Table 1: CommonGen results. All methods are applied to the GPT2-large models;

	BLEU-4		ROUGE-L		CIDEr		SPICE		Constraint	
	dev	*test*	*dev*	*test*	*dev*	*test*	*dev*	*test*	*dev*	*test*
supervised - base models trained with full supervision										
FUDGE	-	24.6	-	40.4	-	-	-	-	-	47.0%
A*esque	-	28.2	-	43.4	-	15.2	-	30.8	-	98.8%
NADO	30.8	-	44.4	-	16.1	-	32.0	-	88.8%	-
GeLaTo	34.0	34.1	46.2	45.9	17.2	17.5	32.2	**33.5**	**100.0%**	**100.0%**
Ctrl-G	**35.1**	**34.4**	**46.7**	**46.4**	**17.4**	**17.6**	**32.7**	33.3	**100.0%**	**100.0%**
unsupervised - base models not trained with keywords as supervision										
A*esque	-	28.6	-	44.3	-	15.6	-	29.6	-	-
NADO	26.2	-	-	-	-	-	-	-	-	-
GeLaTo	30.3	29.0	44.3	43.8	15.6	15.5	30.2	30.3	**100.0%**	**100.0%**
Ctrl-G	**32.1**	**31.5**	**45.2**	**44.8**	**16.0**	**16.2**	**30.8**	**31.2**	**100.0%**	**100.0%**

Table 2: Time (seconds) of generating one example on CommonGen (dev); # of HMM hidden states shown in parentheses. Beam size for GeLaTo and Ctrl-G is 128.

	unsupervised			*supervised*		
# of concepts	3	4	5	3	4	5
A*esque	472.9	542.5	613.9	8.5	9.6	11.4
GeLaTo (4096)	69.8 ± 32.3	97.9 ± 39.5	143.0 ± 44.4	49.8 ± 20.8	88.7 ± 30.5	127.6 ± 30.4
Ctrl-G (4096)	1.1 ± 0.3	1.9 ± 0.5	4.6 ± 1.4	1.2 ± 0.4	2.3 ± 0.8	5.7 ± 1.7
Ctrl-G (32768)	4.1 ± 0.9	9.0 ± 2.0	22.3 ± 5.4	4.7 ± 1.6	11.0 ± 3.8	27.6 ± 8.3

Table 3: Evaluation results of interactive text editing. *K&W* indicates that the model should adhere to both keyphrase (*K*) and word count range (*W*) constraints. Table shows the human evaluation score (*Quality*), constraint success rate (*Success*), and overall satisfaction rate (*Overall*), which is the proportion of examples satisfying constraints with Quality scores above 3.

	Continuation					*Insertion*				
	None	*K*	*W*	*K&W*	*Avg.*	*None*	*K*	*W*	*K&W*	*Avg.*
Quality										
TULU2	3.80	3.77	3.87	3.88	3.83	2.68	2.64	2.78	2.74	2.71
GPT3.5	4.40	4.32	**4.44**	**4.36**	4.38	2.27	2.22	2.27	2.31	2.27
GPT4	**4.48**	**4.44**	**4.44**	4.26	**4.40**	**3.79**	3.33	3.53	3.10	3.44
Ctrl-G	4.13	3.98	4.27	3.96	4.08	**3.77**	**3.56**	**3.73**	**3.59**	**3.67**
Success										
TULU2	-	35%	33%	1%	23%	-	12%	20%	3%	12%
GPT3.5	-	36%	62%	31%	43%	-	22%	54%	10%	29%
GPT4	-	56%	55%	59%	57%	-	60%	20%	27%	36%
Ctrl-G	-	**100%**	**100%**	**100%**	**100%**	-	**100%**	**100%**	**100%**	**100%**
Overall										
TULU2	-	30%	31%	1%	21%	-	7%	10%	1%	6%
GPT3.5	-	36%	62%	31%	43%	-	0%	5%	2%	2%
GPT4	-	56%	55%	57%	56%	-	41%	17%	14%	24%
Ctrl-G	-	**89%**	**97%**	**90%**	**92%**	-	**76%**	**78%**	**82%**	**79%**

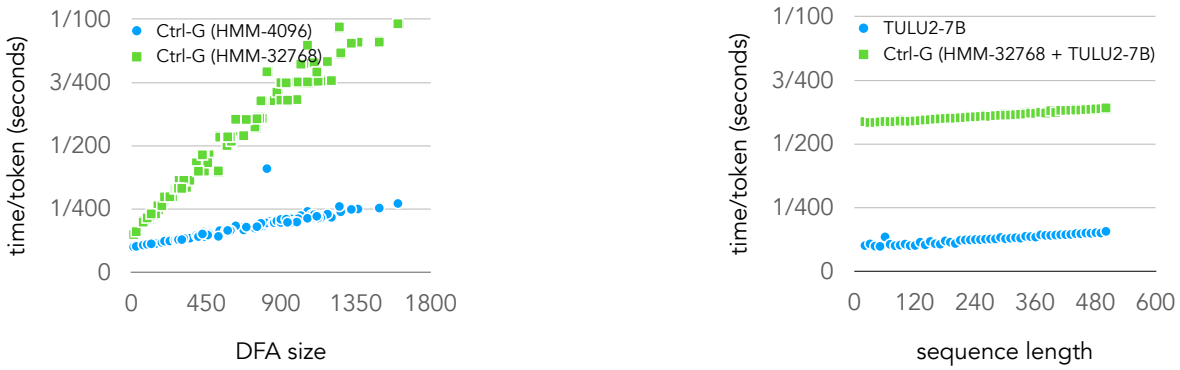

Figure 1: Runtime analysis of Ctrl-G on the task of Interactive Text Editing with TULU2-7B as the base model; Left: the generation time per token scales linearly w/ respect to DFA size. Right: the overhead of Ctrl-G per token (diff. between two lines) stays constant w/ respect to sequence length.