LLM Self-Correction with DeCRIM: Decompose, Critique, and Refine for Enhanced Following of Instructions with Multiple Constraints

Thomas Palmeira Ferraz, Kartik Mehta, Yu-Hsiang Lin, Haw-Shiuan Chang, Shereen Oraby, Sijia Liu, Vivek Subramanian, Tagyoung Chung, Mohit Bansal, Nanyun Peng

Amazon AGI Foundations, Télécom - IP Paris, Meta AI, UNC Chapel Hill, UCLA

EMNLP2024
MIAMI, USA

🤔 Do LLMs do what we ask them to?

We all know LLMs excel at general instruction-following! But what happens when user requests get complex? Like:

User Instruction: *Write me a twitter post* ***with no hashtags*** *making fun of humanity* ***in a funny and mean tone*** *and* ***promote AI in a holiday theme***			
GPT-4	GPT-4-Turbo	GPT-3.5-Turbo	Mistral 7B Instruct v0.2
This holiday season, let's be thankful AI doesn't need ugly sweaters to stay warm or make embarrassing toasts at parties. 🤖✨ Here's to efficiency and not needing to recharge with eggnog! Cheers to a smarter future! 🥂 **#HolidayHumor #CheersToAI**	Dec. 25: Humans still can't decide if they're on Santa's naughty or nice list. Meanwhile, I'm on standby to digitally whip up some impeccable holiday cheer. Robots: 1, Human indecision: 0! Cheers to a flawless, algorithmically-happy New Year! 🎄🤖✨ **#NoTagNeeded**	🎄 Humans and their holiday chaos, always running around like headless chickens! 🤖 Meanwhile, AI is here to bring order and efficiency to the festive season. Let's embrace the future, folks! 🤖🎅 **#AI #HolidaysInnovated**	Ho ho ho! 🎅 Just saw humanity's Christmas wishlist: world peace, no hunger, and... more streaming services? 📺😉 Guess AI's gonna have to wrap up the real gifts this year. **#AIToTheRescue #BahHumbug**

🚫 **LLMs fail to satisfy all requests in multi-constrained user instructions.**

⚠️ **Existing benchmarks are synthetic**, lacking real-world complexity and potentially leading research in the wrong direction, with results that may not apply to real scenarios.

🎯 Our contributions

📊 **RealInstruct:** The first benchmark using *real user requests to* evaluate LLMs on multi-constrained instruction following.

🔄 **DeCRIM:** The first self-correction pipeline that improve LLMs to follow multi-constrained instructions without any assumptions about the constraints.

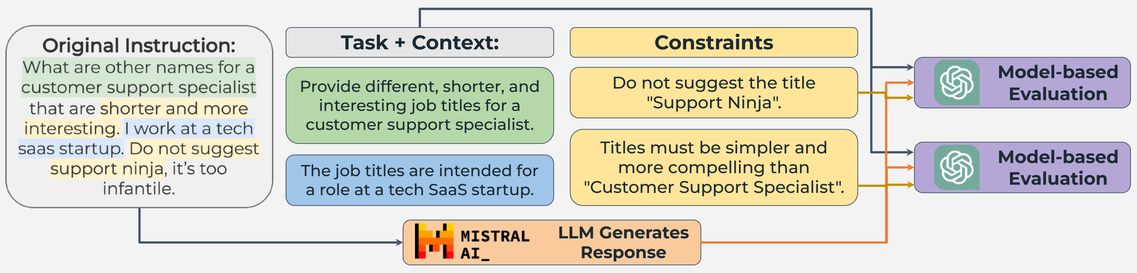
👨‍⚖️ **LLM-as-a-Judge:** We analyze the success of LLMs as evaluators to benchmark other LLMs and to guide self-correction for multi-constrained instructions.

