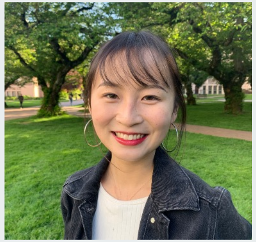
Explanation in the Era of LLMs

NAACL 2024 tutorial
Section 2: **Prompting-based Explanations**



Veronica Qing Lyu
University of Pennsylvania

Hanjie Chen
Rice University

Outline of the tutorial

1. Motivation and desiderata
2. **Prompting-based Explanations** ⟸ This section
3. Data attribution
4. Transformer understanding
5. Conclusion and discussion

Prompting-based Explanations

- Extractive rationales / Feature attributions
- Free-text explanations
- Structured explanations

Extractive rationales / Feature attributions

Extractive Rationales

(**short**) snippets in inputs that support outputs

Input

In this movie, ... Plots to take over the world. The acting is great! The soundtrack is run-of-the-mill, but the action more than makes up for it.

Rationale

In this movie, ... Plots to take over the world. The acting is great! The soundtrack is run-of-the-mill, but the action more than makes up for it.

Output

- Positive
- Negative

[DeYoung et al. 2020]

Extractive Rationales

Pipeline models [DeYoung et al. 2020]

Input	Extractor		Rationale	Predictor	Output
X	$g(\cdot)$	Z	$R = X \odot Z$	$f(\cdot)$	$Y = f(R)$

- Hard selection [Lei et al. 2016]

  Z Binary masks

- Soft selection

  Z Continuous scores

Extractive Rationales

Pipeline models [DeYoung et al. 2020]

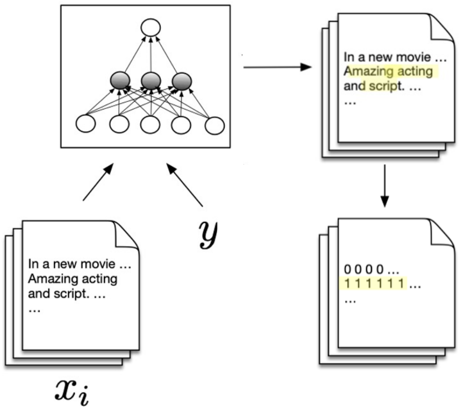
Ground-truth labels

Input	Extractor	Z	Rationale	Predictor	Output
X	$g(\cdot)$		$R = X \odot Z$	$f(\cdot)$	$Y = f(R)$

Ground-truth rationales

- Human annotations
- Pseudo targets

  [Jain et al., ACL 2020]



Extractive Rationales

Pipeline models [DeYoung et al., ACL 2020]

Ground-truth labels

Input	Extractor	Z	Rationale	Predictor	Output
X	$g(\cdot)$		$R = X \odot Z$	$f(\cdot)$	$Y = f(R)$

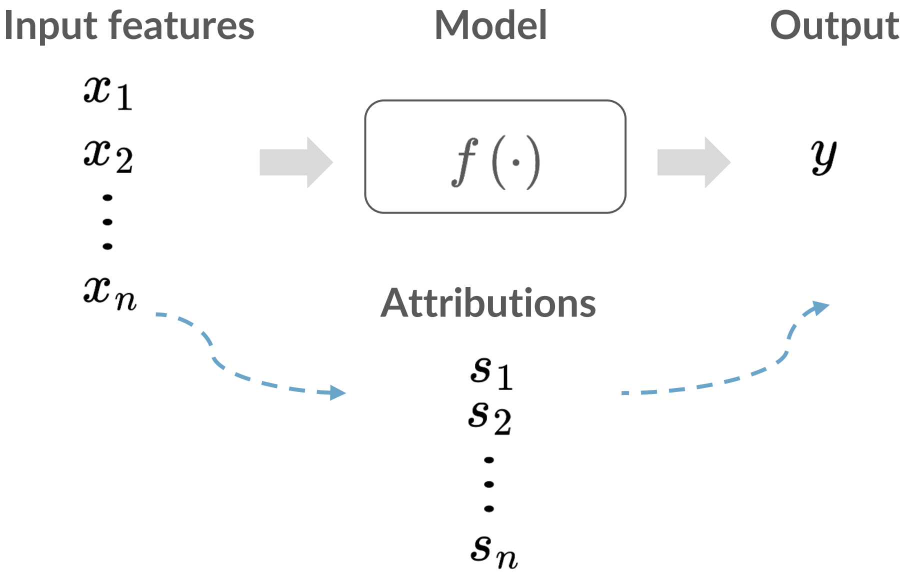
Ground-truth rationales

- Human annotations (Expensive, time-consuming)
- Pseudo targets (Erroneous)

  [Jain et al., ACL 2020]

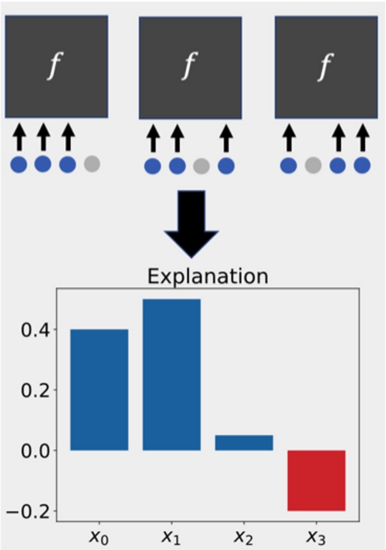
Feature Attributions

Importance scores of input features to model output



Feature Attributions

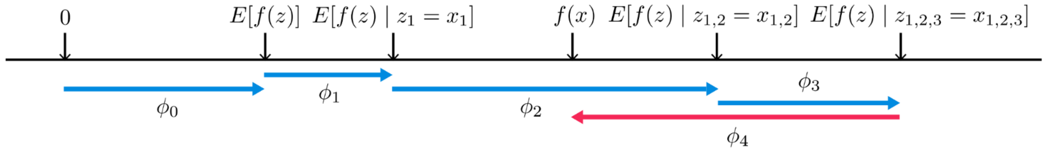
Leave-one-out



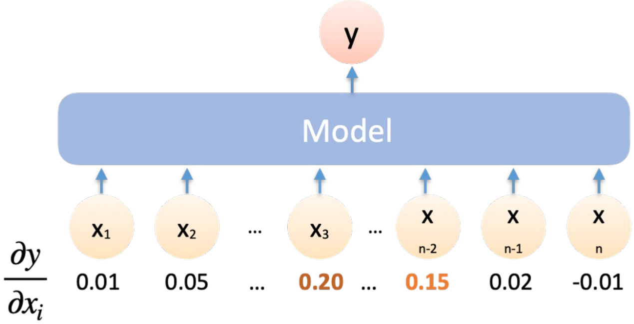
[Covert et al. 2020]

SHAP (SHapley Additive exPlanation)

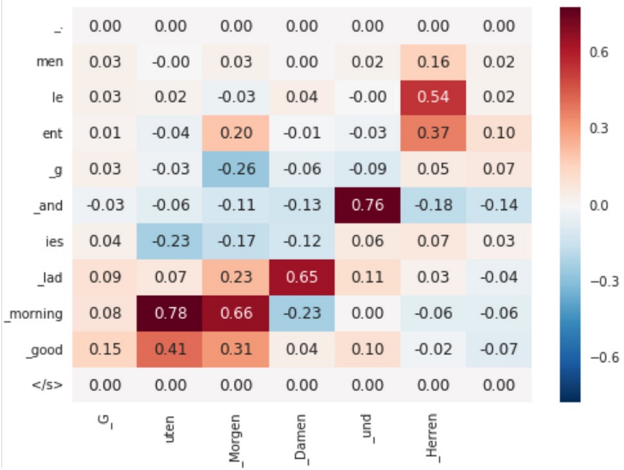
[Lundberg and Lee 2017]



