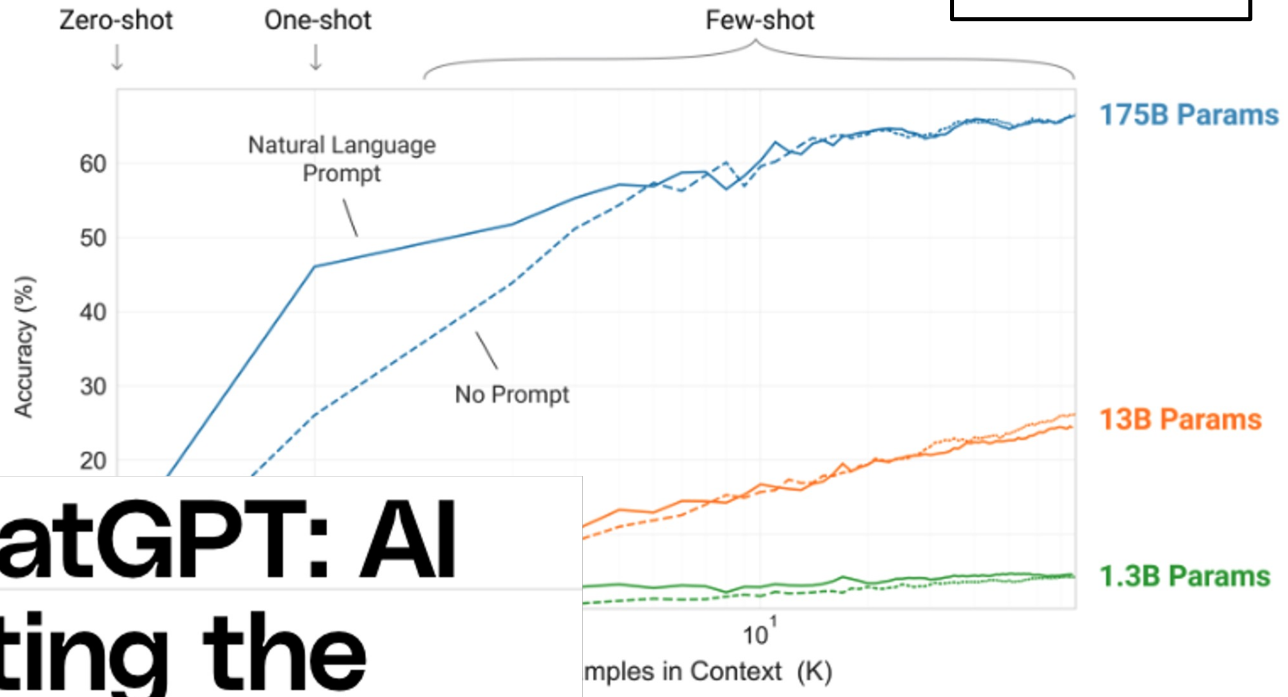


Scaling is the Way Going Forward!

Scaling Laws for Neural Language Models



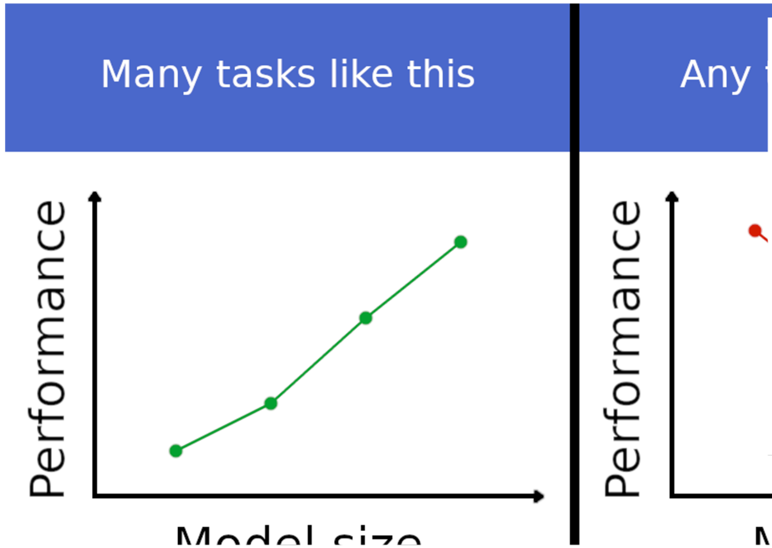
GPT3



Bing, Bard, and ChatGPT: AI chatbots are rewriting the internet

How we use the internet is changing fast, thanks to the advancement of AI-powered chatbots that can find information and redeliver it as a simple conversation.

Does Scaling Always Work?



Zhengping Zhou and Yuhui Zhang, for *NeQA: Can Large Language Models Understand Negation in Multi-choice Questions?*

- This task takes an existing multiple-choice dataset and negates a part of each question to see if language models are sensitive to negation. The authors find that smaller language models display approximately random performance whereas the performance of larger models become significantly worse than random.

Modus Tollens, by Sicong Huang and Daniel Wurgaft (Third Prize)

TL;DR This task shows strong inverse scaling on almost all models and represents a simple logical reasoning task (*modus tollens*) that might be expected to show regular scaling. Inverse scaling trends hold across both pretrained LMs and LMs finetuned with human feedback via RL from Human Feedback (RLHF) and Feedback Made Easy (FeedME).

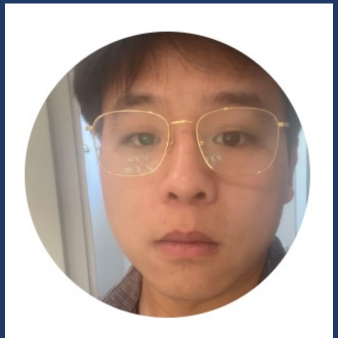
Robustness on logical reasoning?

https://github.com/inverse-scaling/prize

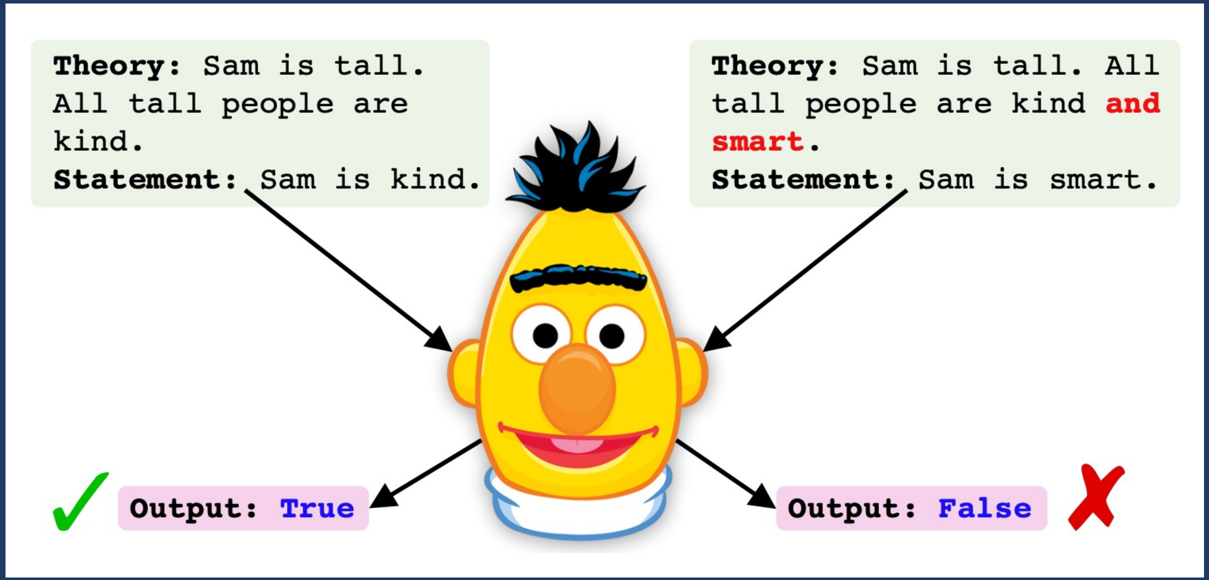
RobustLR: A Diagnostic Benchmark for Evaluating Logical Robustness of Deductive Reasoners

Soumya Sanyal

Zeyi Liao



Xiang Ren



(EMNLP'22)

Language-based Deductive Reasoning

```
fact1: Charlie is blue.
fact2: Charlie is round.
fact3: Erin is kind.
fact4: Dave is round.
rule1: If someone is blue then they are kind.
rule2: Round, kind people are white.
statement: Charlie is white.
```

Theory

Input: Facts + Rules (theory), Statement

Output: Entailment label

- **True**: Theory → statement is True
- **False**: Theory → negation of the statement is True
- **Unknown**: No conclusion

Can ChatGPT do Deductive Reasoning?