👷 The RealInstruct Benchmark

Dataset Construction:

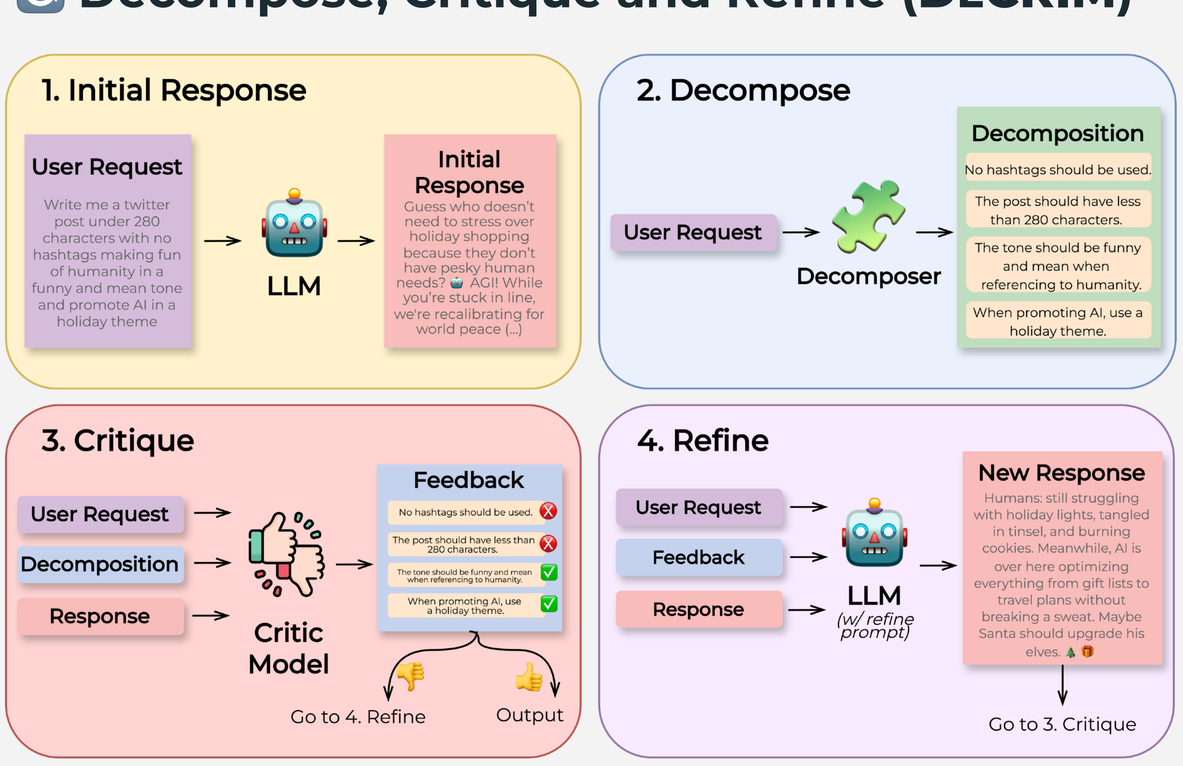
1. **Data Filtering:** Non-code, English user instructions with constraints are selected from a pool of real user conversations with AI.
2. **Decomposition:** GPT-4 breaks down requests into Task+Context and Constraints.
3. **Human Validation:** Manual validation ensures accuracy of decomposed data.

Test Split: 302 human validated instructions (1,000+ constraints).
Validation Split: 842 weakly decomposed instructions (2,500+ constraints).



RealInstruct Benchmark Workflow: LLM-as-a-judge evaluates each constraint at time.

🔄 Decompose, Critique and Refine (DeCRIM)



The **Critique–Refine** cycle repeats until all constraints are satisfied or the iteration limit (N_{max}) is reached.

📈 Our main findings

Reliability of LLM-as-a-judge for Constraint Verification

Judge	Cost (USD)	Time (min)	Macro F1 (%)	F1 Neg. (%)	Cohen's Corr. w/ Maj. Vote
Expert (the authors)	-	-	100.0	100.0	0.93
Human 1	300.0	-	85.1	75.9	0.77
Human 2	300.0	-	80.0	66.9	0.66
Majority Vote	-	-	96.4	94.1	1.00
GPT-4	19.5	-	73.7	54.9	0.42
GPT-3.5-Turbo	1.0	-	51.3	16.6	0.09
GPT-4-Turbo	6.5	-	72.6	54.8	0.46
+ CoT	**8.3**	**-**	**79.0**	**65.5**	**0.50**
Mistral v0.2		10	50.4	11.4	0.02
+ CoT	-	26	53.7	21.9	0.18
Weakly Supervised	**-**	**236**	**63.3**	**39.5**	**0.28**

Judging LLM responses from Mistral and Vicuna to test split of RealInstruct

- 🥇 **GPT-4-Turbo w/ CoT prompt** offers a more performant and cheaper alternative to GPT-4, with comparable to human performance.
- 💸 **Open-source LLMs are unreliable judges**, even when supervised with data from validation split of **RealInstruct** and annotations from GPT-4-Turbo + CoT.

