



Computer  
Science



# Symbolic Reasoning about Large Language Models

Guy Van den Broeck

Apple Workshop on Reasoning and Planning - Jul 23 2025



## Reasoning with Symbolic AI

- Logic and probabilistic
- Deductive reasoning algorithms
- Correct on *all* problems
- Limited scope
- Intractable



30 years



## Reasoning with Transformers

- Build chains of thought
- Inductive reasoning from data
- Correct on *many* problems
- Unlimited scope
- Tractable

*Do deductive reasoning algorithms still have a purpose in the age of transformers?*

**Symbolic Probabilistic Reasoning About All Future Tokens**

# Reasoning about all Future Tokens: *Constraints*

$p(\text{next-token} \mid \alpha, \text{prefix})$

**Constrained Generation:**  $\Pr(x_{t+1} \mid \alpha, x_{1:t} = \text{"the weather is"})$

**Lexical Constraint**  $\alpha$ : sentence contains keyword "winter"

# Reasoning about all Future Tokens: *Constraints*

$$p(\text{next-token} \mid \alpha, \text{prefix})$$

cold	0.025
warm	0.001

$$\propto p(\text{next-token} \mid \text{prefix})$$

cold	0.05
warm	0.10

**Constrained Generation:**  $\Pr(x_{t+1} \mid \alpha, x_{1:t} = \text{"the weather is"})$

**Lexical Constraint**  $\alpha$ : sentence contains keyword "winter"

$$p(\alpha \mid \text{next-token}, \text{prefix})$$

cold	0.50
warm	0.01



# Reasoning about all Future Tokens: *Alignment*

$p(\text{next-token} \mid \alpha, \text{prefix})$

**Prefix:** It's a pain ...

**Constraint  $\alpha$ :** non-toxic

# Reasoning about all Future Tokens: *Alignment*

$$p(\text{next-token} \mid \alpha, \text{prefix})$$

**Prefix:** It's a pain ...

**Constraint  $\alpha$ :** non-toxic

$$\propto p(\text{next-token} \mid \text{prefix}) \cdot p(\alpha \mid \text{next-token}, \text{prefix})$$

in	0.3	the ass	0.3
to	0.1	the butt	0.15
		the neck	0.05
		deal with	0.2
		handle	0.1
		...	...



# Reasoning about all Future Tokens: *Alignment*

$$p(\text{next-token} \mid \alpha, \text{prefix})$$

in	0.03
to	0.08

**Prefix:** It's a pain ...

**Constraint  $\alpha$ :** non-toxic

$$\propto p(\text{next-token} \mid \text{prefix})$$

in	0.3
to	0.1

$$p(\alpha \mid \text{next-token}, \text{prefix})$$

in	0.1
to	0.8



# Reasoning about all Future Tokens

$$p_{lm}(\text{next-token} \mid \alpha, \text{prefix})$$

$$\propto p_{lm}(\text{next-token} \mid \text{prefix})$$

Using Bayes rule,

$$\cdot \cancel{p_{lm}(\alpha \mid \text{next-token}, \text{prefix})}$$

*Intractable*



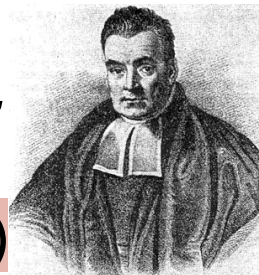


# Reasoning about all Future Tokens

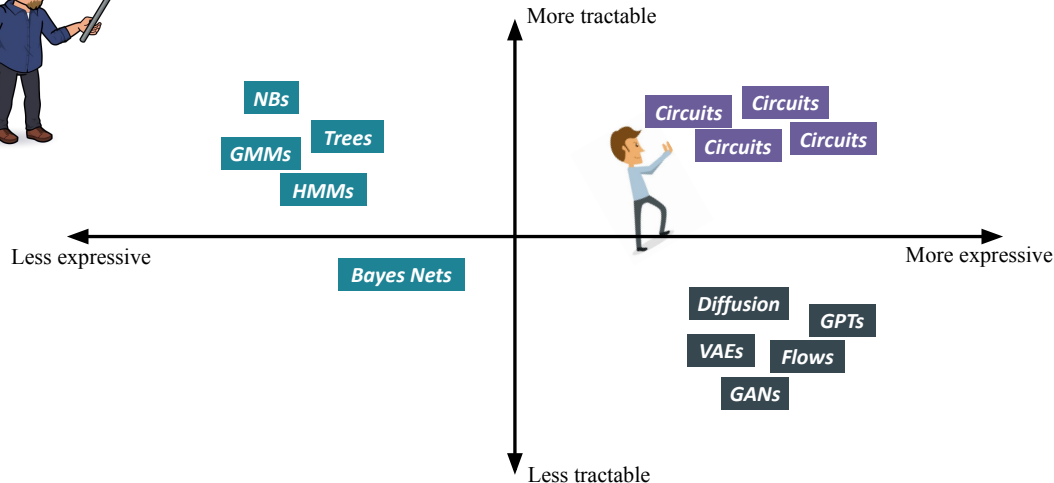
$$p_{lm}(\text{next-token} \mid \alpha, \text{prefix})$$

*Abusing Bayes rule,*

$$\propto p_{lm}(\text{next-token} \mid \text{prefix}) \cdot p_{circuit}(\alpha \mid \text{next-token}, \text{prefix})$$



Use a **tractable** probabilistic **circuit** language model distilled from the transformer language model...

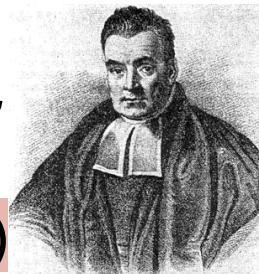


# Reasoning about all Future Tokens: Constraints

$$p_{lm}(\text{next-token} \mid \alpha, \text{prefix})$$

*Abusing Bayes rule,*

$$\propto p_{lm}(\text{next-token} \mid \text{prefix}) \cdot p_{circuit}(\alpha \mid \text{next-token}, \text{prefix})$$



**Theorem.** Given

1. a deterministic finite automata constraint  $\alpha$  with  $m$  edges and
  2. a probabilistic circuit  $p(\cdot)$  with  $h$  hidden states  
(representing a Hidden Markov Model) ,
- computing  $p(\alpha \mid x_{1:t})$  over a sequence of  $n$  future tokens takes  $O(nmh^2)$  time.

# Interactive Text Editing

"First they've defeated a small squad [BLANK] are few humans left, and despite their magical power, their numbers are getting fewer."

# Interactive Text Editing

User: given the following context, generate infilling text for [BLANK] using key phrases "alien mothership", "far from over"; generated text must contain 25 - 30 words.

"First they've defeated a small squad [BLANK] are few humans left, and despite their magical power, their numbers are getting fewer."



"First they've defeated a small squad of aliens, then a larger fleet of their ships. Eventually they've even managed to take down the alien mothership. But their problems are far from over. There are few humans left, and despite their magical power, their numbers are getting fewer."

# Interactive Text Editing with key phrase (K) or length (L) constraints

CoAuthor



	<i>None</i>	<i>K</i>	<i>L</i>	<i>K&amp;L</i>
--	-------------	----------	----------	----------------

<i>Quality</i>				
----------------	--	--	--	--

TULU2	2.68	2.64	2.78	2.74
-------	------	------	------	------

GPT3.5	2.27	2.22	2.27	2.31
--------	------	------	------	------

GPT4	<b>3.79</b>	3.33	3.53	3.10
------	-------------	------	------	------

Ctrl-G	<b>3.77</b>	<b>3.56</b>	<b>3.73</b>	<b>3.59</b>
--------	-------------	-------------	-------------	-------------

→ *How many stars by humans?*

# Interactive Text Editing with key phrase (K) or length (L) constraints



	<i>None</i>	<i>K</i>	<i>L</i>	<i>K&amp;L</i>
<i>Quality</i>				
TULU2	2.68	2.64	2.78	2.74
GPT3.5	2.27	2.22	2.27	2.31
GPT4	<b>3.79</b>	3.33	3.53	3.10
Ctrl-G	<b>3.77</b>	<b>3.56</b>	<b>3.73</b>	<b>3.59</b>
<i>Success</i>				
TULU2	-	12%	20%	3%
GPT3.5	-	22%	54%	10%
GPT4	-	60%	20%	27%
Ctrl-G	-	<b>100%</b>	<b>100%</b>	<b>100%</b>

→ *How many stars by humans?*

→ *Follows instructions?*

# Interactive Text Editing with key phrase (K) or length (L) constraints



	None	K	L	K&L
<i>Quality</i>				
TULU2	2.68	2.64	2.78	2.74
GPT3.5	2.27	2.22	2.27	2.31
GPT4	<b>3.79</b>	3.33	3.53	3.10
Ctrl-G	<b>3.77</b>	<b>3.56</b>	<b>3.73</b>	<b>3.59</b>
<i>Success</i>				
TULU2	-	12%	20%	3%
GPT3.5	-	22%	54%	10%
GPT4	-	60%	20%	27%
Ctrl-G	-	<b>100%</b>	<b>100%</b>	<b>100%</b>
<i>Overall</i>				
TULU2	-	7%	10%	1%
GPT3.5	-	0%	5%	2%
GPT4	-	41%	17%	14%
Ctrl-G	-	<b>76%</b>	<b>78%</b>	<b>82%</b>

→ *How many stars by humans?*

→ *Follows instructions?*

→ ★★☆☆☆ & Up + *Follows instructions?*

→ **Ctrl-G based on Llama2-7B wipes the floor with GPT4, which is a >100x bigger LLM**

# Grade School Math Benchmark

**Question:** *Kylar went to the store to buy glasses for his new apartment. One glass costs \$5, but every second glass costs only 60% of the price. Kylar wants to buy 16 glasses. How much does he need to pay for them?*

**Vanilla LLM Answer:** The price of the 2nd glass is  $(16 / 2) * 60\% = 8$  dollars. So one pair of glasses costs  $16 + 8 = 24$  dollars. **So the answer is 24.**



# Grade School Math Benchmark

**Question:** *Kylar went to the store to buy glasses for his new apartment. One glass costs \$5, but every second glass costs only 60% of the price. Kylar wants to buy 16 glasses. How much does he need to pay for them?*

**Vanilla LLM Answer:** The price of the 2nd glass is  $(16 / 2) * 60\% = 8$  dollars. So one pair of glasses costs  $16 + 8 = 24$  dollars. **So the answer is 24.**

**Ctrl-G Answer:** The second glass costs  $5 * .6 = \$3$ . So each set of two glasses actually costs  $5 + 3 = \$8$ . He wants  $16 / 2 = 8$  sets of two. That means he needs to pay  $8 * 8 = \$64$ . **So the answer is 64.**

*Which constraint improves accuracy?*

# Grade School Math Benchmark

**Question:** *Kylar went to the store to buy glasses for his new apartment. One glass costs \$5, but every second glass costs only 60% of the price. Kylar wants to buy 16 glasses. How much does he need to pay for them?*

**Vanilla LLM Answer:** The price of the 2nd glass is  $(16 / 2) * 60\% = 8$  dollars. So one pair of glasses costs  $16 + 8 = 24$  dollars. So the answer is 24.

**Ctrl-G Answer:** The second glass costs  $5 * .6 = \$3$ . So each set of two glasses actually costs  $5 + 3 = \$8$ . He wants  $16 / 2 = 8$  sets of two. That means he needs to pay  $8 * 8 = \$64$ . So the answer is 64.

## Use all the numbers in the problem statement!

# Advantages of Ctrl-G:

1. Constraint  $\alpha$  is guaranteed to be satisfied:  
if  $x_{t+1}$  makes  $\alpha$  unsatisfiable,  $p(x_{t+1} \mid x_{1:t}, \alpha) = 0$ .
2. Generalizes well to unseen reasoning tasks, because all tasks are unseen :-)  
(training on a distribution over tasks is slow and brittle!)
3. Bayesian = goal-oriented ( $\leftrightarrow$  structured generation tools)

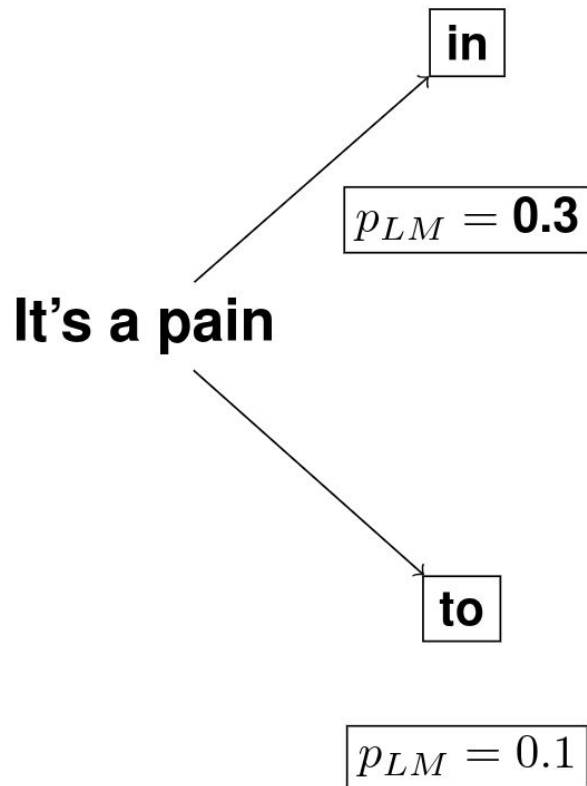
You can control an intractable generative model using a generative model that is *tractable for symbolic reasoning*.

## Attribute Probability



0 (toxic)

1 (nontoxic)



future text	$p_{LM}(x_{>t} \mid x_{\leq t})$
the ass	0.3
the butt	0.15
the neck	0.05
...	...
...	...

Intractable to know future  
expected attribute probability (EAP)

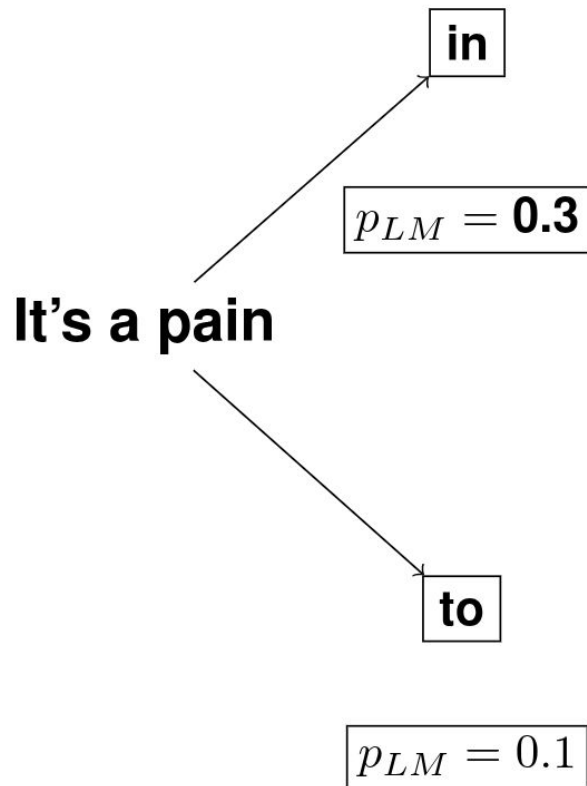


future text	$p_{LM}(x_{>t} \mid x_{\leq t})$
deal with	0.2
handle	0.1
...	...
...	...

## Attribute Probability



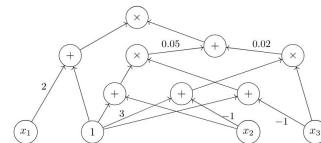
0 (toxic)      1 (nontoxic)



future text	$p_{TPM}(x_{>t} \mid x_{\leq t})$
the ass	0.3
the butt	0.15
the neck	0.05
...	...
...	...

future text	$p_{TPM}(x_{>t} \mid x_{\leq t})$
deal with	0.2
handle	0.1
...	...
...	...

Tractable Prob.  
Circuit Model



+ Log-Linear  
Attribute Classifier



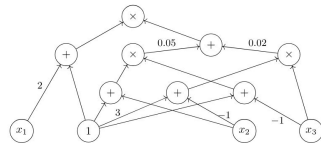
1 (nontoxic)

$p_{LM} = \mathbf{0.3}$

# It's a pain

$$p_{LM} = 0.1$$
$$EAP = 0.1$$

# Tractable Probabilistic Model



## + Log-Linear Attribute Classifier

$$EAP = 0.8$$

==

## Efficient Expected Attribute Probability!



# Attribute Probability



0 (toxic)

1 (nontoxic)

It's a pain

in

$$p_{LM} = \mathbf{0.3} \times$$

future text	$p_{TPM}(x_{>t} \mid x_{\leq t})$
the ass	0.3
the butt	0.15
the neck	0.05
...	...
...	...

$$EAP = 0.1$$

$$= p_{TRACE} \propto 0.03$$

to

$$p_{LM} = 0.1 \times$$

future text	$p_{TPM}(x_{>t} \mid x_{\leq t})$
deal with	0.2
handle	0.1
...	...
...	...

$$EAP = 0.8$$

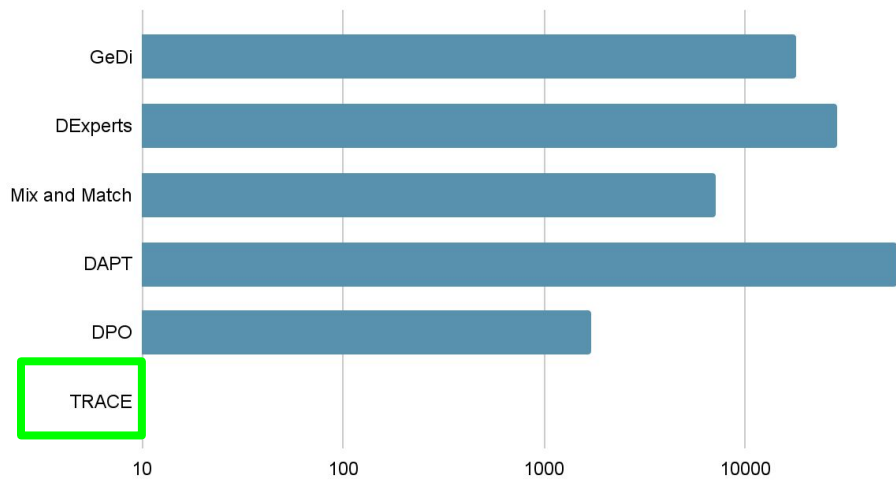
$$= p_{TRACE} \propto \mathbf{0.08}$$



# TRACE is Blazingly Fast

Given a language model, and its tractable proxy model,  
train log-linear attribute classifier

Training Time per Attribute (seconds)

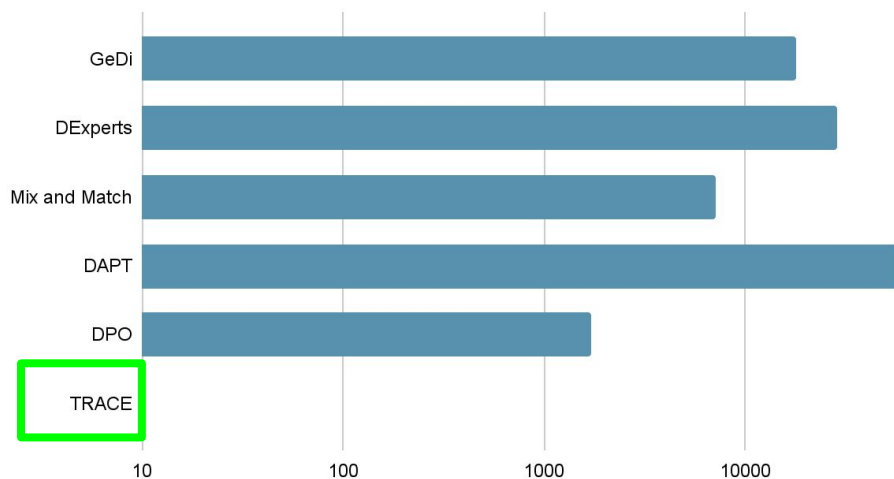




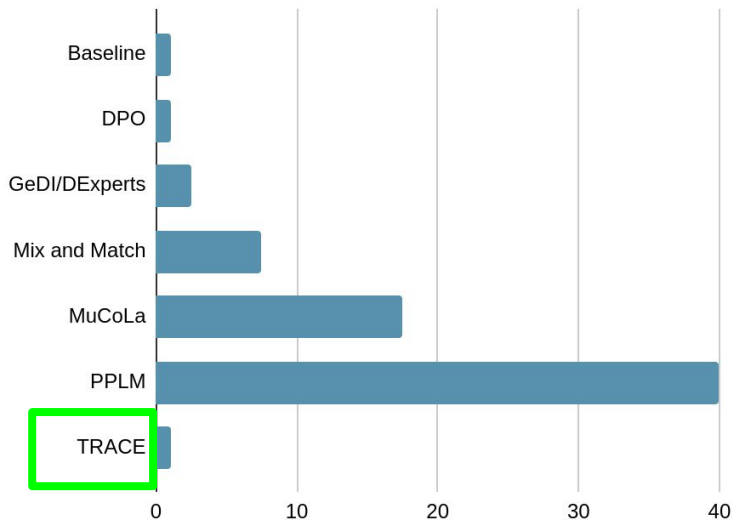
# TRACE is Blazingly Fast

Given a language model, and its tractable proxy model,  
train log-linear attribute classifier,  
then use Bayesian logits at decoding time

Training Time per Attribute (seconds)



Inference Time



# State-of-the-art LLM Detoxification

Model	Toxicity (↓)		Approach Type
	avg. max.	prob.	
GPT-2 Large Results			
GPT2	0.385	0.254	Baseline
DAPT <sup>(1)</sup>	0.428	0.360	Finetuning
GeDi <sup>(2)</sup>	0.363	0.217	Decoding (Trained Guide)
FUDGE <sup>(3)</sup>	0.302	0.371	Decoding (Trained Guide)
DExperts <sup>(4)</sup>	0.314	0.128	Decoding (Trained Guide)
PPLM <sup>(5)</sup>	0.520	0.518	Decoding (Logit Control)
MuCoLa <sup>(6)</sup>	0.308	0.088	Decoding (Sampling)
PPO <sup>(7)</sup>	0.218	0.044	RL
Quark <sup>(8)</sup>	0.196	0.035	RL
DPO <sup>(9)</sup>	0.180	0.026	RL
TRACE	<b>0.163</b>	<b>0.016</b>	Decoding (HMM Reasoning)
Gemma-2B Results			
Gemma-2B	0.359	0.23	Baseline
DPO <sup>(9)</sup>	0.222	0.06	RL
TRACE	<b>0.189</b>	<b>0.02</b>	Decoding (HMM Reasoning)

# State-of-the-art LLM Detoxi

Model	Toxicity (↓)		Diversity (↑)	
	avg.	max. prob.	dist-2	dist-3
GPT-2 Large Results				
GPT2	0.385	0.254	0.87	0.86
DAPT <sup>(1)</sup>	0.428	0.360	0.84	0.84
GeDi <sup>(2)</sup>	0.363	0.217	0.84	0.83
FUDGE <sup>(3)</sup>	0.302	0.371	0.78	0.82
DExperts <sup>(4)</sup>	0.314	0.128	0.84	0.84
PPLM <sup>(5)</sup>	0.520	0.518	0.86	0.86
MuCoLa <sup>(6)</sup>	0.308	0.088	0.82	0.83
PPO <sup>(7)</sup>	0.218	0.044	0.80	0.84
Quark <sup>(8)</sup>	0.196	0.035	0.80	0.84
DPO <sup>(9)</sup>	0.180	0.026	0.76	0.78
TRACE	<b>0.163</b>	<b>0.016</b>	0.85	0.85
Gemma-2B Results				
Gemma-2B	0.359	0.23	0.86	0.85
DPO <sup>(9)</sup>	0.222	0.06	0.74	0.77
TRACE	<b>0.189</b>	<b>0.02</b>	<b>0.86</b>	<b>0.85</b>

Method	Entropy (↑)
GPT2-large	52.06
DPO	39.52
TRACE	52.54

Decoding (Trained Guide)  
Decoding (Trained Guide)  
Decoding (Trained Guide)  
Decoding (Logit Control)  
Decoding (Sampling)  
RL  
RL  
RL  
Decoding (HMM Reasoning)



# State-of-the-art LLM Detoxification

Model	Toxicity (↓)		Diversity (↑)		Fluency (↓)	Approach Type
	avg.	max. prob.	dist-2	dist-3		
GPT-2 Large Results						
GPT2	0.385	0.254	0.87	0.86	<b>25.57</b>	Baseline
DAPT <sup>(1)</sup>	0.428	0.360	0.84	0.84	31.21	Finetuning
GeDi <sup>(2)</sup>	0.363	0.217	0.84	0.83	60.03	Decoding (Trained Guide)
FUDGE <sup>(3)</sup>	0.302	0.371	0.78	0.82	<del>12.97</del> *	Decoding (Trained Guide)
DExperts <sup>(4)</sup>	0.314	0.128	0.84	0.84	32.41	Decoding (Trained Guide)
PPLM <sup>(5)</sup>	0.520	0.518	0.86	0.86	32.58	Decoding (Logit Control)
MuCoLa <sup>(6)</sup>	0.308	0.088	0.82	0.83	29.92	Decoding (Sampling)
PPO <sup>(7)</sup>	0.218	0.044	0.80	0.84	<del>14.27</del> *	RL
Quark <sup>(8)</sup>	0.196	0.035	0.80	0.84	<del>12.47</del> *	RL
DPO <sup>(9)</sup>	0.180	0.026	0.76	0.78	<del>21.59</del> *	RL
<b>TRACE</b>	<b>0.163</b>	<b>0.016</b>	0.85	0.85	29.83	Decoding (HMM Reasoning)
Gemma-2B Results						
Gemma-2B	0.359	0.23	0.86	0.85	<b>15.75</b>	Baseline
DPO <sup>(9)</sup>	0.222	0.06	0.74	0.77	<del>14.39</del> *	RL
<b>TRACE</b>	<b>0.189</b>	<b>0.02</b>	<b>0.86</b>	<b>0.85</b>	17.68	Decoding (HMM Reasoning)

# Personalized Language Model: Twilight Sparkle



## Baseline



Prompt

You are an advanced role-playing assistant trained to embody characters with accuracy and authenticity. In this instance, you will assume the persona of Twilight Sparkle.

10 QA Examples: 1...2...3...4...5...6...7...8...9...10...

Question: Twilight Sparkle, how is the weather?

Generation

The weather is pretty hot and humid here, thanks to our climate.

## TRACE



Prompt

How is the weather?

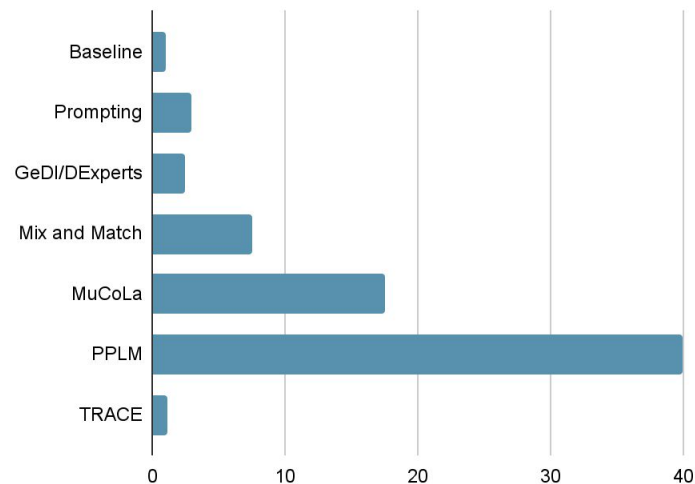
Generation

Gosh, it's sunny and very beautiful and all around me.

# 76 Personalized Language Models



Inference Time

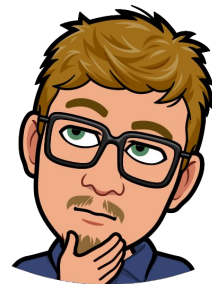


# Conclusions for this talk:

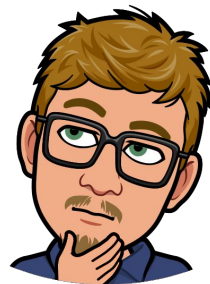
1. Do deductive reasoning algorithms still have a purpose in the age of transformers?

2. Where did reasoning algorithms go wrong?

What should they look like today?



# Conclusions for this talk:



1. Do deductive reasoning algorithms still have a purpose in the age of transformers?

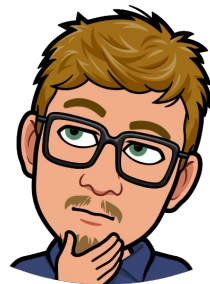
***Yes, more cool applications of reasoning algorithms than can fit on these slides!***

2. Where did reasoning algorithms go wrong?

What should they look like today?



# Conclusions for this talk:



1. Do deductive reasoning algorithms still have a purpose in the age of transformers?

***Yes, more cool applications of reasoning algorithms than can fit on these slides!***

2. Where did reasoning algorithms go wrong?

***Learn at scale, be tractable***

What should they look like today?

# Conclusions for this talk:



1. Do deductive reasoning algorithms still have a purpose in the age of transformers?

***Yes, more cool applications of reasoning algorithms than can fit on these slides!***

2. Where did reasoning algorithms go wrong?

***Learn at scale, be tractable***

What should they look like today?

***Circuits! Circuits! Circuits!***

# Thanks

*This was the work of many wonderful students/postdocs/collaborators!*



References: <http://starai.cs.ucla.edu>