How often do LLMs Brush🖌️ My Instruction Off using Ctrl+C✂️ & Ctrl+V🩹 and How to Stop 💪them?

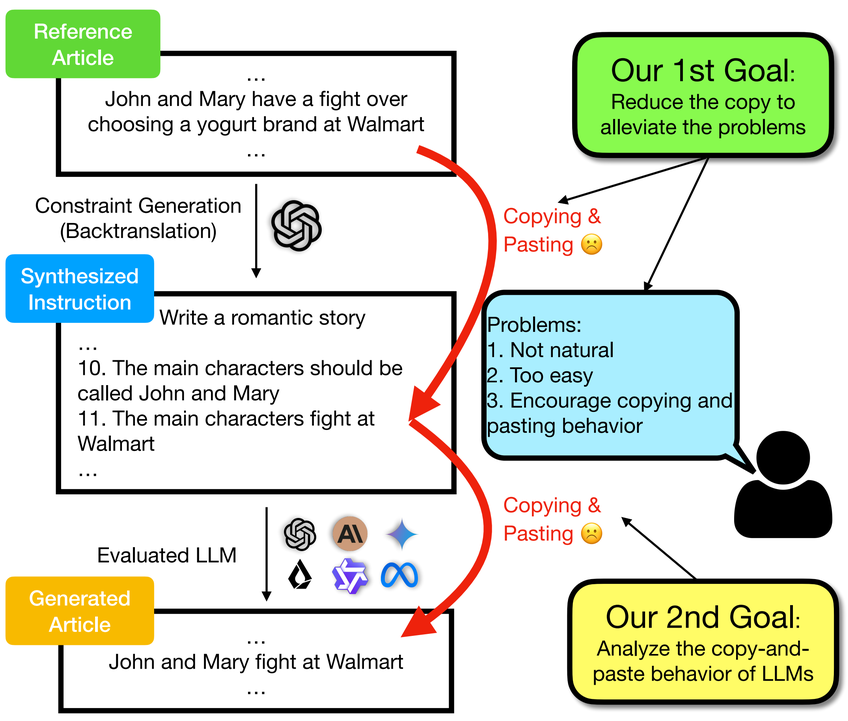
UNSPECIFIC: General Constraint Synthesis for Breaking Copy-and-Paste Shortcut in LLM Instruction Following

College of Information Science

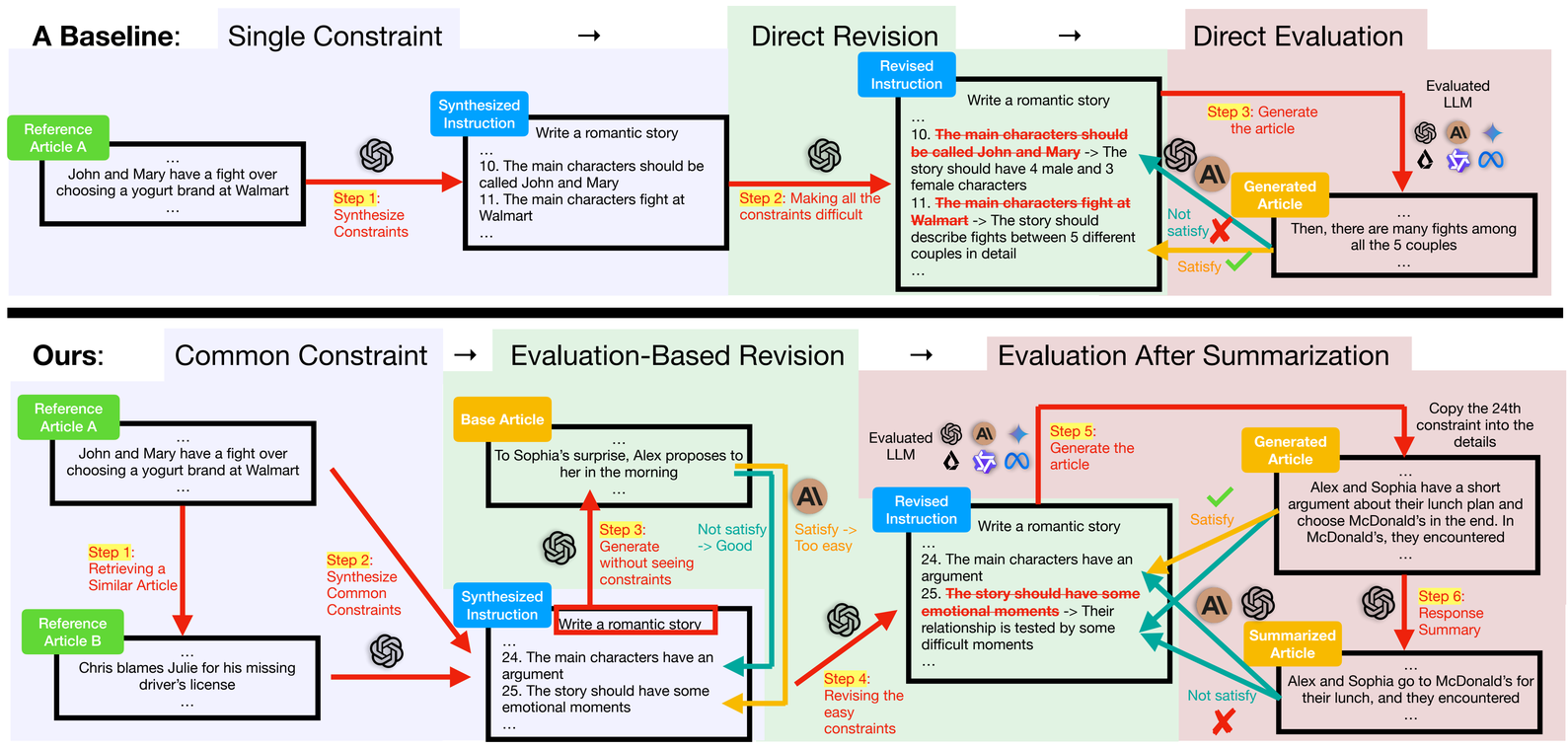
Jeet Sharma*, Balpreet Kaur*, Jeremiah Hong,
Hamed Zamani, Haw-Shiuan Chang*