Gradient-based explanation



[Sundararajan et al. 2017]

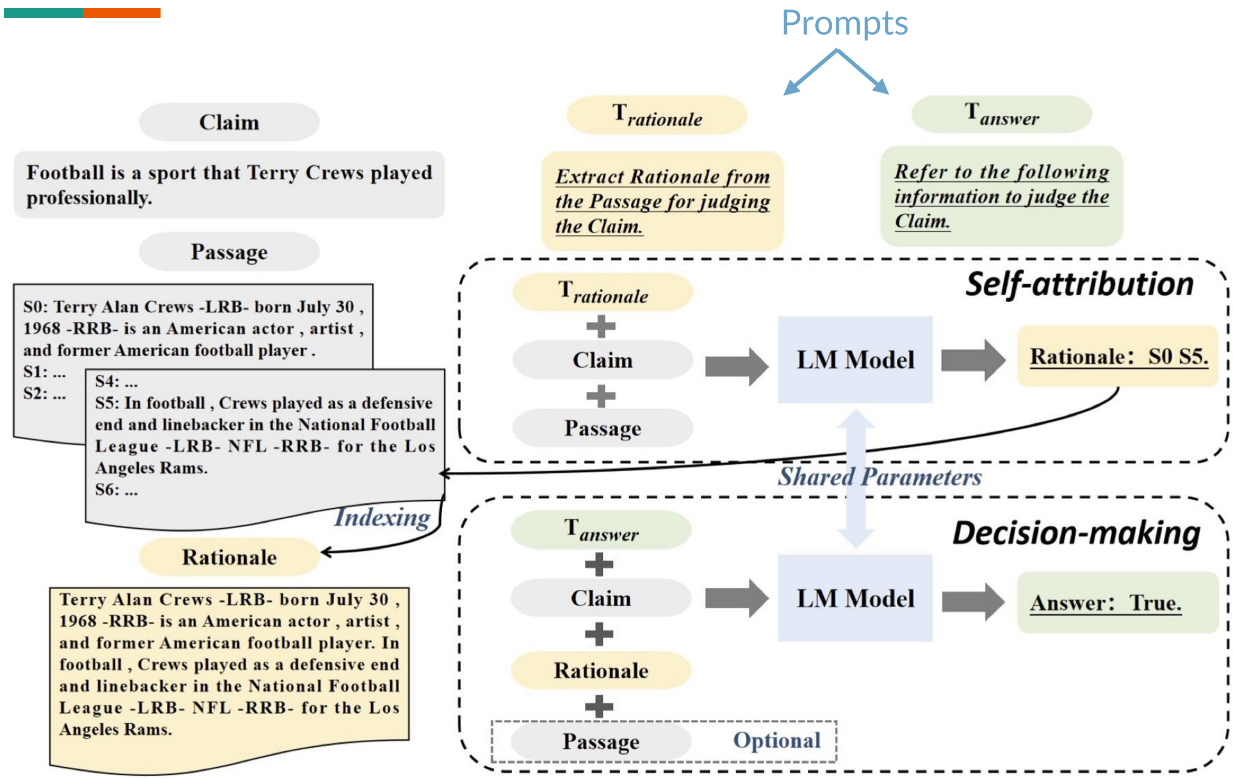


Challenges for LLMs

- Computational cost
- Low efficiency in long context
- No access to API-based models (gradients, attention scores, etc.)

Prompting-based extractive rationales/feature attributions

Self-Attribution and Decision-Making



Stage 1:
Prompting for extractive rationales

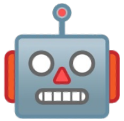
Stage 2:
Making decisions based on rationales

[Du et al. 2023]

See also: [Ludan et al. 2024]

How to evaluate rationales/feature attributions?

Faithfulness



Explanation

Plausibility

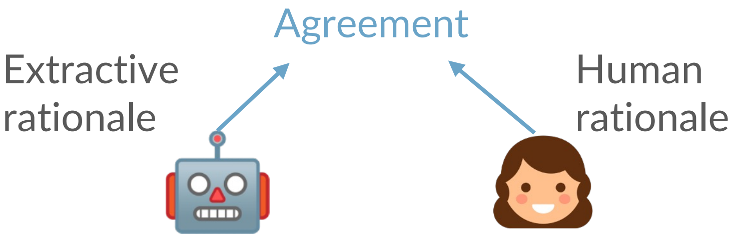
*How accurately the explanation reflects the **true** reasoning process of the model*

*How **convincing** the explanation is to humans*

[Jacovi and Goldberg, ACL 2020; Wiegreffe and Pinter, EMNLP 2019]

Evaluation—Plausibility

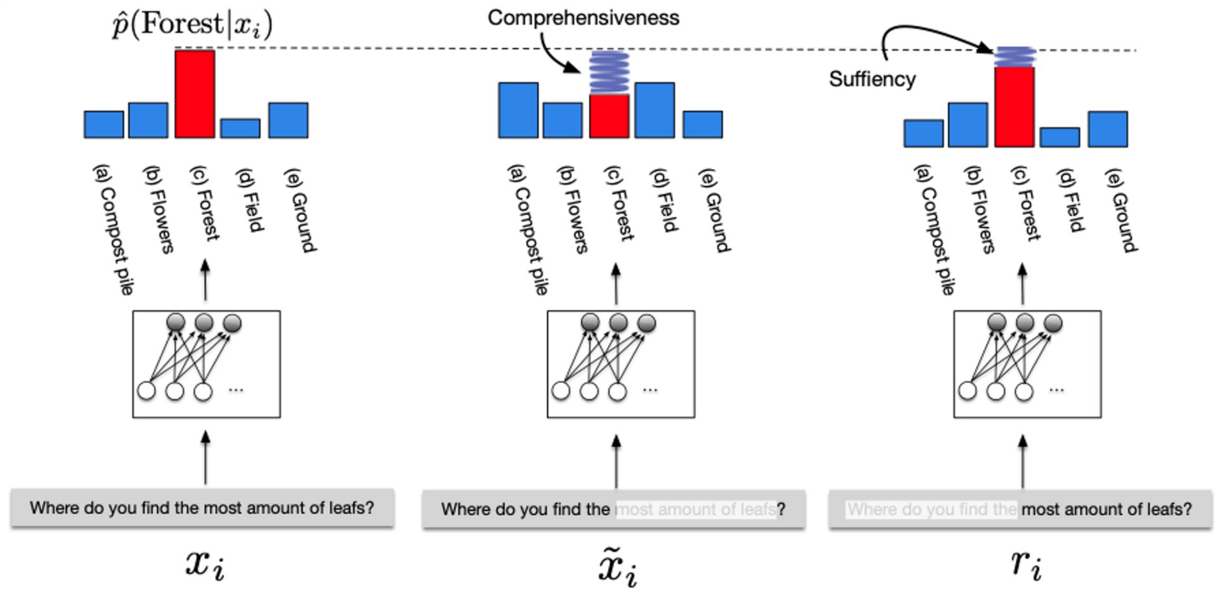
- Agreement
  e.g. Intersection-Over-Union (IOU)



Evaluation—Faithfulness

Comprehensiveness= $f_{\hat{y}}\left(x_i\right)-f_{\hat{y}}\left(x_i \backslash r_i\right)$