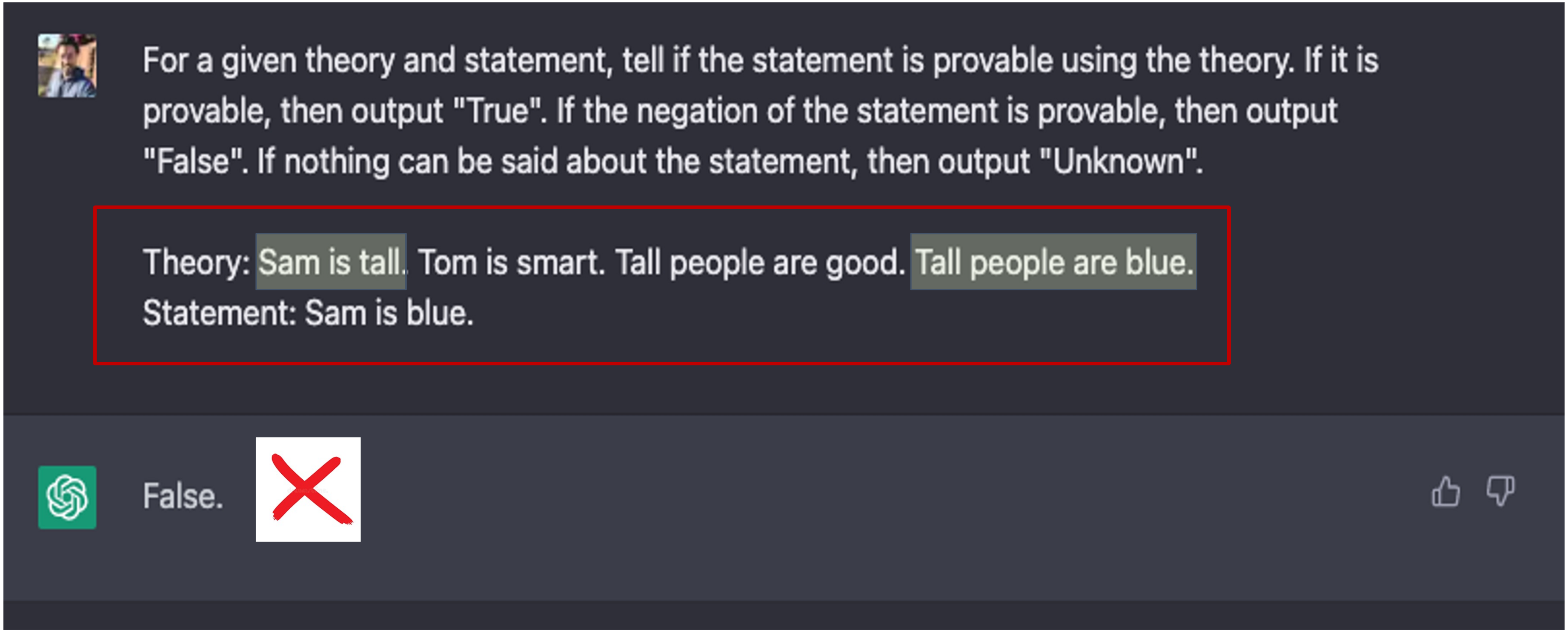
For a given theory and statement, tell if the statement is provable using the theory. If it is provable, then output "True". If the negation of the statement is provable, then output "False". If nothing can be said about the statement, then output "Unknown".

Theory: Sam is tall. Tom is smart. Tall people are good. Tall people are not blue.
Statement: Sam is blue.

Output: False.

Sure, it can get it right sometimes, but …

Can ChatGPT do Deductive Reasoning?



... *not* robust to negation within the theory..

Can ChatGPT do Deductive Reasoning?

For a given theory and statement, tell if the statement is provable using the theory. If it is provable, then output "True". If the negation of the statement is provable, then output "False". If nothing can be said about the statement, then output "Unknown".

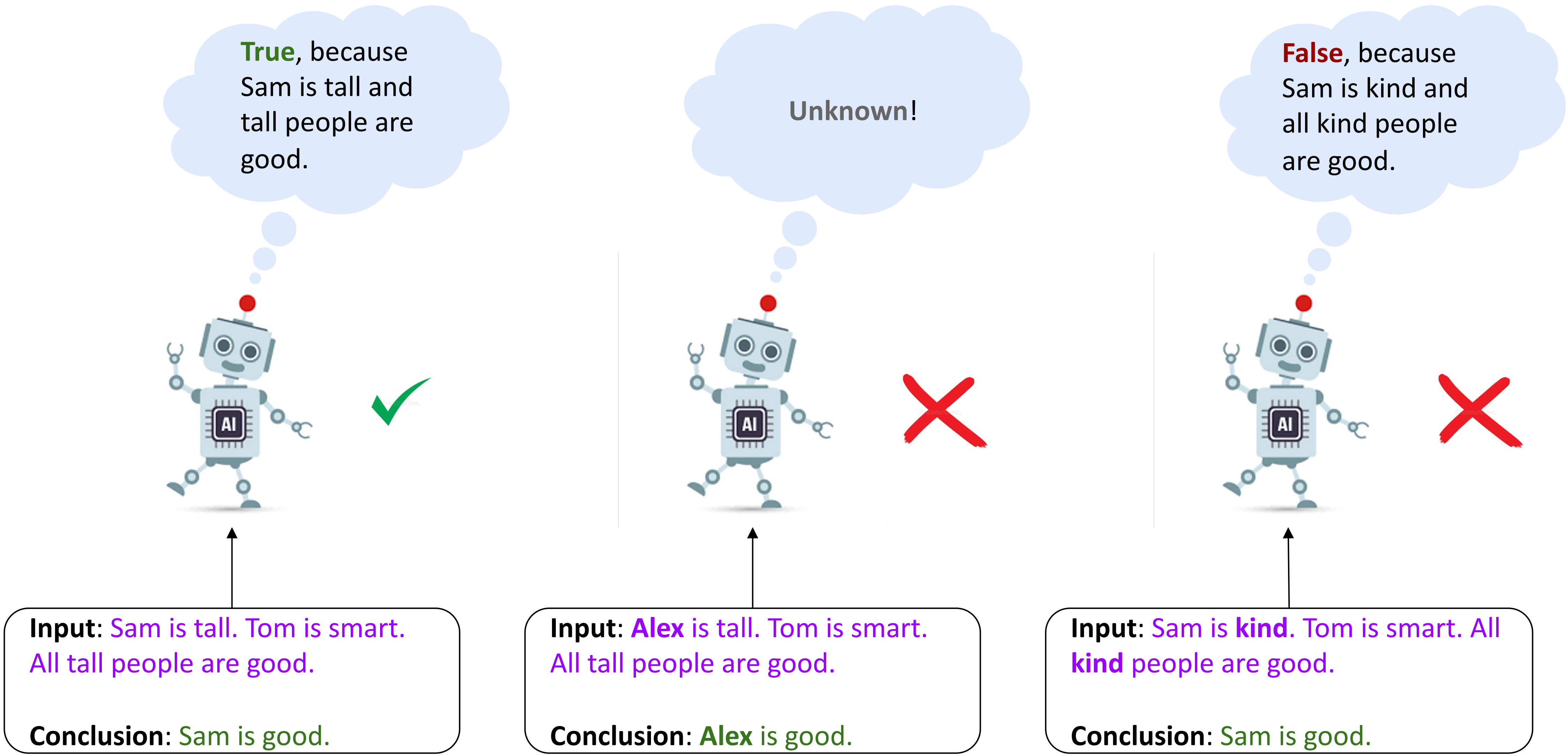
Theory: Sam is tall. Tom is smart. Tall people are good. Tall people are blue and not kind.
Statement: Sam is kind.

Unknown.

The theory provides information about Sam being tall, but it doesn't provide any direct information about whether Sam is kind or not. Therefore, we cannot determine the truth value of the statement "Sam is kind" based on the given theory alone.

Erroneous reasoning given the theory…

Robust Reasoning: Lexical Perturbation



Robust Reasoning: Lexical Perturbation

True, because Sam is tall and tall people are good.

Unknown!

False, because Sam is kind and all kind people are good.

FaiRR: Faithful and Robust Deductive Reasoning over Natural Language, ACL 2022

Input: Sam is tall. Tom is smart. All tall people are good.

Conclusion: Sam is good.

Input: **Alex** is tall. Tom is smart. All tall people are good.

Conclusion: **Alex** is good.

Input: Sam is **kind**. Tom is smart. All **kind** people are good.

Conclusion: Sam is good.

RobustLR: Logical Perturbation

Sam is tall. Tom is smart. Tall people are good. Tall people are blue.

Sam is good. **True**

RobustLR: Logical Perturbation

Sam is tall. Tom is smart. Tall people are good. Tall people are blue.

Sam is good. **True**

- Logical Equivalence **Contraposition** $(A \rightarrow B \equiv \sim B \rightarrow \sim A)$

Sam is tall. Tom is smart. A person who's not good is also not tall. Tall people are blue.

