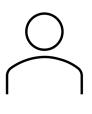
# Tractable Control for Autoregressive Langua

Honghua Zhang\* Meihua Dang\* University of California, Los Angeles

Nanyun Violet Peng

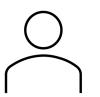
### Prompting is Not All You Need



Generate a sentence using "frisbee", "caught" and "dog", following the given order.



After a perfect throw, the <u>frisbee</u> glided through the air, and the <u>dog</u>, with incredible agility, <u>caught</u> it mid-flight.



That's not correct. Generate a sentence using "frisbee", "caught" and "dog". The keywords should appear in the order as specified.



Here's the correct sentence: The <u>dog caught</u> the <u>frisbee</u> in mid-air. showing off its amazing catching skills in mid-air, showing off its amazing catching skills.

ChatGPT fails to follow simple logical constraints!

## **Constrained Generation is Challenging**

**Logical Constraint** α: e.g., text contains keyword "winter"

**Constrained Autoregressive Generation.** Our goal is to generate from

$$p(x_{1:n} | \alpha) = \prod_{t} p(x_{t+1} | x_t, \alpha)$$

at each step, e.g., suppose we have generated the first t tokens  $x_{1:t}$  as "the weather is"; then we generate the next token  $x_{t+1}$  from

$p(x_{t+1}   x_{1:t}, \alpha) \propto p(x_{t+1}   x_{1:t}) \cdot p(\alpha   x_{t+1}, x_{1:t})$					
$x_{t+1}$	$p_{lm}(x_{t+1}   x_{1:t})$		$x_{t+1}$	$p_{lm}(\alpha   x_{t+1}, x_{1:t})$	
cold	0.05		cold	?	
warm	0.10		warm	?	
Off-the-shelf LM $\checkmark$ Steers LM to satisfy $\alpha$ . $\checkmark$ distribution. Intractable for LLMs.					
Requires marginalization over all					

|suffixes  $x_{t+1:n}$  containing "winter".

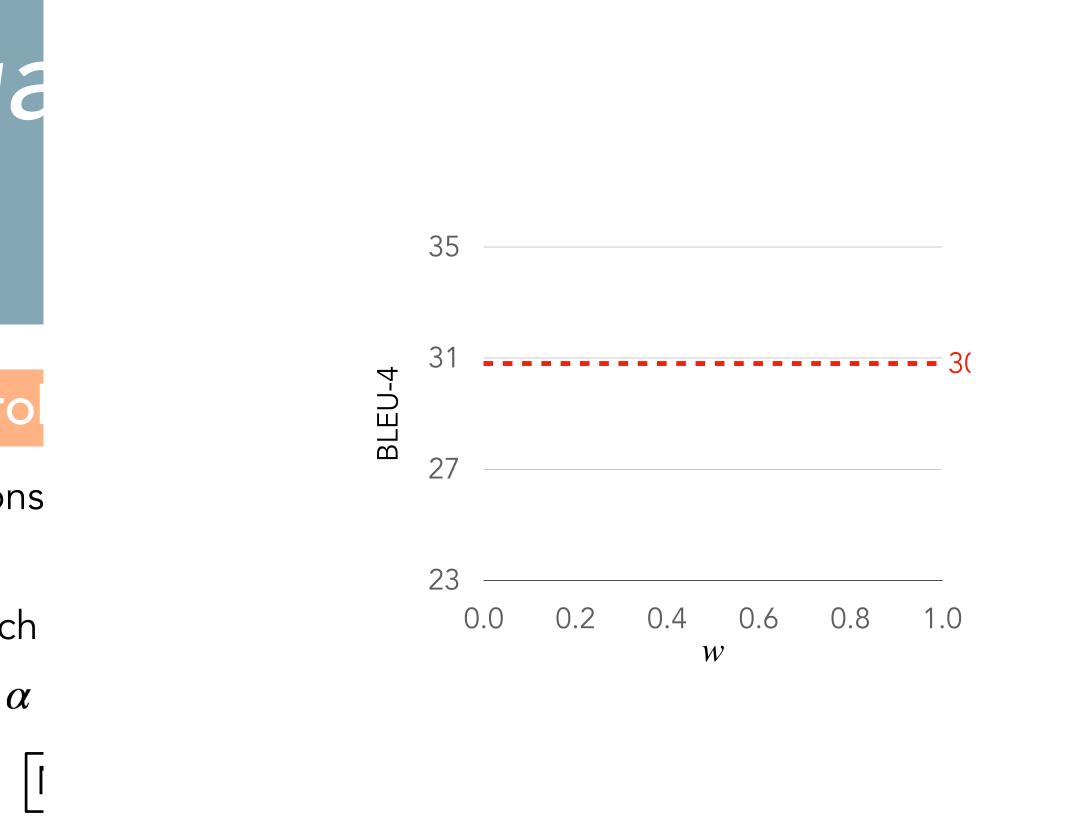
**References:** [1] Lu, S., Meng, T., and Peng, N. Insnet: An efficient, flexible, and performant insertion-based text generation model. In NeurIPS, 2022. [2] Lu, X., West, P., Zellers, R., Le Bras, R., Bhagavatula, C., and Choi, Y. Neurologic decoding: (un) supervised neural text generation with predicate logic constraints. In Proceedings of NAACL, 2021. [3] Lu, X., Welleck, S., West, P., Jiang, L., Kasai, J., Khashabi, D., Le Bras, R., Qin, L., Yu, Y., Zellers, R., Smith, N. A., and Choi, Y. NeuroLogic a\*esque decoding: Constrained text generation with lookahead heuristics. In Proceedings of NAACL, 2022. [4] Meng, T., Lu, S., Peng, N., and Chang, K.-W. Controllable text generation with neurally-decomposed oracle. In NeurIPS, 2022.

Guy Van den Broeck

GeLaTo	Pro
We propose <b>GeLaTo</b> ( <b>Ge</b> nerating <b>La</b> nguage with <b>T</b> ractable Constraints) to guide autoregressive generation from LLMs.	Con
Tractable Probabilistic Models (TPMs) are generative models $p_{tpm}(x_{1:n})$ that allow efficient conditioning. We use hidden Markov models (HMMs) as an example.	each a
Step 1. Distilling an HMM from LM	
- Train $p_{hmm}$ on $D \sim p_{lm}$ to minimize their KL-divergence.	Effic
Step 2. Probabilistic Reasoning with Constraints	HM
- Compute $p_{hmm}(\alpha x_{1:t}, x_{t+1})$ to approximate $p_{lm}(\alpha x_{1:t}, x_{t+1})$ ; then generate from:	Assı
$p_{gelato}(x_{t+1}   x_{1:t}, \alpha) \propto p_{lm}(x_{t+1}   x_{1:t}) \cdot p_{hmm}(\alpha   x_{t+1}, x_{1:t})$	com
*GeLaTo can also help prompting!	$p(\alpha_t)$
Given some prompt $\pi$ that represents $\alpha$ , e.g., "keywords = XXX", we can combine $p_{hmm}$ and $p_{lm}$ by taking their weighted geometric mean:	
$p_{gelato}(x_{t+1}   x_{1:t}, \pi, \alpha) \propto p_{lm}(x_{t+1}   x_{1:t}, \pi)^{1-w} \cdot p_{hmm}(x_{t+1}   x_{1:t}, \alpha)^{w}$	
*GeLaTo can enforce various logical constraints	Exp
<ol> <li>Keywords appear (in any order/form of inflections)</li> <li>(Some) keywords are generated following a specific order.</li> <li>(Some) keywords must appear at specified positions.</li> <li>(Some) keywords must not appear in the generated text.</li> </ol>	Con Inpu Outp Outp
Advantages of GeLaTo	N
1. Logical constraint $\alpha$ is <b>guaranteed</b> to be satisfied. - When generating next token $x_{t+1}$ , if $x_{t+1} = w$ would	Un. InsN Neu

w would not be generated. 2. The training of  $p_{hmm}$  does not depend on  $\alpha$ , which is only imposed during generation. Once  $p_{hmm}$  is trained, GeLaTo generalizes to any tractable constraints.

make  $\alpha$  unsatisfiable, then  $p_{gelato}(x_{t+1}|\alpha, x_{1:n}) = 0$ ; hence



icient Probabilistic Reasoning for HMMs

*Ms* define distributions over  $x_{1:n}$  and latent variables  $z_{1:n}$ :

$$p(x_{1:n}, z_{1:n}) = \prod_{t} p(z_{t+1} | z_t) p(x_t | z_t)$$

sume  $\alpha$  only contains single-token keywords, then we can mpute  $p(\alpha_{t:n})$ , as well as  $p(\alpha_{1:n}, x_{1:t})$ , by

$$\sum_{t:n} |z_t| = \sum_{w \in vocab} \sum_{z_{t+1}} p((\alpha \setminus w)_{t+1:n} |z_{t+1}) p(z_{t+1} | z_t) p(x_t = w | z_t)$$

The time complexity for sampling from  $p_{gelato}(x_{1:n}|\alpha)$  is  $O(2^m n)$ 

#### periments

mmonsense Generation (CommonGen)