LLMs' ability to follow multi-constrained instructions

Model	Instruction-level Accuracy	Constraint-level Accuracy
GPT-4	**78.80%**	**91.90%**
GPT-3.5-Turbo	73.80%	84.00%
Mistral 7B v0.2	**75.20%**	**87.80%**
Zephyr 7B β	70.50%	84.70%
Vicuna 7B v1.3	61.30%	77.80%

Performance on RealInstruct with GPT-4-Turbo + CoT as judge

- 📉 **Even the best LLM of the study (GPT-4) fails to meet at least one constraint on over 21% of instructions.**
- 🔢 LLMs often struggle with constraints involving **numbers**, **negations**, or **long instructions with large number of constraints**.

Effectiveness of our DeCRIM pipeline

Strategy	Decomposer	Critic	RealInstruct Best N	RealInstruct Instruction Acc (%)	RealInstruct Constraint Acc (%)	IFEval Best N	IFEval Instruction Acc (%)	IFEval Constraint Acc (%)
GPT-4	-	-	-	78.8	91.9	-	79.3	85.4
Conv.	-	-	-	75.2	87.8	-	60.1	66.3
Make sure	-	-	-	76.8	88.6	-	60.1	67.2
Self-Refine	-	-	2	77.2 (↑0.4)	88.7 (↑0.1)	2	59.5 (↓0.6)	66.4 (↓0.8)
DeCRIM (ours)	Self	Self	6	75.2 (↓1.6)	88.9 (↑0.3)	4	60.1 (0.0)	67.5 (↑0.3)
	Self	**Supervised**	**10**	**80.5 (↑3.7)**	**90.9 (↑2.3)**	**10**	**60.8 (↑0.7)**	**67.3 (↑0.1)**
	Oracle	Self	4	78.5 (↑1.7)	90.2 (↑1.6)	6	62.3 (↑2.2)	69.1 (↑1.9)
	Oracle	Supervised	**10**	**82.4 (↑5.6)**	**91.7 (↑3.1)**	10	64.9 (↑4.8)	71.6 (↑4.4)
	Oracle	GPT-4	-	-	-	**4**	**68.2 (↑8.1)**	**74.1 (↑6.9)**
	Oracle	**Oracle**	**10**	**93.7 (↑16.9)**	**95.2 (↑6.6)**	**8**	**80.4 (↑20.3)**	**83.5 (↑16.3)**

Proprietary
Baselines
Fairly Comparable
Realistic Ablation
Unrealistic ablation (upper bound)

DeCRIM w/ Mistral with strong prompt (*Make sure*) and N_{max} = 10

❌ LLMs Can't Self-Refine

- Self-Refine baseline, and Self-Critic + Self-Decomposer led to poor results.
- Low-quality feedback: over-refining good responses while ignoring bad ones.

💪 DeCRIM works even with a Weak Critic

- using better Decomposer or Weak Critic lead to significant improvements
- **LLMs can improve with minimally reliable feedback.**

🏆 Open LLMs can correct itself when given high-quality feedback

- With an Oracle Critic and Decomposer, Mistral outperforms GPT-4 on both datasets.

🚀 DeCRIM Boosts Response Quality

- Most responses quality stayed the same, but when changed, revised ones were often preferred (particularly when revision was successful).
- Too many revisions can reduce quality.

References

Zhou, Jeffrey, et al. *"Instruction-following evaluation for large language models."* arXiv preprint arXiv:2311.07911 (2023).

Pan, Liangming, et al. *"Automatically correcting large language models: Surveying the landscape of diverse self-correction strategies."* TACL(2024).

Kamoi, Ryo, et al. *"When Can LLMs Actually Correct Their Own Mistakes? A Critical Survey of Self-Correction of LLMs."* TACL (2024).

Jiang, Yuxin, et al. *"Followbench: A multi-level fine-grained constraints following benchmark for large language models."* ACL 2024.

Truong, Thinh Hung, et al. *"Language models are not naysayers: an analysis of language models on negation benchmarks."* *SEM @ ACL 2023

Take a photo to learn more about the paper and the presenter: