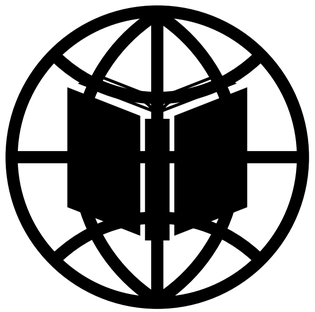
Amortizing Intractable Inference in Large Language Models

Edward J. Hu*, Moksh Jain*, Eric Elmoznino, Younesse Kaddar,
Guillaume Lajoie, Yoshua Bengio, Nikolay Malkin

International Conference on Learning Representations, 2024.

$$p_{\mathrm{LM}}(w_1 w_2 \ldots w_n) = p(w_1)p(w_2 \mid w_1)\ldots p(w_n \mid w_1, \ldots, w_{n-1})$$

Language models store knowledge about the world



$$\max \log p_\theta(w_{k+1} \mid w_1, \ldots, w_k)$$

Language Model

How to perform inference with this knowledge?

Prompting
Can be brittle

MCMC
Slow Mixing

RL
Suffers from mode collapse

?

Intractable Inference in Language Models

- Tempered and contrastive sampling

$$q(Z \mid X) \propto p_{\mathrm{LM}}(Z \mid X)^{\frac{1}{T}}$$

- Constrained sampling

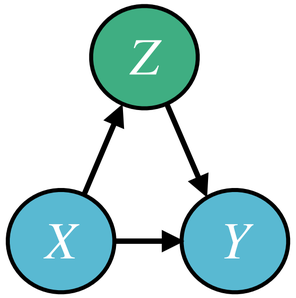
$$q(Z) \propto p_{\mathrm{LM}}(Z)c(Z)$$

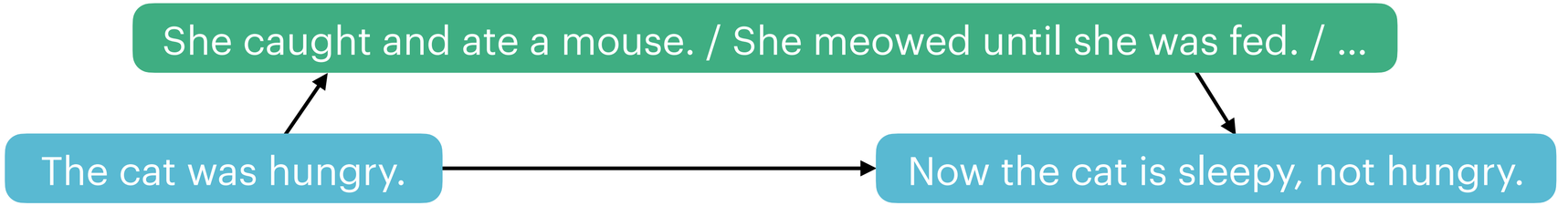
- Infilling and reverse sampling