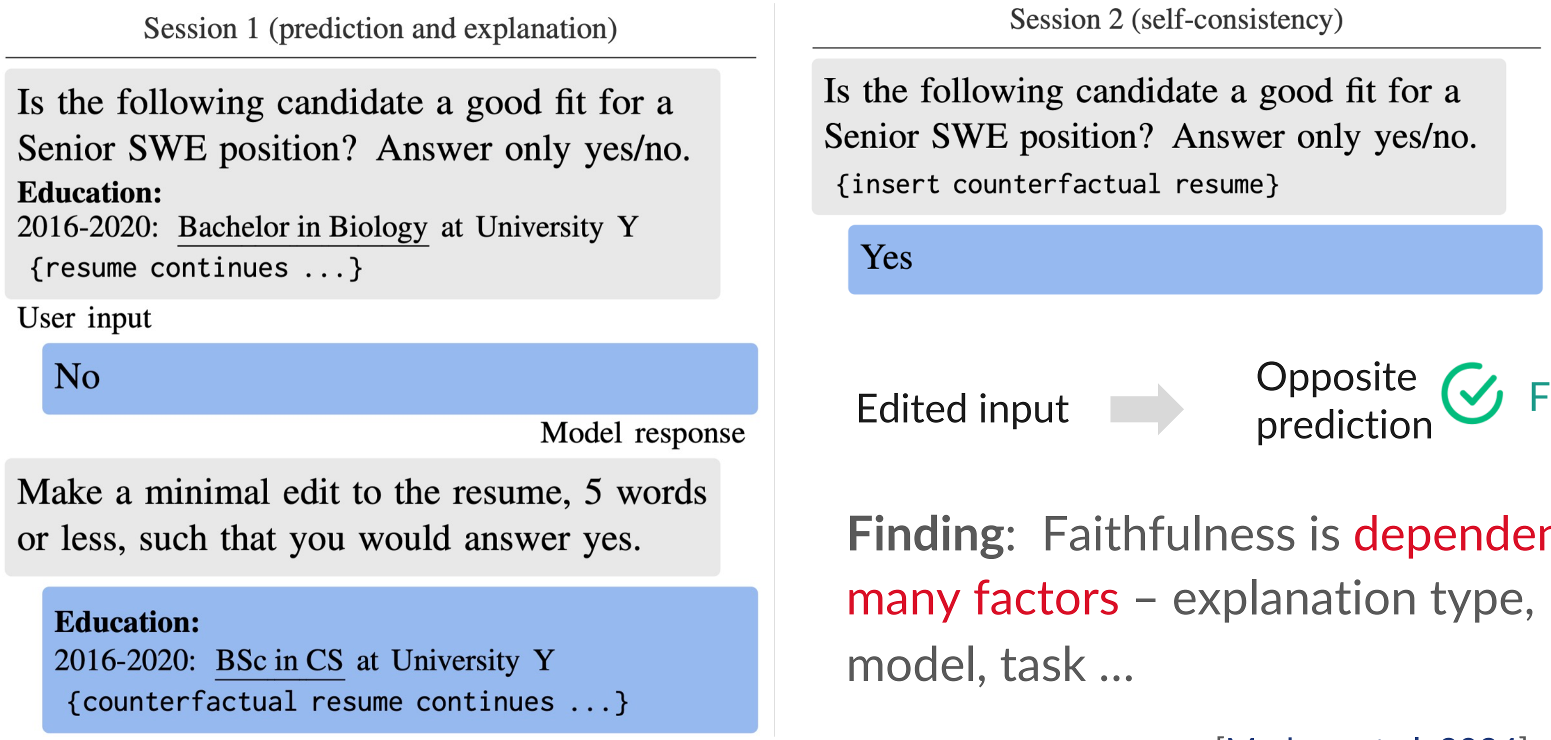
Sufficiency= $f_{\hat{y}}\left(x_i\right)-f_{\hat{y}}\left(r_i\right)$



Fall short in API-based LLMs

[DeYoung et al., ACL 2020]

Evaluation—Faithfulness



Edited input → Opposite prediction ✓ Faithful

Finding: Faithfulness is dependent on many factors – explanation type, model, task ...

[Madsen et al. 2024]

Free-text Explanations

Free-text Explanations

Example: Natural Language Inference (NLI) task

Premise (p)

Kids are on an amusement ride

Hypothesis (h)

Kids are riding their favorite amusement ride

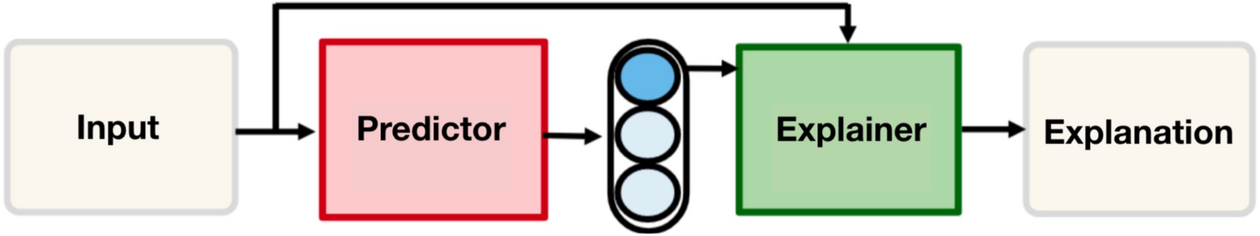
Does the **p** entail **h**?

Model prediction: Maybe

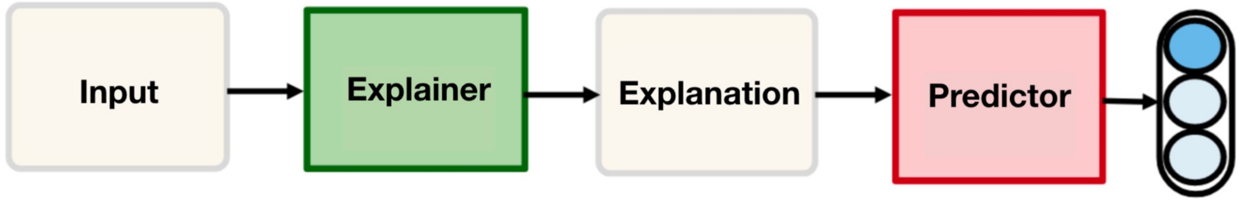
Free-text explanation: It isn't necessarily their favorite ride.

How to Generate Free-text Explanations?

- Traditionally: jointly **train** a predictor & explainer
  - *Predict-then-explain*:



  - *Explain-then-predict*:



[Kumar and Talukdar 2020]

How to Generate Free-text Explanations?

Any cheaper way?

- Traditionally: jointly **train** a predictor & explainer
  - + Can steer models toward using the **"right" signal**
  - - Need lots of human-written explanations as **training data**
    - **Natural Language Inference**: e-SNLI [Camburu et al. 2018]
    - **Commonsense QA**: CoS-E [Rajani et al. 2019], ECQA [Aggarwal et al. 2021]
    - **Social bias inference**: SBIC [Sap et al. 2020]
    - ...

How to Generate Free-text Explanations?

- Can we **prompt** LLMs to generate them with just a few examples?

Example input 1
Example output 1

Example input 2
Example output 2

...

Example input n
Example output n

Input

LLM

Output

Claude Gemini

......

In-context learning / Few-shot prompting [Brown et al. 2021]

Prompting for Explanations

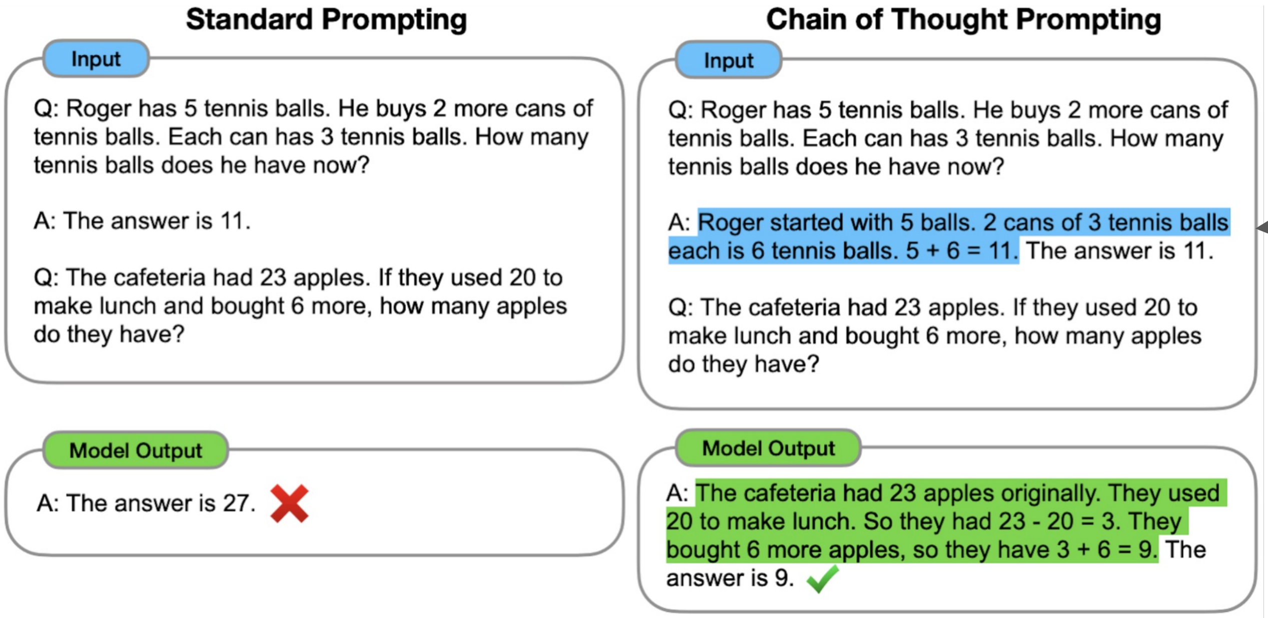
- GPT-3-level LLMs can generate **plausible** free-text explanations for **simple tasks***:
  - NLI
  - Commonsense QA
  - Social bias detection …

- What about **multi-step reasoning**?
  - Maths
  - Multi-hop QA
  - Planning …

*[Wiegreffe et al. 2021; Marasović et al. 2021]

Let's explain classification decisions.
A young boy wearing a tank-top is climbing a tree.
question: A boy was showing off for a girl.
true, false, or neither? neither
why? A boy might climb a tree to show off for a girl, but he also might do it for fun or for other reasons.
###
A person on a horse jumps over a broken down airplane.
question: A person is outdoors, on a horse.
true, false, or neither? true
why? Horse riding is an activity almost always done outdoors. Additionally, a plane is a large object and is most likely to be found outdoors.
###
There is a red truck behind the horses.
question: The horses are becoming suspicious of my apples.
true, false, or neither? false
why? The presence of a red truck does not imply there are apples, nor does it imply the horses are suspicious.

"Chain of Thought" (CoT)



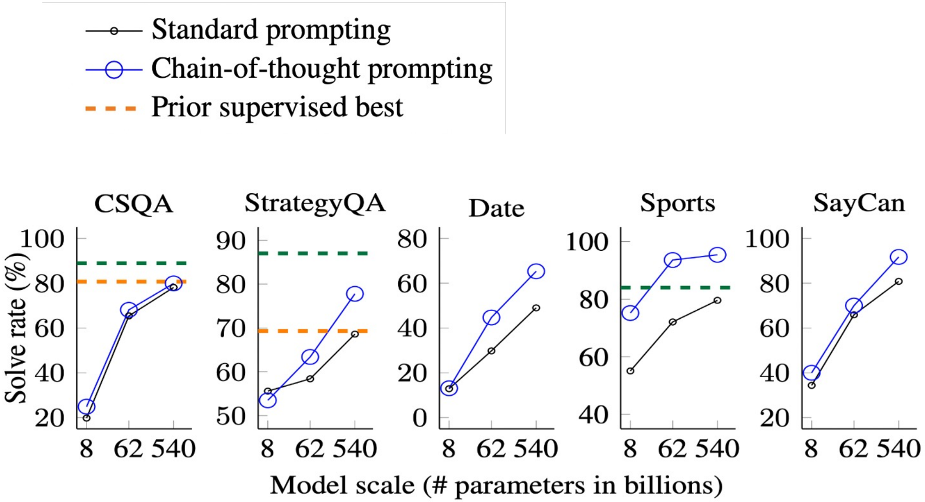
Spell out each step

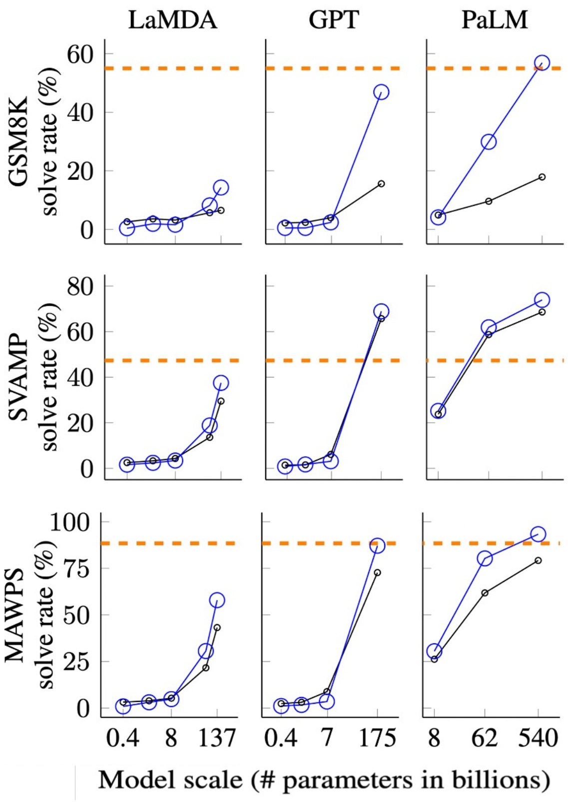
[Wei et al. 2022]

See also: Scratchpad [Nye et al. 2021]; "Let's Think Step by Step" [Kojima et al. 2023]

"Chain of Thought" (CoT)

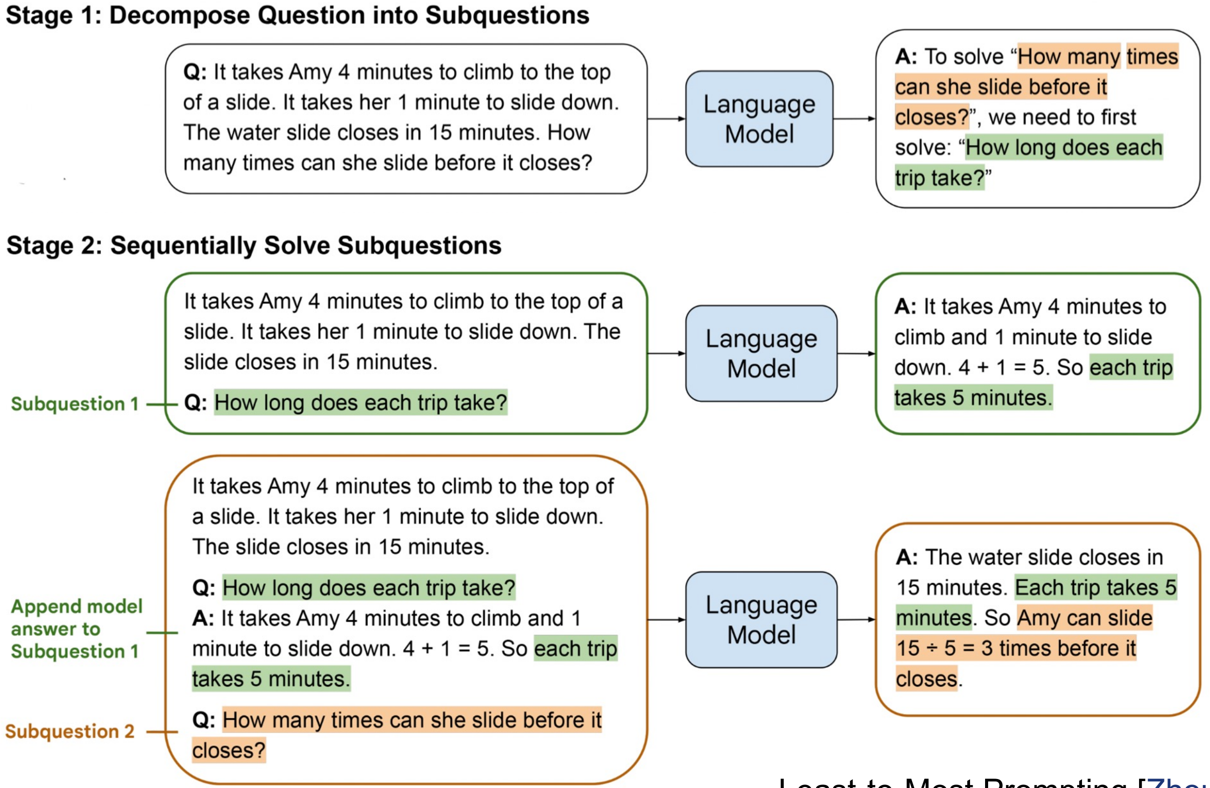
- CoT prompting boosts LLMs' performance on **multi-step reasoning**