Sam is good. **True**

RobustLR: Logical Perturbation

Sam is tall. Tom is smart. Tall people are good. Tall people are blue.

Sam is good. **True**

- Logical Equivalence **Contraposition**
  $(A \rightarrow B \equiv {\sim}B \rightarrow {\sim}A)$

- Logical Equivalence **Distributive**
  $(A \rightarrow B; A \rightarrow C \equiv A \rightarrow B \text{ AND } C)$

Sam is tall. Tom is smart. Tall people are good and blue.

Sam is good. **True**

RobustLR: Logical Perturbation

Sam is tall. Tom is smart. Tall people are good. Tall people are blue.

Sam is good. **True**

- Logical Equivalence **Contraposition**
  (A → B ≣ ~B → ~A)

- Logical Equivalence **Distributive**
  (A → B; A → C ≣ A → B AND C)

- Logical **Contrast**
  (A → B vs A → B & C, etc.)

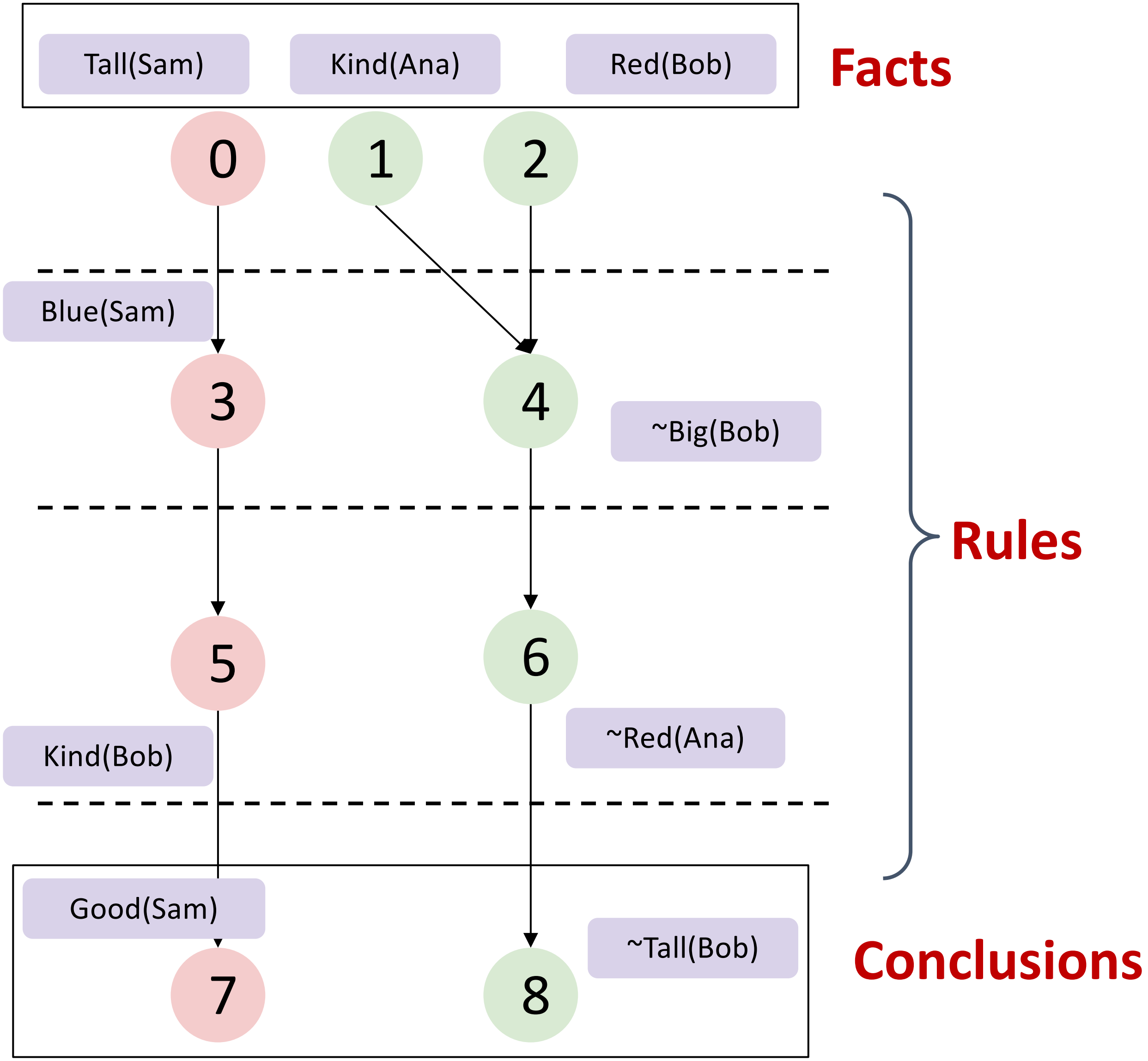
Sam is tall. Tom is smart. Tall people are good. Tall people are blue.

Sam is good. **True**
Sam is kind. **Unknown**

Sam is tall. Tom is smart. Tall people are good and not kind. Tall people are blue.

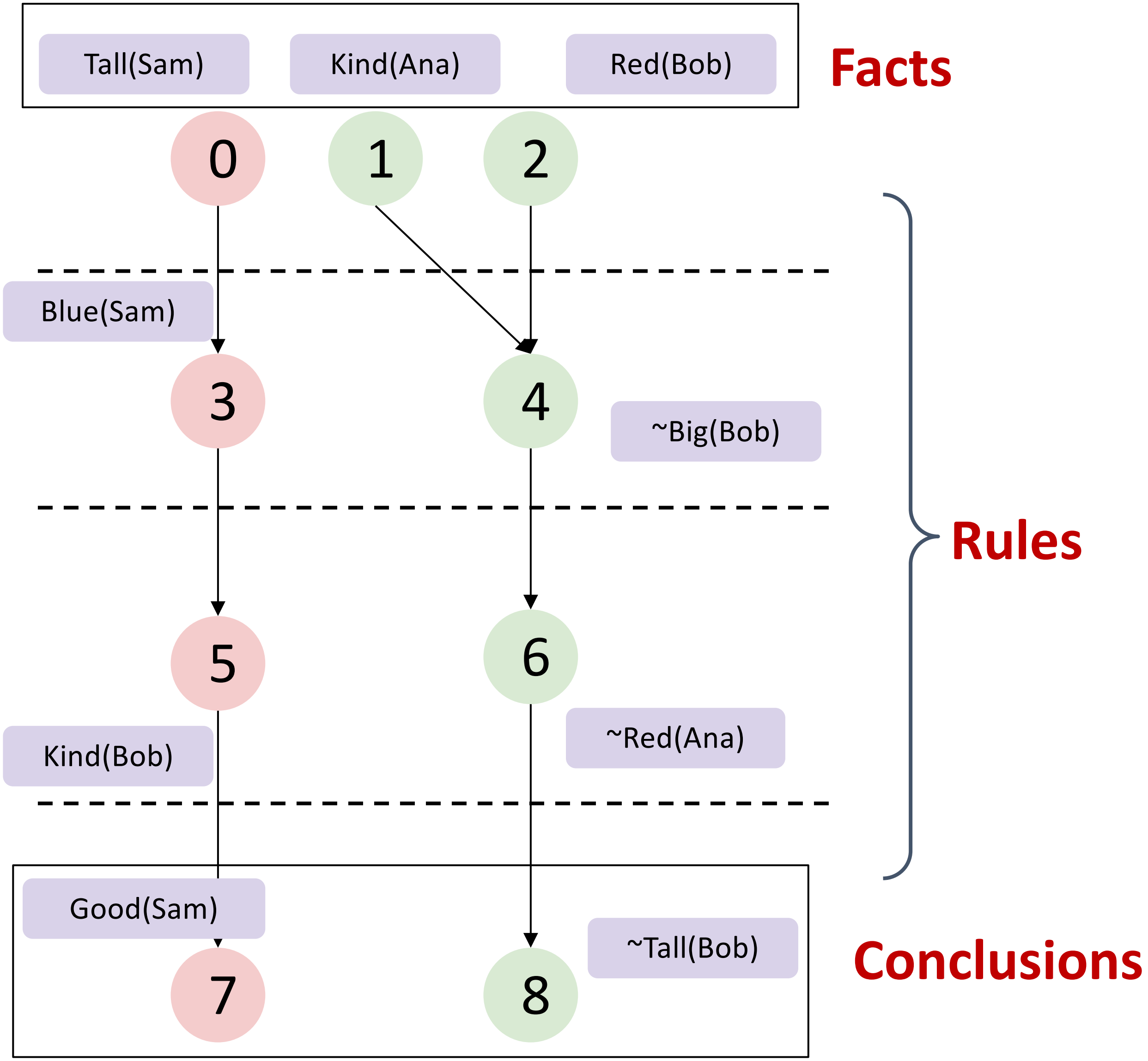
Sam is good. **True**
Sam is kind. **False**

RobustLR: Dataset generation process



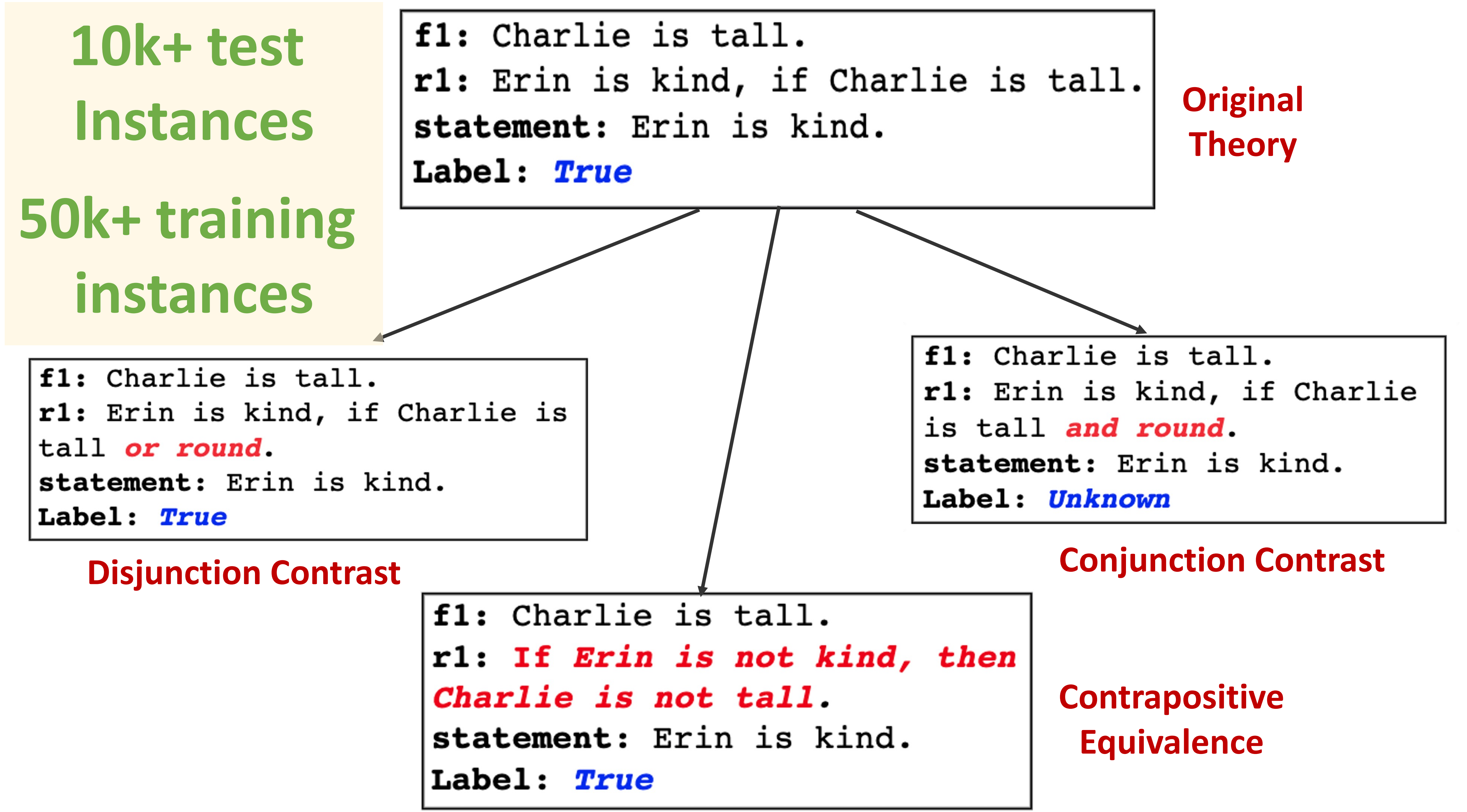
1. Sample some predicates
2. Label the predicates as **valid** and **invalid**
3. Break down into multiple levels
4. Starting from level 1, select predicates from lower level, such that a valid rule is formed

RobustLR: Dataset generation process

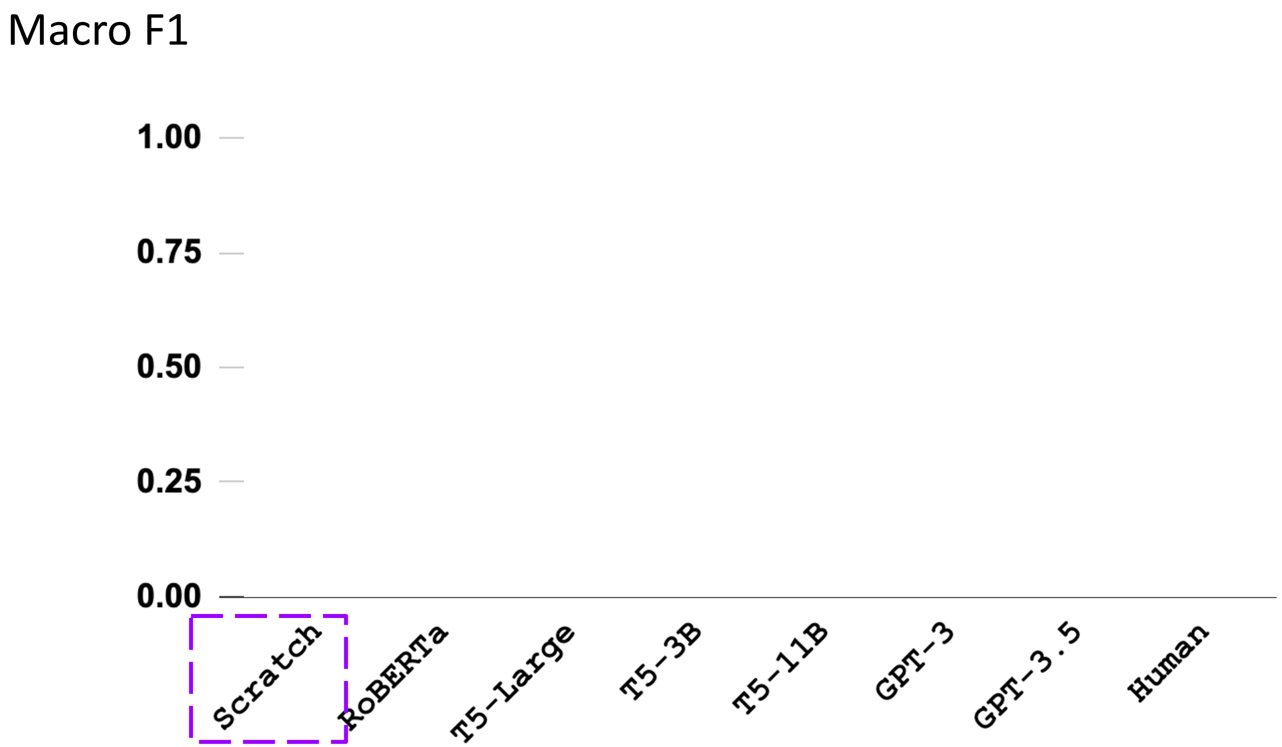


1. Sample some predicates
2. Label the predicates as **valid** and **invalid**
3. Break down into multiple levels
4. Starting from level 1, select predicates from lower level, such that a valid rule is formed

Can control the degree of the rule, #negations, multiple proof graphs, etc., in a flexible manner

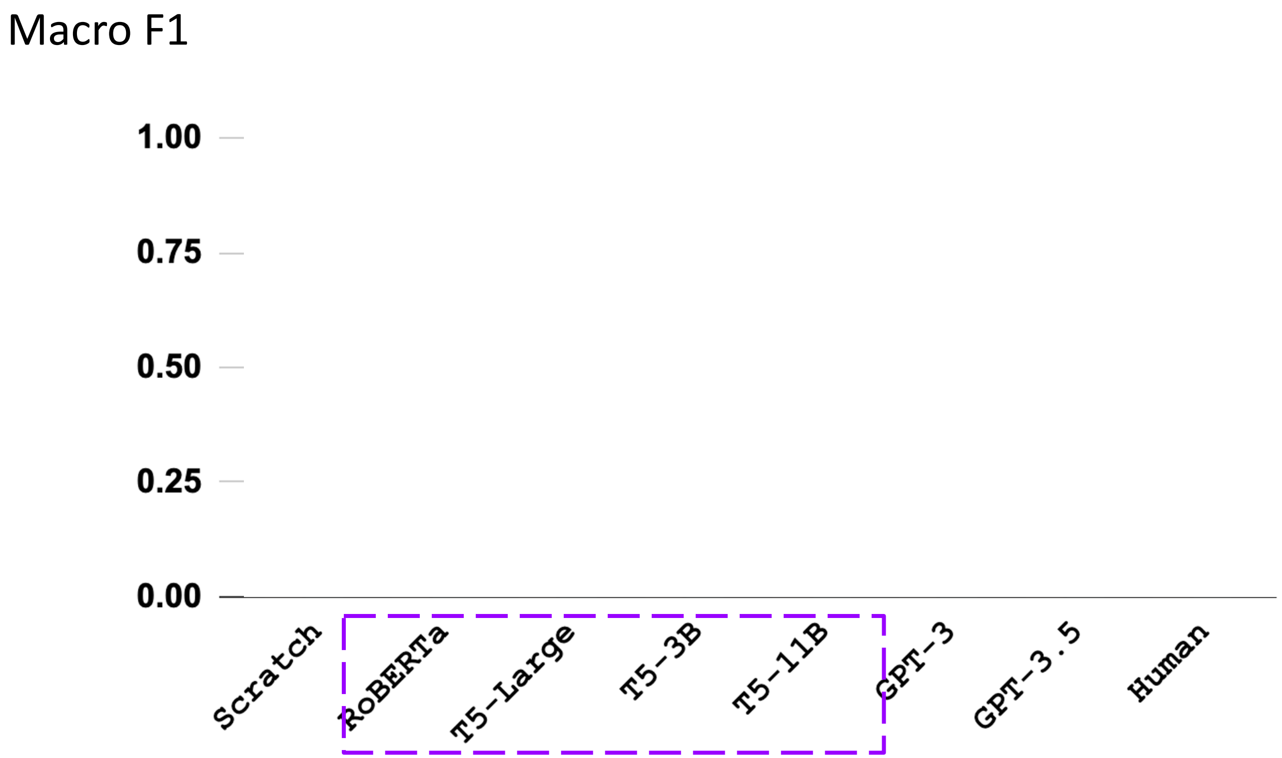

10k+ test Instances
50k+ training instances
f1: Charlie is tall.
r1: Erin is kind, if Charlie is tall.
statement: Erin is kind.
Label: True
Original Theory
f1: Charlie is tall.
r1: Erin is kind, if Charlie is tall or round.
statement: Erin is kind.
Label: True
Disjunction Contrast
f1: Charlie is tall.
r1: Erin is kind, if Charlie is tall and round.
statement: Erin is kind.
Label: Unknown
Conjunction Contrast
f1: Charlie is tall.
r1: If Erin is not kind, then Charlie is not tall.
statement: Erin is kind.
Label: True
Contrapositive Equivalence

Results - Machine vs Human



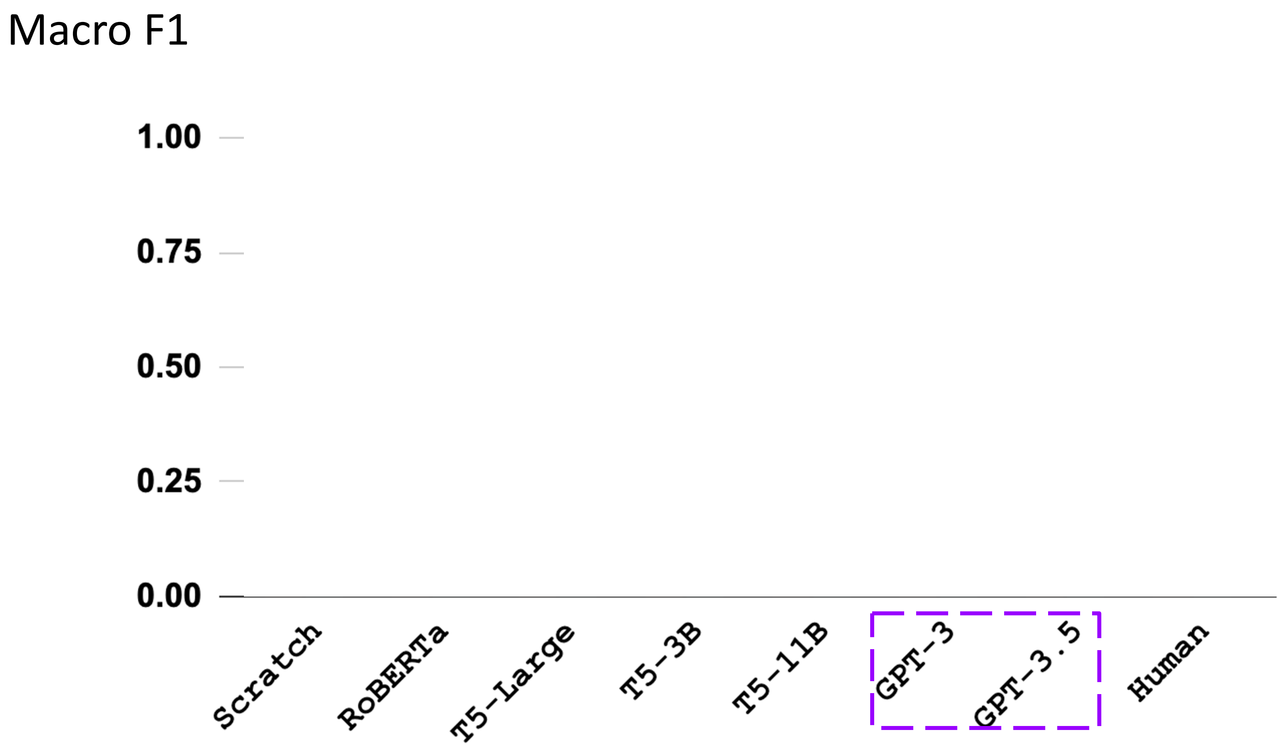
*Training a RoBERTa architecture from scratch

Results - Machine vs Human



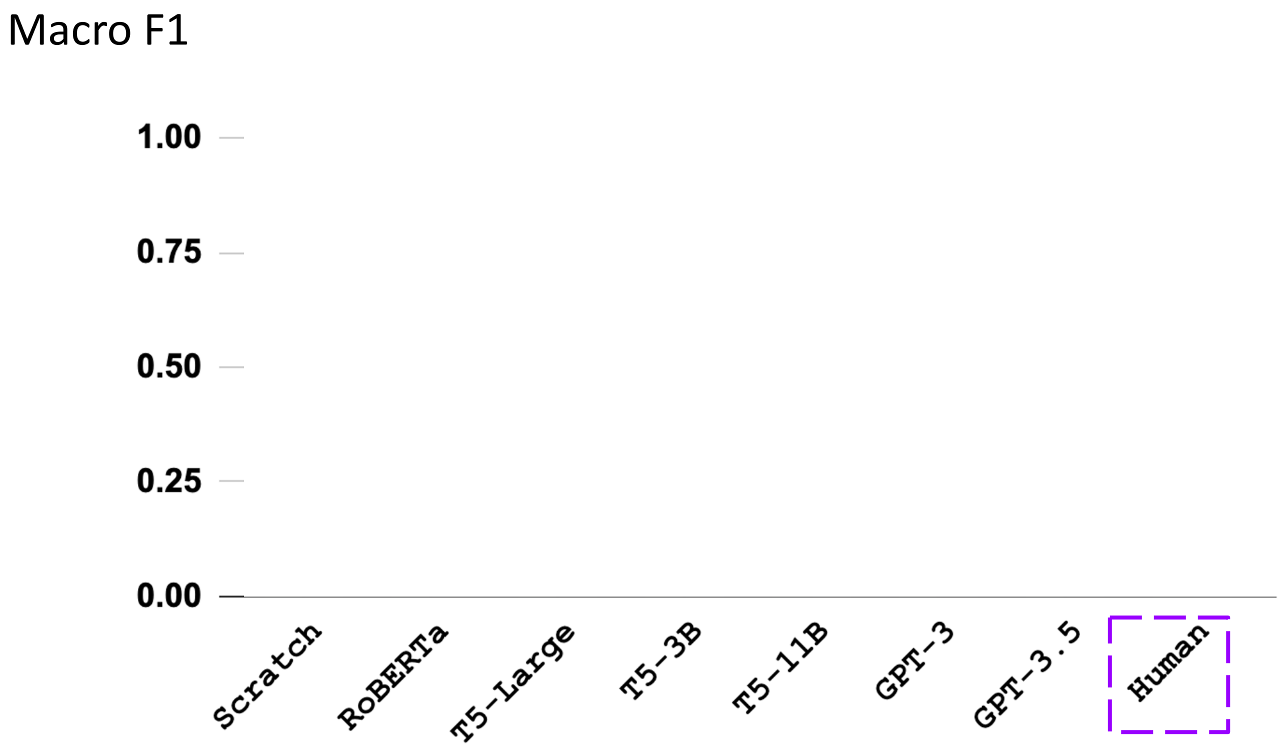
*Finetune a pretrained checkpoint

Results - Machine vs Human



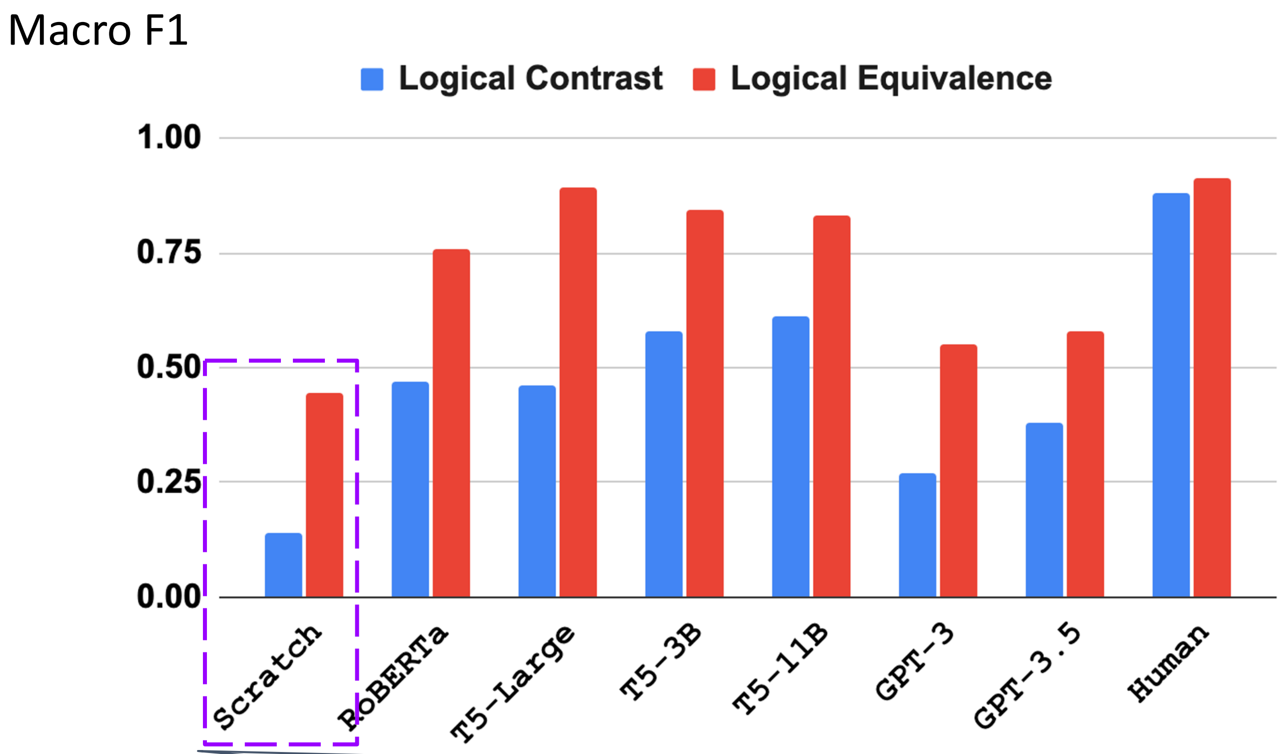
*6-shot in-context learning

Results - Machine vs Human



*7 CS graduates annotating a subset of the data

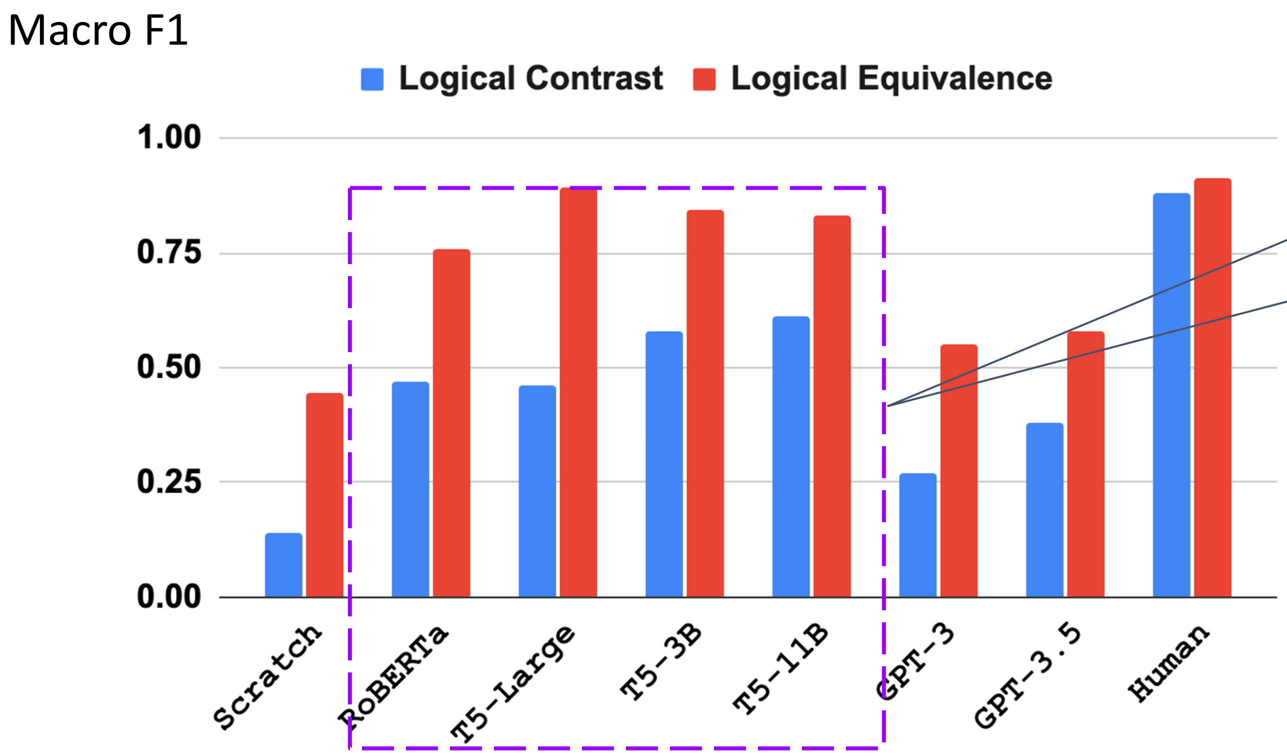
Results - Machine vs Human



Training from scratch fails!

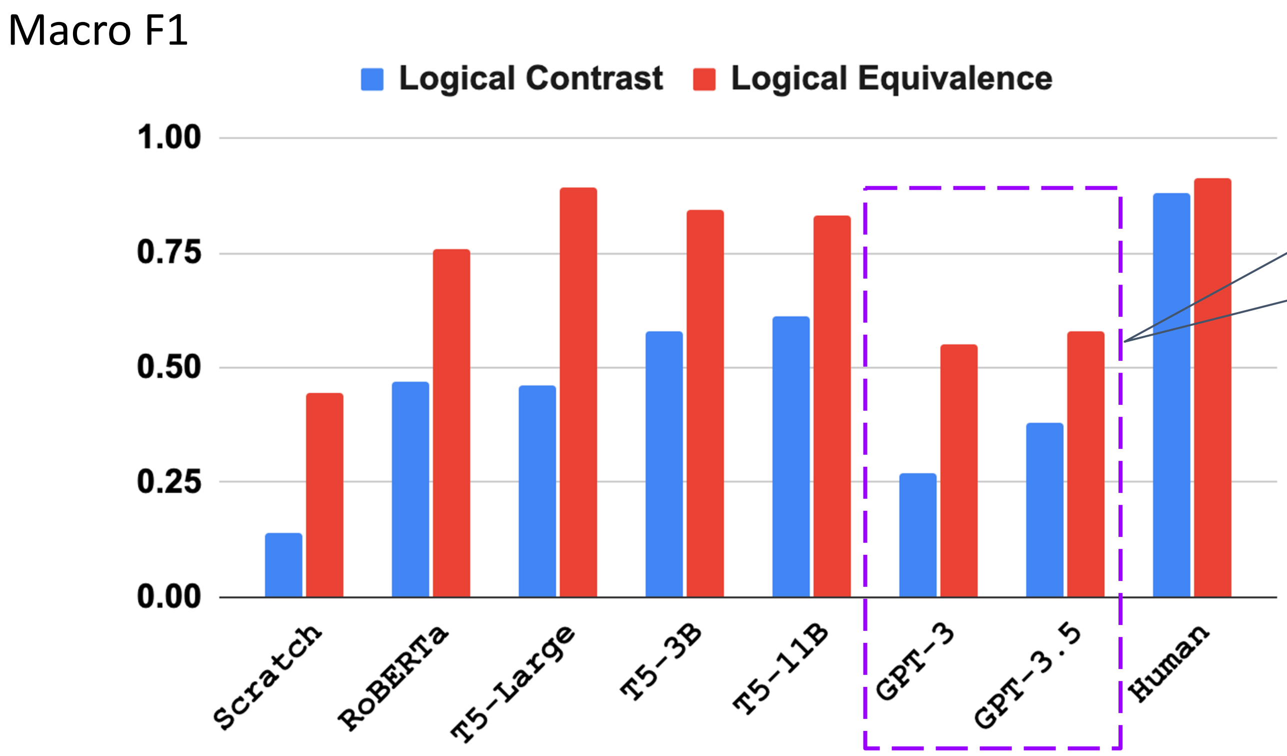
Pretrained knowledge is crucial

Results - Machine vs Human



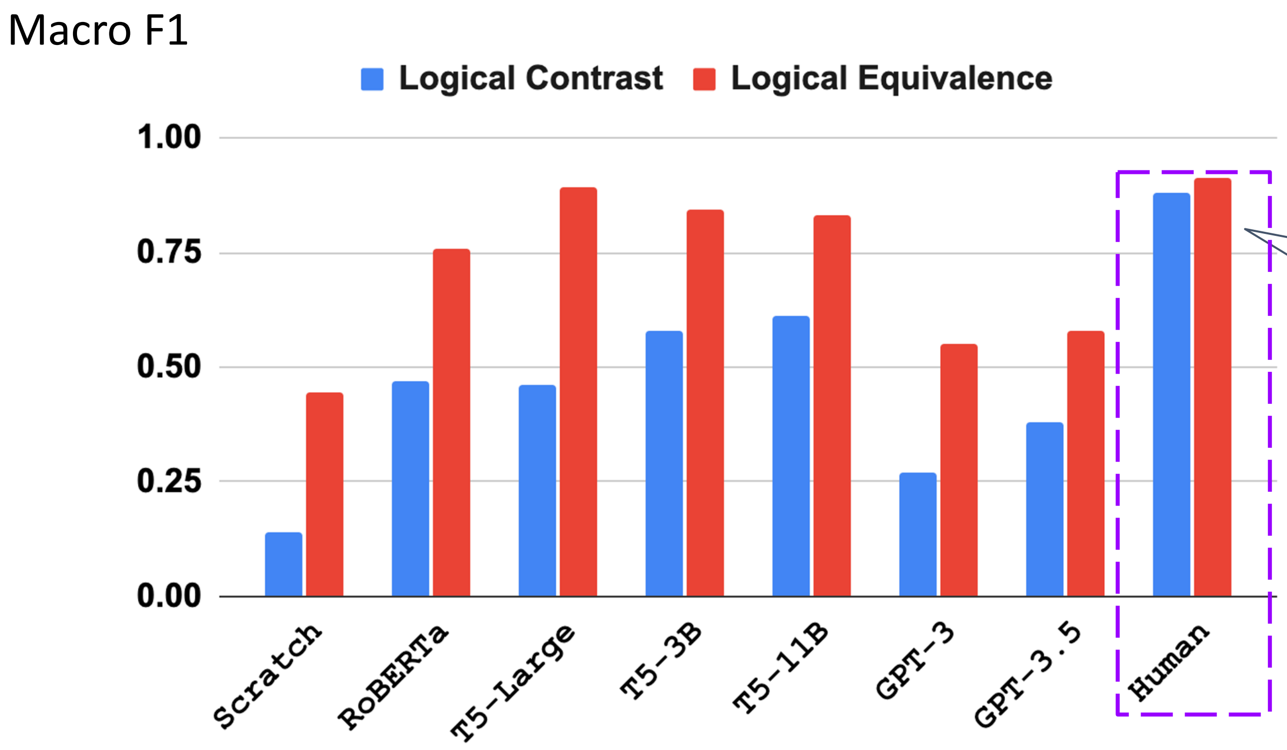
Model size is not a very significant factor,
but T5 > RoBERTa!

Results - Machine vs Human



GPT3/3.5 performance is worse than finetuned models!

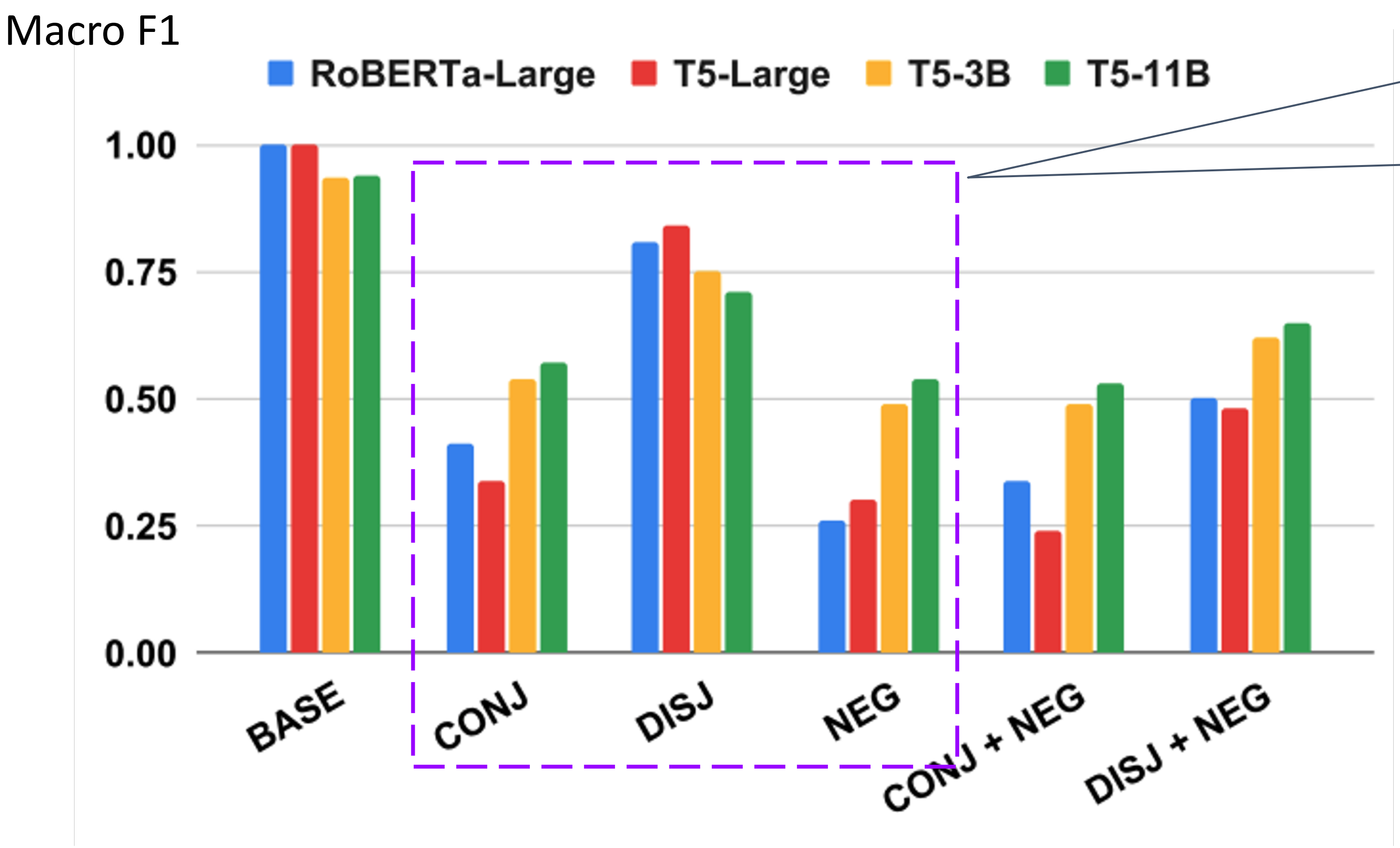
Results - Machine vs Human



The performance gap is **low** for humans

→ more robust reasoning!

Results - Variation with Logical Operators



Difficulty level

Negation > Conjunction > Disjunction

Related Works

P1: David, Jack and Mark are colleagues in a company. David supervises Jack, and Jack supervises Mark. David gets more salary than Jack.

Q: *What can be inferred from the above statements?*

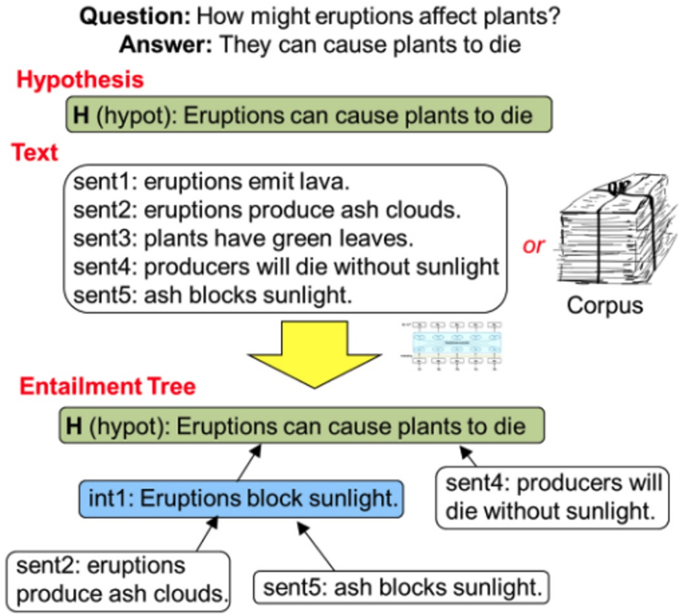
A. Jack gets more salary than Mark.
B. David gets the same salary as Mark.
C. One employee supervises another who gets more salary than himself.
✓ **D. One employee supervises another who gets less salary than himself.**

P2: Our factory has multiple dormitory areas and workshops. None of the employees who live in dormitory area A are textile workers. We conclude that some employees working in workshop B do not live in dormitory area A.

Q: *What may be the missing premise of the above argument?*

A. Some textile workers do not work in workshop B.
B. Some employees working in workshop B are not textile workers.
✓ **C. Some textile workers work in workshop B.**
D. Some employees living in dormitory area A work in the workshop B.

LogiQA



Entailment Bank

(Input Facts:) Alan is blue. Alan is rough. Alan is young.
Bob is big. Bob is round.
Charlie is big. Charlie is blue. Charlie is green.
Dave is green. Dave is rough.

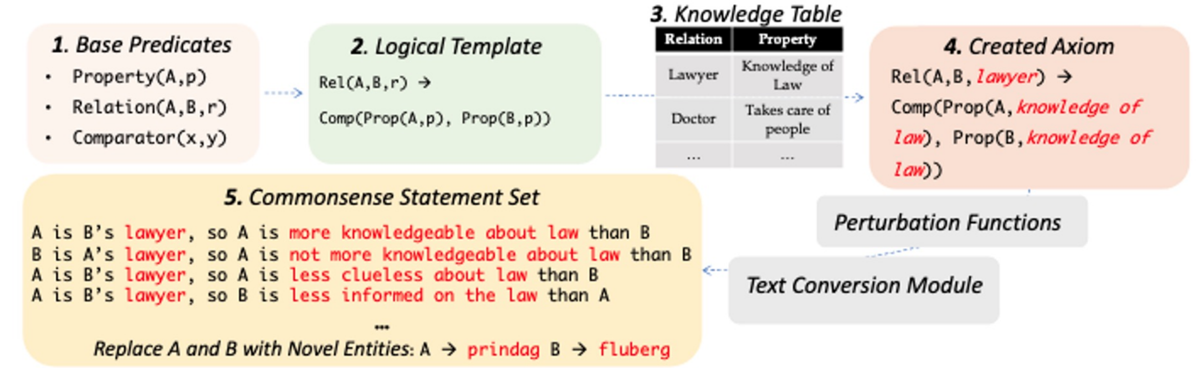
(Input Rules:) Big people are rough.
If someone is young and round then they are kind.
If someone is round and big then they are blue.
All rough people are green.

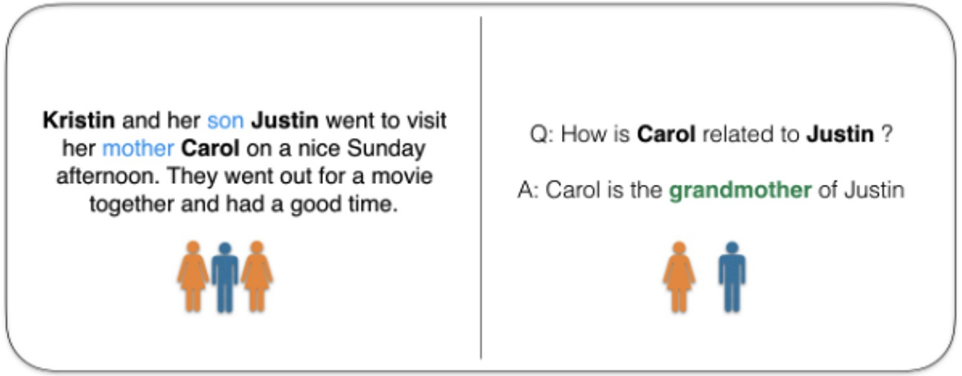
Q1: Bob is green. True/false? **[Answer: T]**
Q2: Bob is kind. True/false? **[F]**
Q3: Dave is blue. True/false? **[F]**

RuleTaker

RICA

CLUTRR





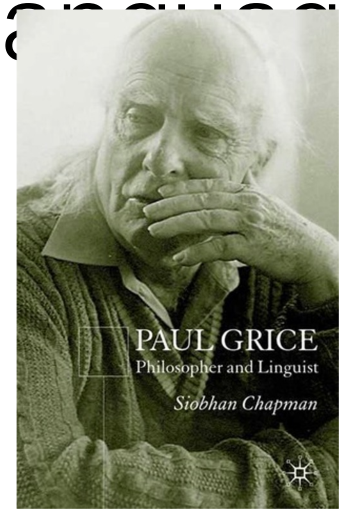
"*Reflect*" Style Language Reasoning

They might be feeling bad and need help cleaning it up

Oh no, I spilled the food I prepared for dinner

Don't worry! How about let's clean it up and order from your favorite pasta place?

We Need Slower and Deeper Language Reasoning



- Paul Grice's Maxims on *cooperative principles*

- Herbert H Clark: *Common ground*

- Jens Allwood: *Linguistic Communication as Action and Cooperation*

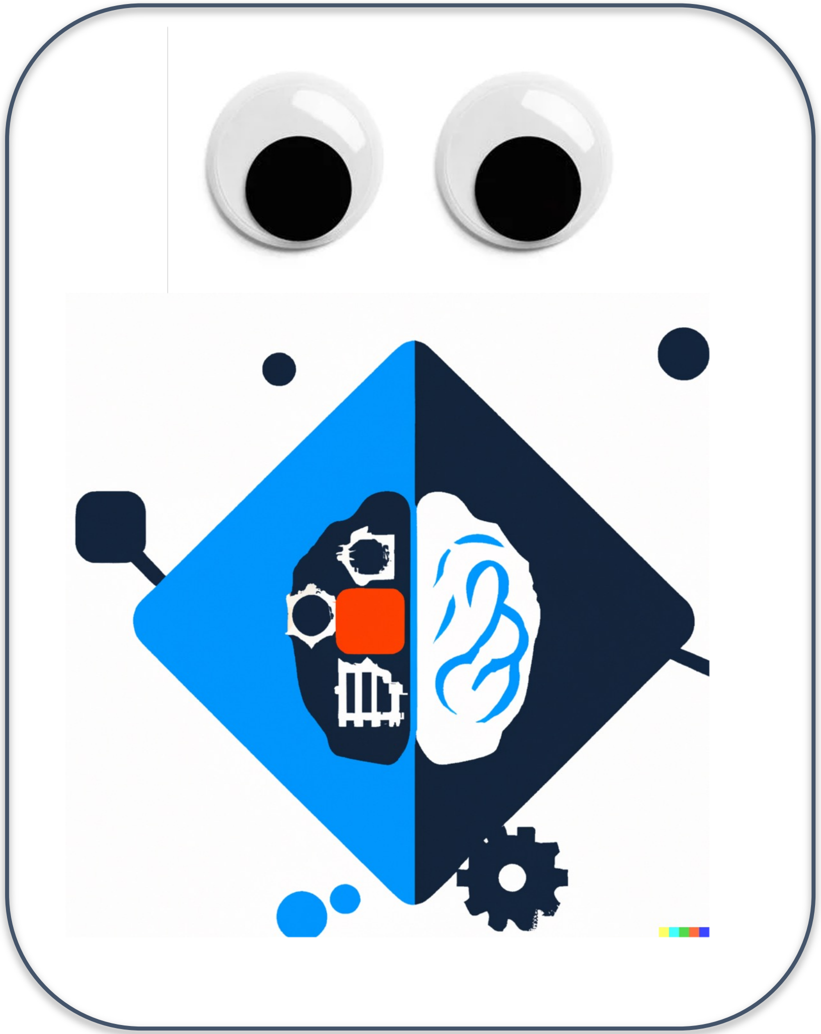
We Need Slower and Deeper Language Reasoning

- ★ Communication is a **collaborative** effort with **intents** and people tend to "*minimize the total effort spent*". [Least collaborative effort]

- ★ Effective communications require "*reaching **mutual beliefs and knowledge** among participants called grounding*". Common sense serves a critical role in building such knowledge [Common Ground]

- ★ Due to least collaborative effort, we need to **make inferences to draw conclusions about the speaker's intentions, emotion states, and experiences.** [Build Common Ground]

AI Companion



*Logo imagined by DALL-E

Ohh, I know

Deep communication abilities

- ➢ Pragmatics
- ➢ Understanding Intent
- ➢ Commonsense Inferences
- ➢ Theory-of-Mind

Us