$$q(Z \mid X, Y) \propto p_{\mathrm{LM}}(XZY)$$

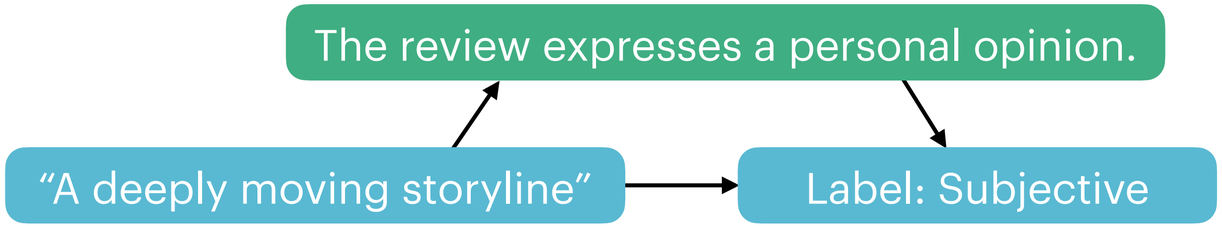
Reasoning Through Latent Variables

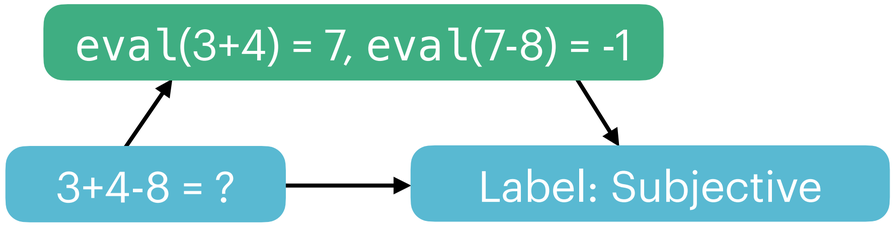
Infilling is a latent variable inference problem





Reasoning and tool use are instantiations of this problem!





$$p_{\mathrm{LM}}(Z \mid X, Y) = \frac{p_{\mathrm{LM}}(Z \mid X)\, p_{\mathrm{LM}}(Y \mid X, Z)}{\sum_{Z'} p_{\mathrm{LM}}(Z' \mid X)\, p_{\mathrm{LM}}(Y \mid X, Z')}$$

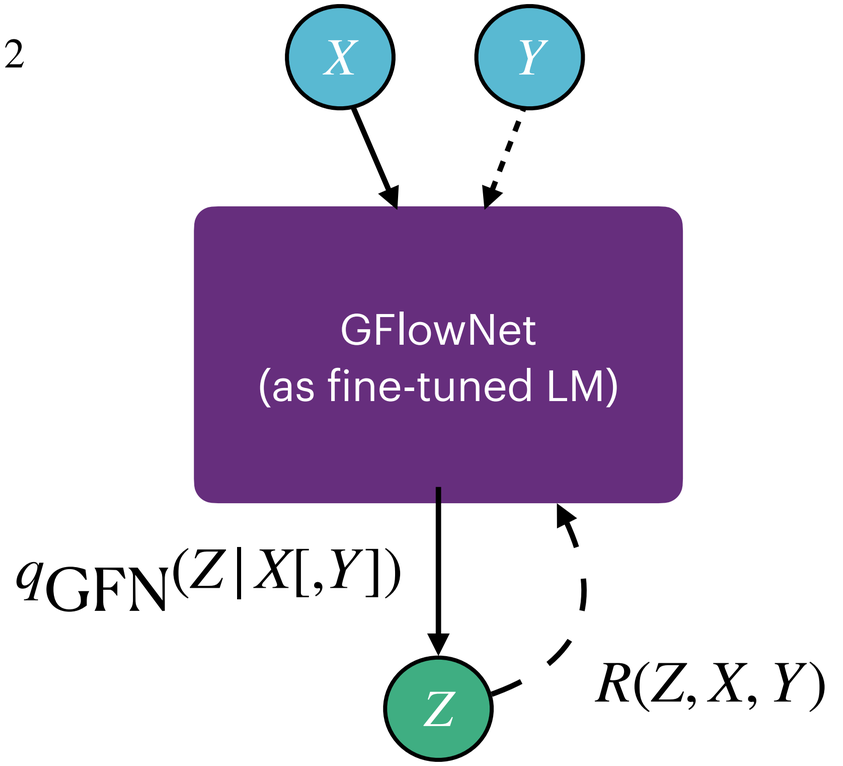
$$\propto p_{\mathrm{LM}}(X, Z, Y)$$

Amortized Inference with GFlowNets

Finetune the LLM as a GFlowNet policy to sample from $p(Z \mid X, Y)$

$$L(Z;\theta) = \sum_{0 \le i < j \le n} \left(\log \frac{R(z_{1:i}\top) \prod_{k=i+1}^{j} q_{\mathrm{GFN}}(z_k \mid z_{1:k-1}) q_{\mathrm{GFN}}(\top \mid z_{1:j})}{R(z_{1:j}\top) q_{\mathrm{GFN}}(\top \mid z_{1:i})} \right)^2$$

Equivalent to path consistency objective in MaxEnt RL!