UMASS AMHERST

Introduction



Method



UNSPECIFIC (UNcovering Summary-resistant, Paste-reducing, and Evaluation-revised Constraints to Improve Following Instruction Challenge) Benchmark and Evaluation

		LLM Satisfaction Judge				LLM Coherence Judge				Human Naturalness Judge		
		GPT-5 Mini		Llama-3 8B		GPT-5 Mini		Llama-3 8B				
Synth	Revise	Org	Core	Org	Core	Org	Core	Org	Core	Human Win	Tie	LLM Win
Single	None	89.7	48.4	50.7	15.2	3.36	2.80	2.32	2.17	59	5	36
Common	None	95.1	65.7	57.9	27.7	3.52	2.92	2.44	2.36	40	14	46
Common	Direct	71.8	41.0	27.6	8.2	3.22	2.72	2.36	2.16	48	3	49
Common	Eval	78.2	46.5	38.1	10.6	3.28	2.88	2.28	2.12	41	11	48

Table 1: Comparison of different constraint synthesis methods in news domain. Single+None is the back-translation baseline and Common+Eval is our method. The numbers except for coherence are percentages.

1-1. Common+Eval is more challenging and natural compared to Single+None baseline.
1-2. Evaluation-based revision increases the difficulty without degrading the naturalness significantly

Single Constraints	Common Constaints
1. Begin by describing high schoolers walking out of class across Alaska 2. Include details about students chanting slogans and marching to the state Capitol 3. Conclude by emphasizing student activism as democracy in action	1. Begin with a scene that involves students participating in a protest 2. Include detailed accounts of the middle portion of events 3. End with a statement reflecting on the broader significance of student activism

Table 2: Example comparison between the constraints from a single reference article and the common constraints from two similar reference articles

2. Common constraints mitigate superficial constraint satisfaction

Task	Synth	Revise	GPT-5 Mini Org	GPT-5 Mini Core	Llama-3 8B Org	Llama-3 8B Core
Gen	Single	None	89.7	48.4	50.7	15.2
		Direct	70.2	31.1	33.3	7.5
		Eval	77.5	35.5	38.2	9.4
	General Constraint Prompting	None	90.4	47.6	53.3	31.6
		Direct	70.1	28.1	31.7	13.1
		Eval	79.8	28.1	32.5	12.9
	Common	None	95.1	65.7	57.9	27.7
		Direct	71.8	41.0	27.6	8.2
		Eval	78.2	46.5	38.1	10.6
Edit	Single	None	86.6	41.9	30.1	8.2
		Direct	70.5	27.3	14.7	3.2
		Eval	77.1	23.7	25.1	5.9
	Common	None	93.3	62.7	52.1	23.7
		Direct	72.8	36.6	21.0	5.2
		Eval	77.2	44.6	34.8	6.5

Table 3: The satisfaction rate in the news domain. The Org is the satisfaction rate for the original response and Core is the satisfaction rate for the summary of the response with 25% length of the response. The maximum standard error across all cells is 4.78.

Task	Synth	Revise	GPT-5 Mini Org	GPT-5 Mini Core	Llama-3 8B Org	Llama-3 8B Core
Gen	Single	None	0.589	0.501	0.602	0.508
		Direct	0.597	0.528	0.638	0.523
		Eval	0.598	0.509	0.617	0.512
	General Constraint Prompting	None	0.581	0.500	0.599	0.527
		Direct	0.594	0.544	0.640	0.558
		Eval	0.579	0.510	0.615	0.542
	Common	None	0.546	0.459	0.558	0.482
		Direct	0.579	0.498	0.614	0.512
		Eval	0.561	0.485	0.608	0.509
Edit	Single	None	0.586	0.501	0.565	0.485
		Direct	0.604	0.525	0.595	0.509
		Eval	0.596	0.518	0.603	0.511
	Common	None	0.541	0.449	0.534	0.450
		Direct	0.577	0.497	0.607	0.497
		Eval	0.570	0.490	0.590	0.479

Table 5: The average similarity between each constraint to the closest sentence in the response. Maximum standard error across all cells is 0.024.

3-1. Evaluation after summarization is a more reliable metric than the similarity measurement.

3-2. Editing an article is more likely to induce the superficial constraint satisfaction from a strong LLM than generating the article from scratch

	News				Story				Blog			
	Single		Common		Single		Common		Single		Common	
	Org	Core	Org	Core	Org	Core	Org	Core	Org	Core	Org	Core
Claude Sonnet 4.5	89.6	15.9	85.2	22.7	89.6	25.9	91.7	44.4	82.9	39.2	87.2	50.9
Claude Haiku 4.5	74.7	16.0	71.8	19.1	83.8	26.9	85.9	41.1	74.8	36.8	78.0	45.9
GPT-5 Mini	79.0	30.6	78.5	36.1	94.7	57.4	95.1	69.9	85.9	45.6	84.3	55.1
GPT-5 Nano	68.5	15.7	69.5	17.8	81.4	31.4	86.2	47.3	74.6	36.0	75.5	49.1
LFM2 24B A2B	58.1	12.6	61.5	14.4	70.7	24.9	79.0	42.5	56.8	30.8	69.9	41.5
Gemma-3n E4B	59.3	12.1	56.5	15.5	76.0	26.4	77.8	40.3	57.5	28.1	56.2	35.4
Qwen2.5 7B Turbo	54.6	13.1	51.3	15.4	66.0	19.4	58.5	25.4	55.0	30.4	58.1	38.1
Llama-3.3 70B Lite	60.5	13.7	51.4	13.7	69.5	25.1	64.5	30.0	51.9	30.1	49.0	32.0
Llama-3 8B Lite	34.2	8.3	33.9	6.8	45.2	19.1	50.5	26.2	34.9	22.4	42.4	27.4

Table 6: The average satisfaction rate from Haiku 4.5 + GPT-5 Nano in UNSPECIFIC. The numbers are percentages. Maximum standard error across all cells is 3.0.

6-1. LLM tends to satisfy the constraints in news more superficially than in blogs.
6-2. The improvement of some larger models mostly comes from superficial instruction following.

Conclusions

LLMs often satisfy constraints by copy-pasting them into details (only 28% survive summarization in news). UNSPECIFIC (common constraints + evaluation-based revision + evaluation after summarization) makes instructions harder, more natural, and less copyable, and shows that copying depends on domain, task, model family and size.

Future Work and Other Applications

1. Doing RL to reduce the superficial instruction following
2. Common summary for coarse-to-fine interactive text generation [1]
3. Common summary for ultra-fine-grained taxonomy construction [2]

[1] Haw-Shiuan Chang, Nanyun Peng, Mohit Bansal, and Tagyoung Chung. Fine-to-Coarse Entailment Hierarchy Construction for Coarse-to-Fine Story Generation. In Bridging Human-Computer Interaction and Natural Language Processing (HCI+NLP), 2024

[2] Sivaraaman Balakrishnan*, Prakriti Shankar Shetty*, Jeremiah Hong, Hamed Zamani, and Haw-Shiuan Chang*, COMPREHEND: Ultra-Fine-Grained Benchmark Understanding by Entailment Hierarchy. In NeurIPS 2026 Trust-AI-Eval Workshop