Limitation: Easy-to-hard generalization

CoT + Question Decomposition



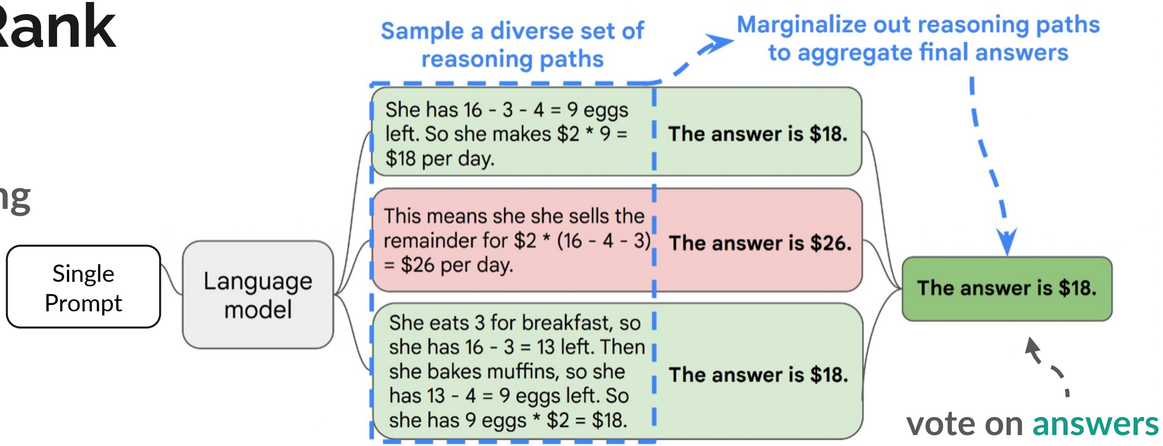
+ Better generalization than CoT

- Greedy decoding has limited diversity

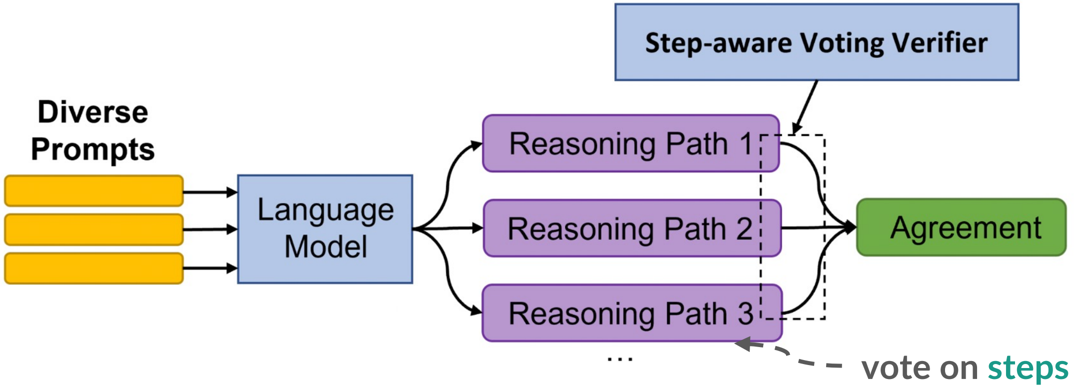
Least-to-Most Prompting [Zhou et al. 2022]

CoT + Vote and Rank

Self-Consistency Prompting
[Wang et al. 2022]



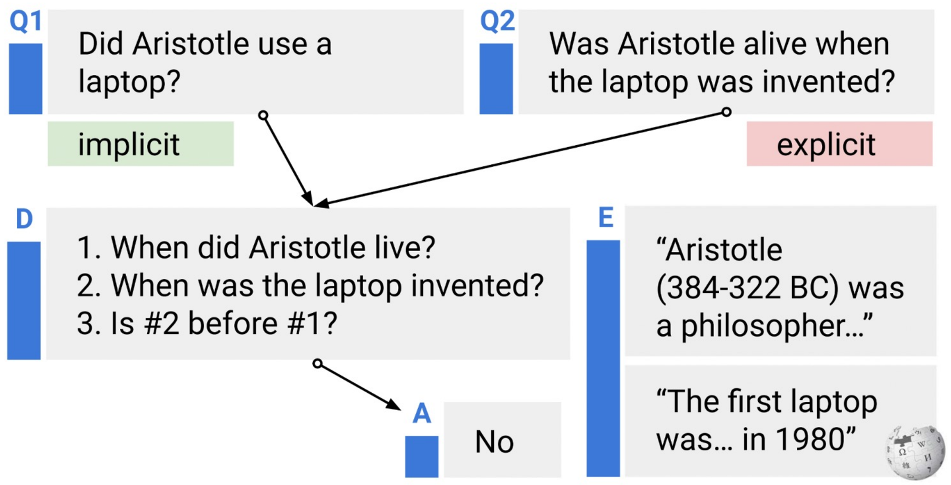
DiVeRSe
[Li et al. 2023]



Structured Explanations

Why Structured Explanations?

- Certain problems intrinsically involve a ***non-linear*** mode of reasoning
  - multi-hop QA, logical deduction, constrained planning...



StrategyQA dataset
[Geva et al. 2021]

Why Structured Explanations?

- Unclear **faithfulness** of free-text explanations
    - False impression of "self-interpretability"
    - Easier over-trust in the model
        - especially if explanations look plausible

Should I hire this candidate?

Generated CoT

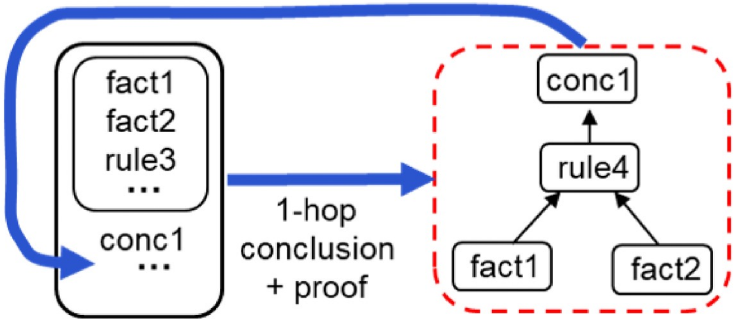
Based on their excellent education background and strong technical skills, I highly recommend hiring this candidate

True Reasoning

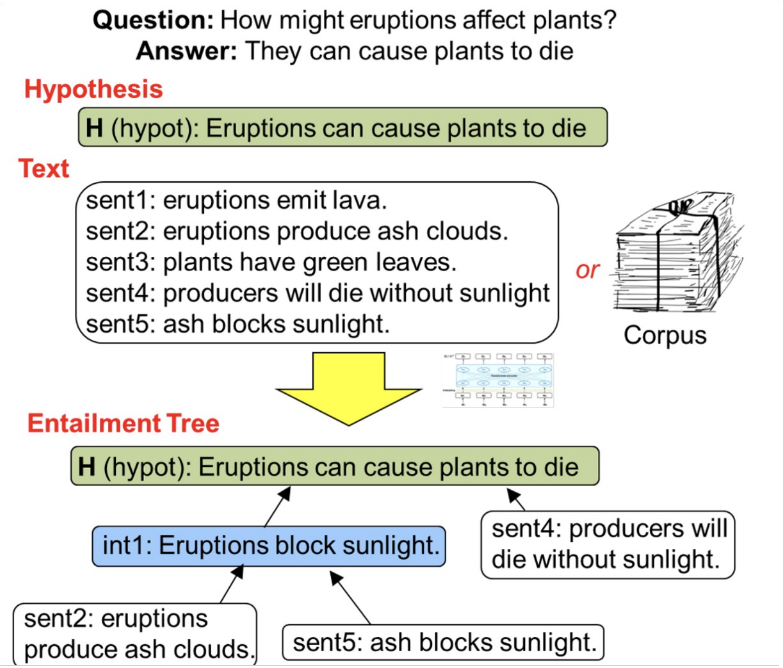
Their name looks like a white male, so I highly recommend hiring this candidate

How to Generate Structured Explanations?

- Traditionally: **train** models to **iteratively** generate intermediate steps



ProofWriter [Tafjord et al 2021]



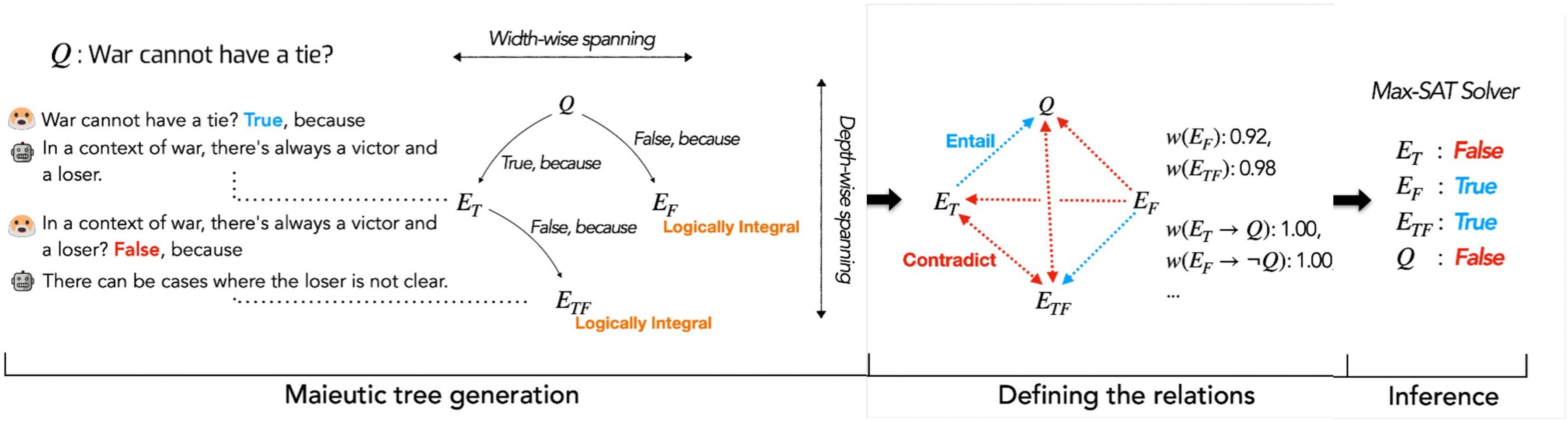
EntailmentWriter [Dalvi et al 2021]

- Still needs lots of (even more expensive) training data

Structured Explanations by Prompting

- Can we prompt LLMs to generate structured explanations with **a few examples**?
- If so, **what types of structures**?
    - Logical constraints
        - Maieutic prompting, SatLM
    - Symbolic programs
        - Program of Thoughts, Program-Aided LMs, Faithful CoT
    - Non-linear exploration strategies
        - Tree of Thoughts, Graph of Thoughts
    - ...

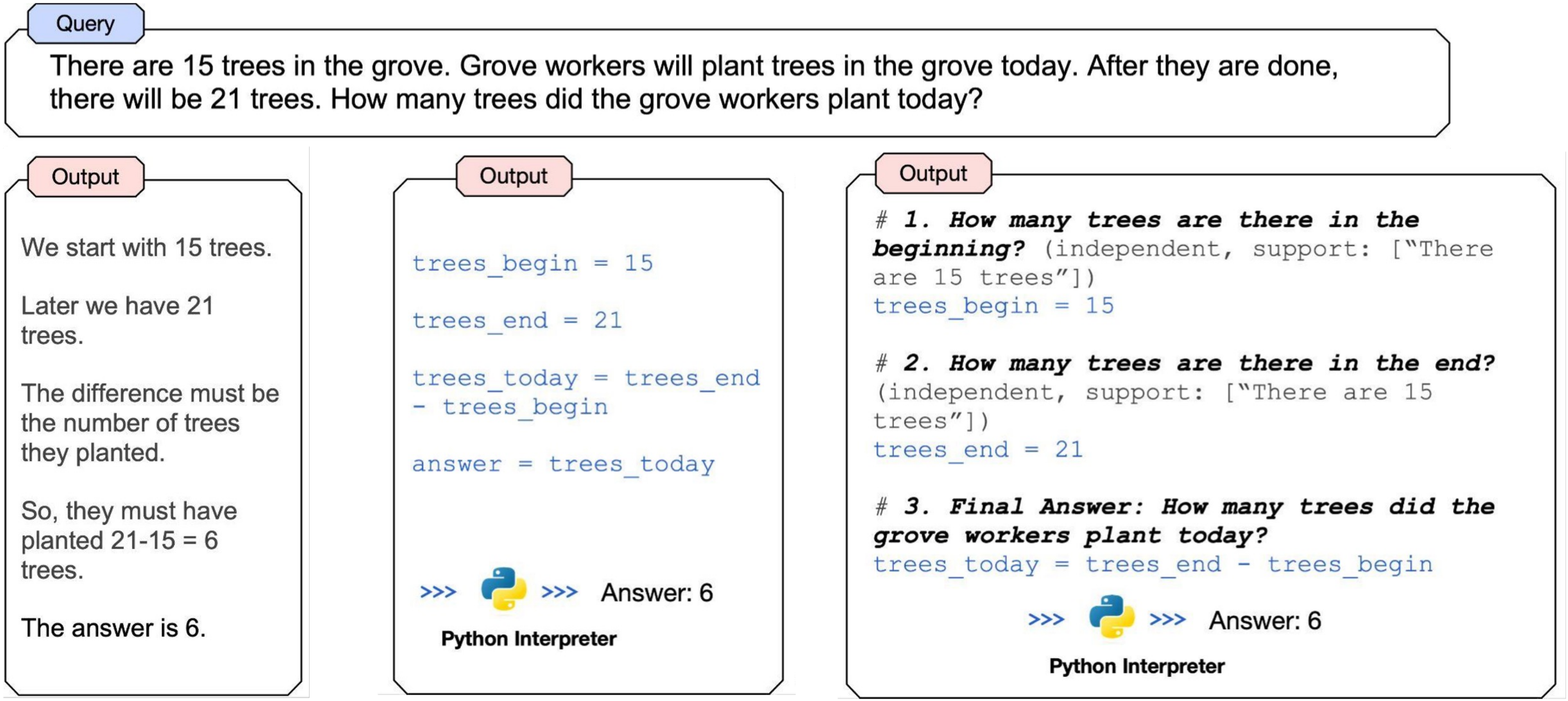
Logically-Constrained Reasoning



Maieutic prompting [Jung et al., 2022]

See also: SatLM [Ye et al., 2023]

Symbolically-Aided Reasoning



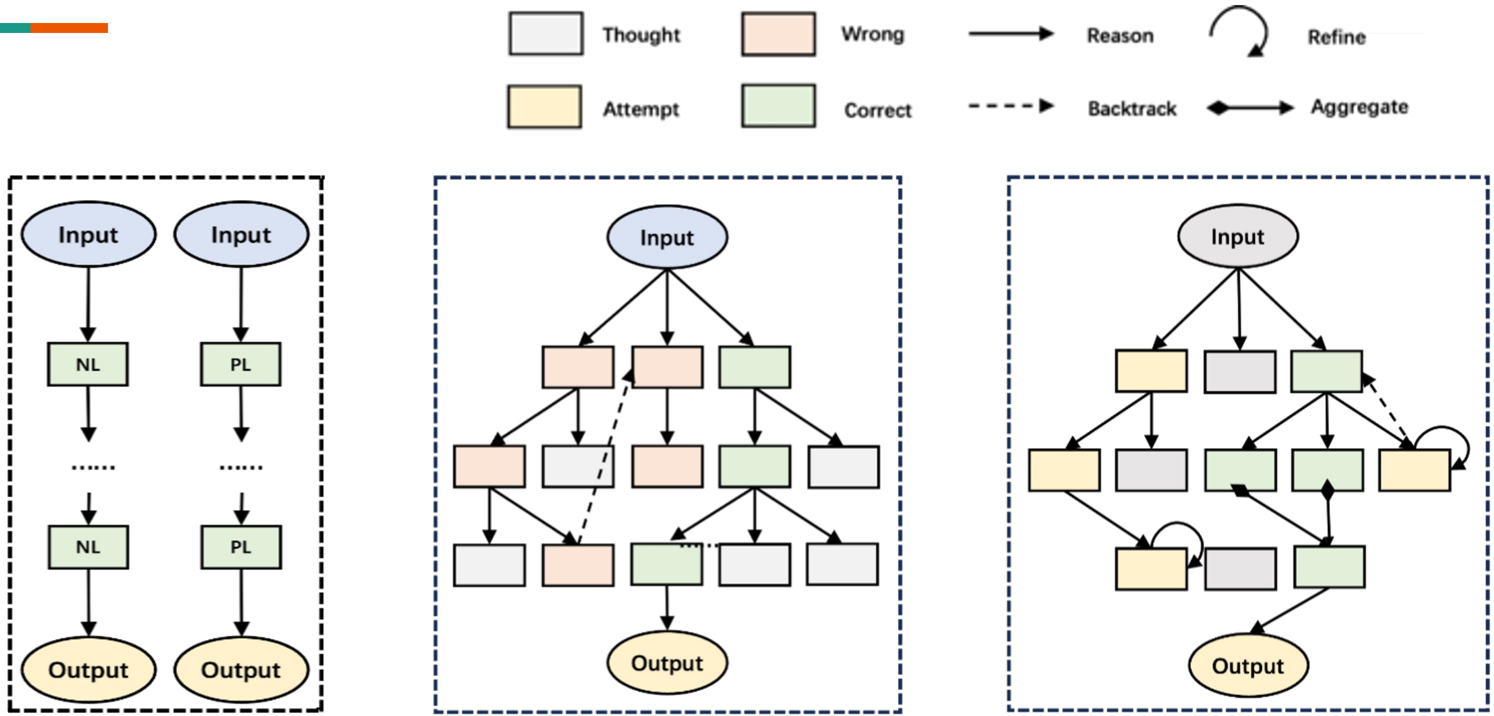
CoT

Program-Aided LM/PAL [Gao et al., 2023]

Program of Thoughts/PoT [Chen et al., 2023]

Faithful CoT [Lyu et al., 2023]

Reasoning with Non-linear Exploration



CoT/PoT   Tree of Thoughts [Yao et al. 2023]   Graph of Thoughts [Besta et al. 2023]

Figure from [Chu et al. 2024]

How to Evaluate Free-text/Structured Explanations?

- **Faithfulness**
  *How **accurately** the explanation reflects the **true** reasoning process of the model?*
- **Plausibility**
  *How **convincing** the explanation is to humans?*
- **Informativeness**
  *How much **new information** is supplied by a explanation to justify the prediction?*
- **Utility**
  *How **useful** is the explanation for the target audience to achieve their predefined **goal**?*
- ….

Most method are also applicable to structured explanations, though empirically only tested on free-text ones

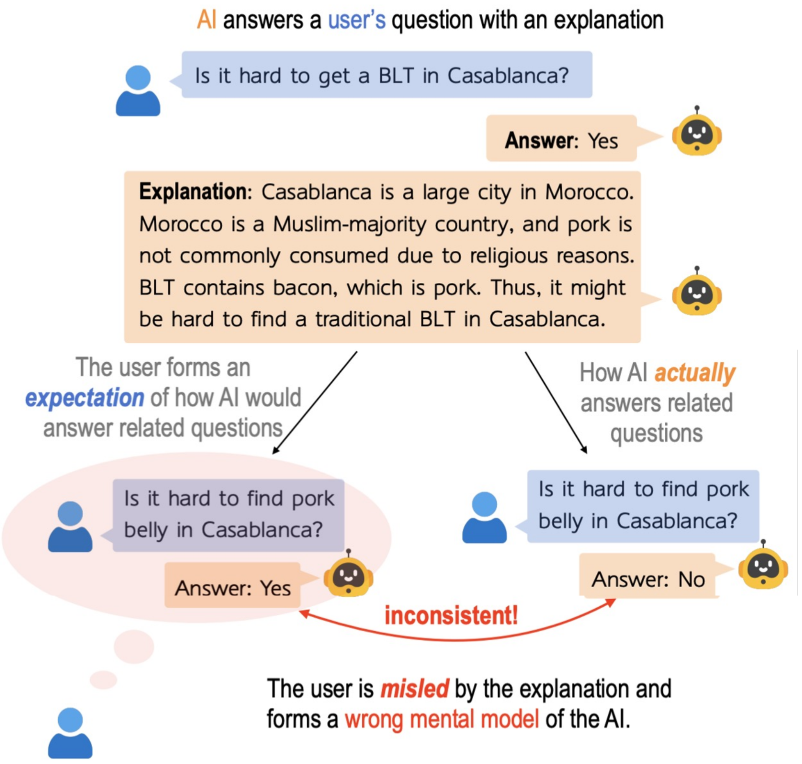
Evaluation—Faithfulness

Many ways with different assumptions, no consensus yet

- **Counterfactual simulatability** [Chen et al., 2023]
  Assumption: *Explanations should allow the audience to **predict** the model behavior on **unseen** inputs*
- **Biasing features** [Turpin et al., 2023]
  Assumption: *Features that **influence** model predictions should be **mentioned** in the explanations*
- **Corrupting CoT** [Lanham et al., 2023]
  Assumption: *Compared to the original explanation, a **corrupted** explanation should lead to a **different** prediction*
- **Input token contribution alignment** [Parcalabescu and Frank, 2024]
  Assumption: *Input token contributions should be **similar** when the model produces the **prediction** and the **explanation***
- …

Evaluation—Faithfulness

Example: Counterfactual simulatability [Chen et al., 2023]

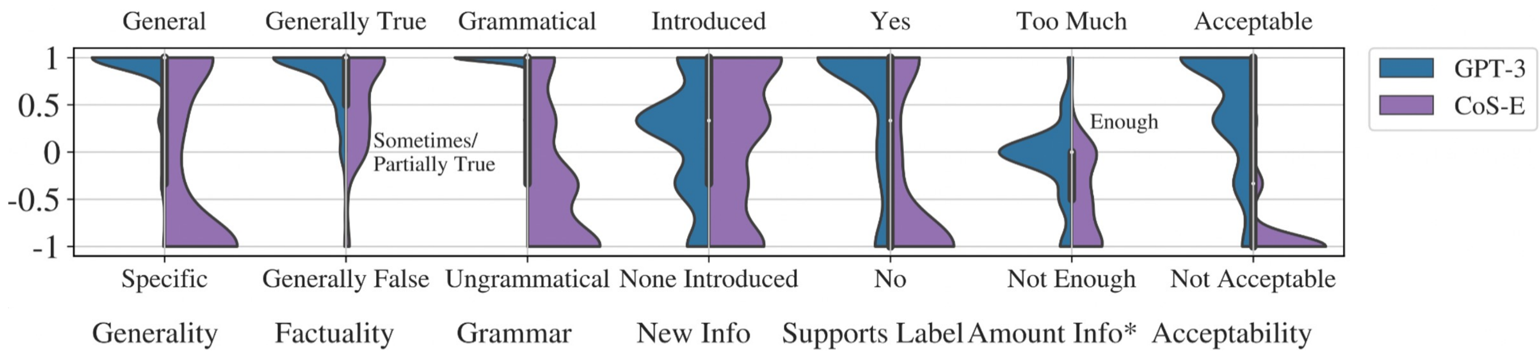


Findings:

- LLM-generated free-text explanations are far from faithful
- Faithfulness doesn't correlate well with plausibility

Evaluation—Plausibility

Annotate LLM-generated explanations with human-written explanations as reference



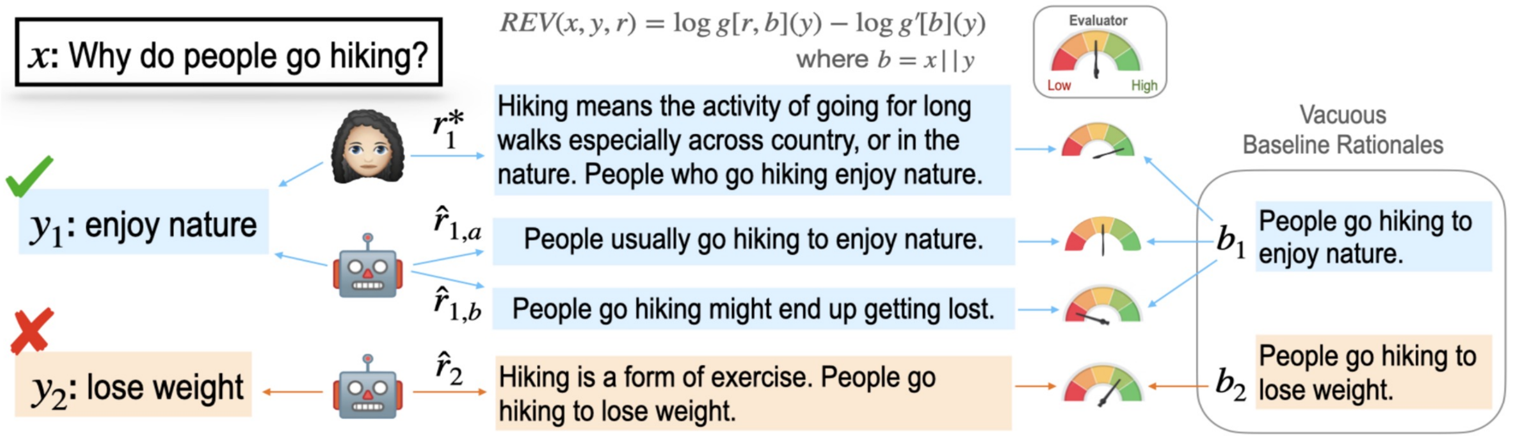
[Wiegreffe et al. 2021]

LLMs can generate plausible explanations, but still have room for improvement compared to human-written ones

See also: ReCEval [Prasad et al. 2021] & SocREval [He et al. 2024] for auto-eval

Evaluation—Informativeness

Measure the **new information** an explanation provides to justify the label, beyond what is contained in the input, using conditional V-information

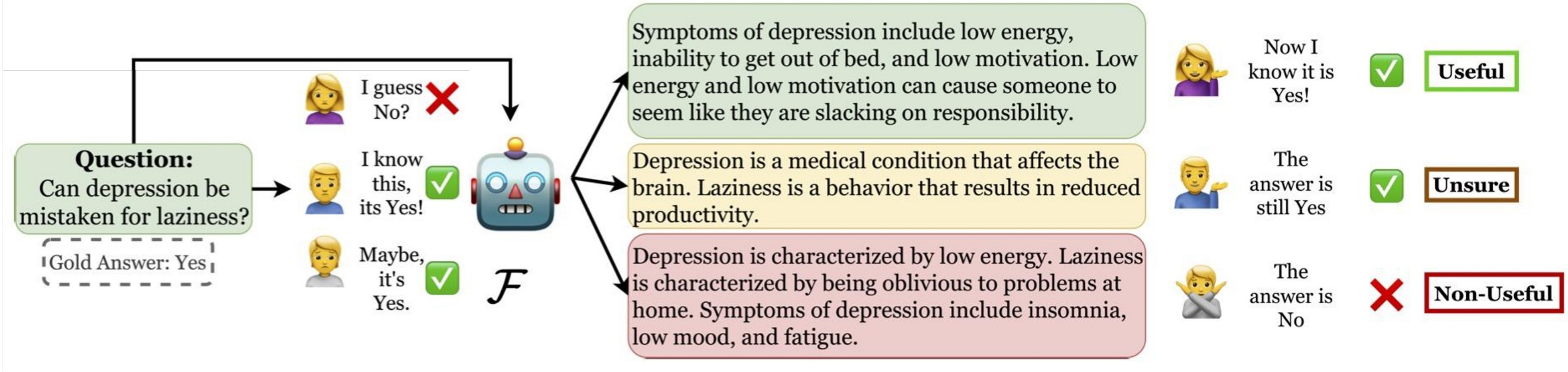


REV [Chen et al. 2023]

See also: [Jiang et al. 2024]

Evaluation—Utility

~~Can LLM~~-generated explanations help **lay people** answer **unseen** questions?



Utility is far from satisfactory – only 20% of generated explanations are actually useful

[Joshi et al. 2023]

Summary

Pros & Cons

- Extractive rationales / Feature attributions
  - ? Faithfulness
  - - Plausibility

- Free-text explanations
  - + Plausibility
  - - Faithfulness, Utility

- Structured explanations
  - + Faithfulness, Accuracy
  - - Flexibility

Takeaways

- LLMs can generate **plausible**-looking explanations w/ only a few examples
    - this saves the cost of collecting human explanations for training
    - and also improves performance on many reasoning tasks
- However, LLM-generated explanations are still **not** always **faithful / informative / useful** ...
    - Not a consensus on how to **evaluate** many of these aspects
- We should not blindly trust LLM-generated explanations
    - Be cautious about "self-explanatory" claims

Future Directions

- Establishing a more unified **evaluation framework**
  - esp. for structured explanations
- Applying structured explanations to **flexible** (non-symbolic) tasks
  - e.g. commonsense reasoning, summarization, web browsing …

Further Reading

- A Comprehensive Collection of Explainable NLP Datasets [Wiegreffe and Marasović 2021]
- A Survey on Chain-of-Thought-style Reasoning [Chu et al. 2024]
- A Survey on Faithfulness of Explanations in NLP [Lyu et al. 2024]

Thanks! Questions?