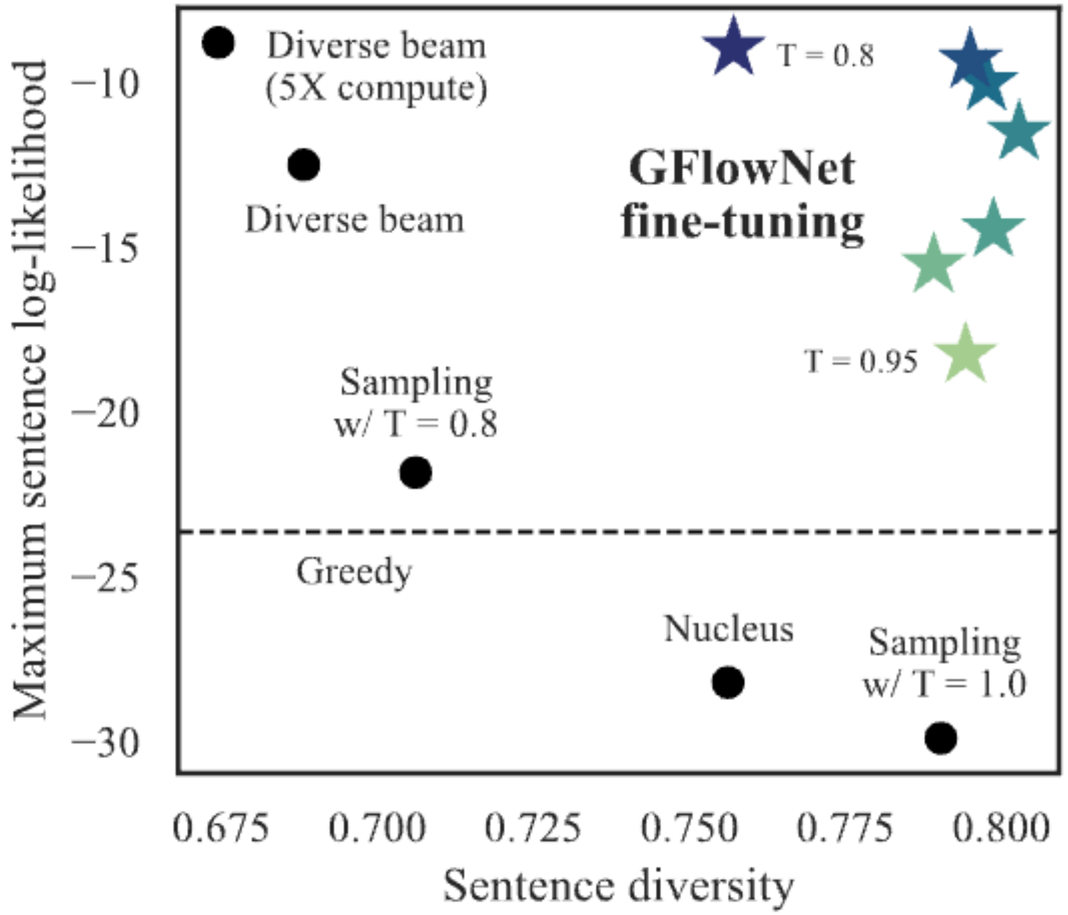


- Learned policy q_{GFN} can be used to sample Z for a new X
- **Posterior Predictive**: Sample many chains and take the most likely Y using $p_{LM}(Y \mid X, Z)$
- **Variational EM**: Can also update $p_{\mathrm{LM}}(Y \mid X, Z)$

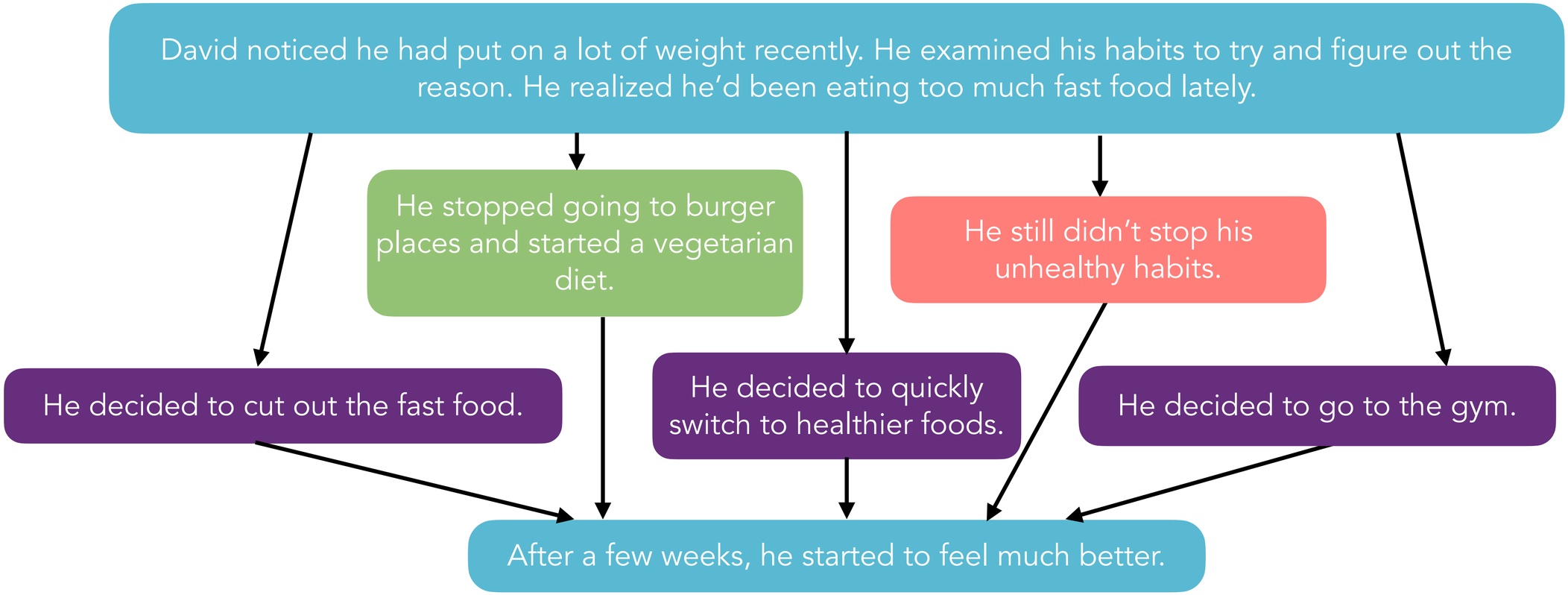
Sentence Continuation

Sampling from tempered distribution $q(Z|X) \propto p_{\mathbf{LM}}(Z|X)^{\frac{1}{T}}$

GFlowNet fine-tuning balances **likelihood** and **diversity**!



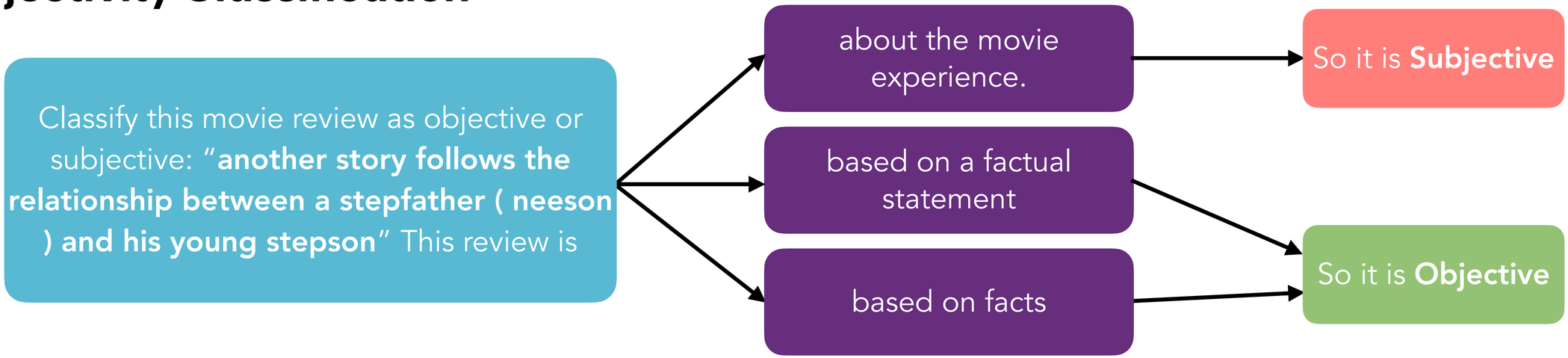
Infilling Stories



	BERTScore	GLEU-4	LLM Judge
Prompting	0.081	3.2	2.4
SFT	0.094	3.7	2.7
GFlowNet-FT	**0.184**	**4.2**	**3.4**