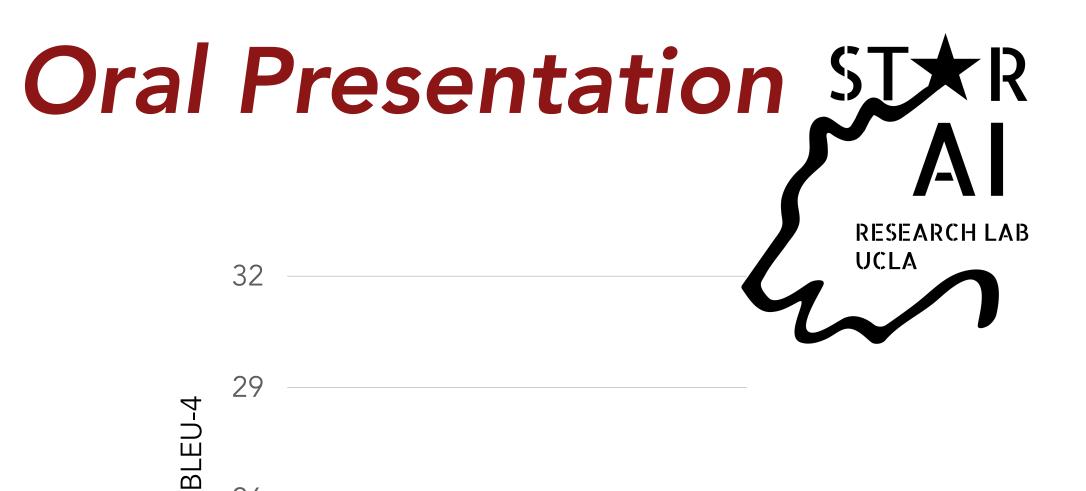
out Concepts: snow, car, drive tput 1: The <u>car drives</u> down a <u>snow</u>-covered road. tput 2: <u>Driving</u> through the <u>snow</u>, the <u>car</u> crashed.

Method	Generation Quality					Constraint		
Ivietiiou	ROUGE-L		BLEU-4		CIDEr		Success Rate	
Unsupervised	dev	test	dev	test	dev	test	dev	test
InsNet	-	-	18.7	-	-	-	100.0	-
NeuroLogic	-	41.9	-	24.7	_	14.4	-	-
A*esque	-	44.3	-	28.6	_	15.6	-	-
NADO	-	-	26.2	-	-	-	-	-
GeLaTo	44.3	43.8	30.3	29.0	15.6	15.5	100.0	100.0
Supervised	dev	test	dev	test	dev	test	dev	test
NeuroLogic	-	42.8	-	26.7	-	14.7	-	93.9 <sup>†</sup>
A*esque	-	43.6	-	28.2	-	15.2	-	$97.9^{\dagger}$
NADO	$44.4^{\dagger}$	_	30.8	_	16.1 <sup>†</sup>	-	$88.8^{\dagger}$	-
GeLaTo	46.2	45.9	34.0	34.1	17.2	17.5	100.0	100.0

Table 1. Automatic evaluation results on CommonGen.

**ICML 2023** 

**Acknowledgements:** This work was funded in part by the DARPA Perceptually-enabled Task Guidance (PTG) Program under contract number HR00112220005, NSF grants #IIS-1943641, #IIS-1956441, #CCF-1837129, an SRA from Meta, a research gift from Amazon Alexa AI, and a gift from RelationalAI. GVdB discloses a financial interest in RelationalAI.



Method	Concepts	Plausibility	Quality	Overall
GPT2	16 2.47	32 <b>2.52</b>	2.65	2.28
NADO	2.71	<b>2.54</b> 64	2.73 <sup>128</sup>	2.54
GeLaTo	2.73	Bear <b>2</b> , <b>52</b>	2.70	<b>2.60</b>

# of concepts	3	4	5
Unsupervised			
A*esque	472.9	542.5	613.9
GeLaTo (16)	$13.5 \pm 4.4$	$21.9\pm5.37$	$39.3 \pm 6.3$
GeLaTo (128)	$69.8 \pm 32.3$	$97.9\pm39.5$	$143.0 \pm 44.4$
Supervised			
A*esque	8.5	9.6	11.4
GPT2 (16)	$5.8 \pm 1.1$	$13.0 \pm 1.6$	$29.3\pm3.2$
GPT2 (128)	$9.4 \pm 1.8$	$21.1 \pm 11.9$	$33.7\pm3.5$
GeLaTo (16)	$11.1 \pm 2.8$	$22.0\pm5.0$	$41.6 \pm 5.6$
GeLaTo (128)	$49.8 \pm 20.8$	$88.7\pm30.5$	$127.6\pm30.4$

Table 3. Time (seconds) for generating one sentence on CommonGen.

#### Yelp!Review and News: fixing order of keywords

The Yelp!Review and News datasets are similar to CommonGen except for that they require keywords to appear in the given order.

Method \Dataset	Yelp!Review	News
InsNet NADO	5.8	5.0 4.5
GeLaTo	<b>6.6</b>	<b>5.4</b>

Table 4. BLEU-4 scores for Yelp!Review and News datasets