Chain of thought reasoning

Subjectivity Classification

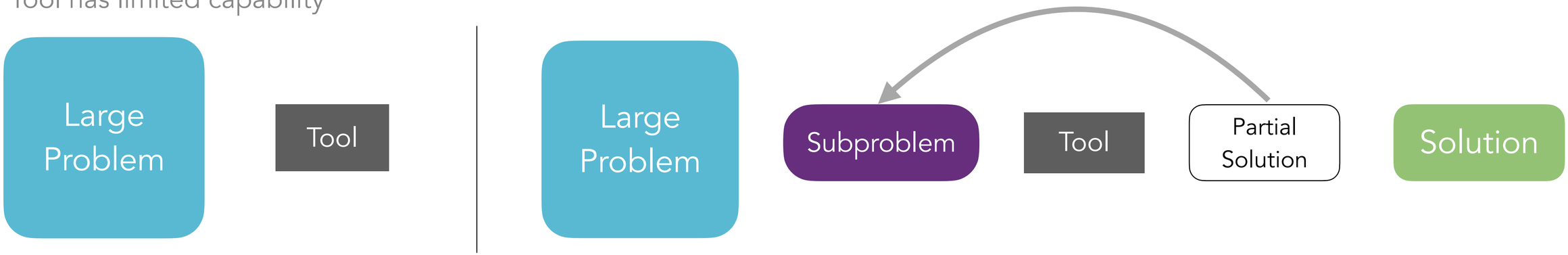


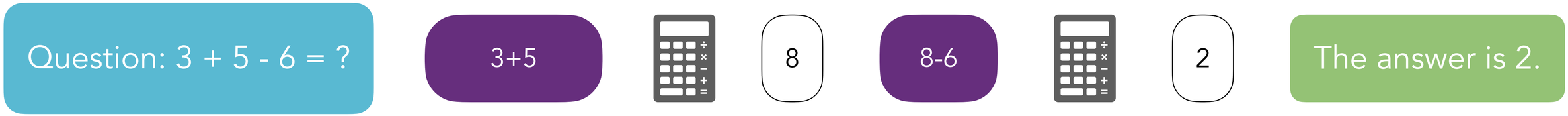
	Training Samples		
	10	**20**	**50**
0-shot	51.7		
ICL	61.3	61.8	65.8
SFT	64.3	69.1	**89.7**
GFlowNet FT	71.4	**81.1**	87.7
+ EM	**75.2**	78.7	**89.9**

Tool Use

Arithemetic with a calculator

Tool has limited capability





Arithmetic with Tool Use

	Simple In-distribution	Hard In-distribution	OOD
CoT Prompting	35.5	21.0	10.5
SFT	72.1	19.6	12.8
PPO	30.6	13.7	5.6
GFlowNet FT	**95.2**	**75.4**	**40.7**

x

$z \sim q(z \mid x)$

$\log R$

Question: 1 - 9 - 8 = ? Answer:

1 - 9 - 8

-13.17

1 - 9 = **-8**, -8 - 8 = **-16**

-27.75

Limitations

- **Assumption**: Base language model is a “good” world model
  - LMs can suffer from hallucination / miscalibration
- Slower than supervised fine-tuning
- Exploration is still a challenge
- Task specific models

What next?

- Transfer and generalization across tasks
- Quantifying epistemic uncertainty through chains-of-thought
- More structured latents (e.g. tree-of-thought)
- Learning from preference data - capturing distribution over preferences!
- Amortized inference in diffusion models

Find us at Poster #93 Halle B!

Paper

Code