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 Paris - IP Paris, Meta AI, UNC Chapel Hill, UCLA

About Me

- PhD student at École Polytechnique / Télécom Paris (IP Paris)
- Master MVA (Applied Math & AI) at ENS Paris-Saclay
- Engineering Degree from Universidade de São Paulo
- Research Internships at Apple, Amazon, Meta and Naver Labs.
- Publications on LLMs, Multilingual NLP, Low-Resource NLP, Zero-shot learning, Speech/Text Translation, Robustness.

🤔 Do LLMs do exactly what we ask them to?

- 👍 LLMs excel at overall instruction-following!
- 🚫 LLMs fail to satisfy all requests in multi-constrained user instructions.
- ⚠️ Existing benchmarks are synthetic
    - Lacking real-world complexity
    - Artificially hard constraints
        - Potentially leading research in the wrong direction, with results that may not apply to real scenarios.

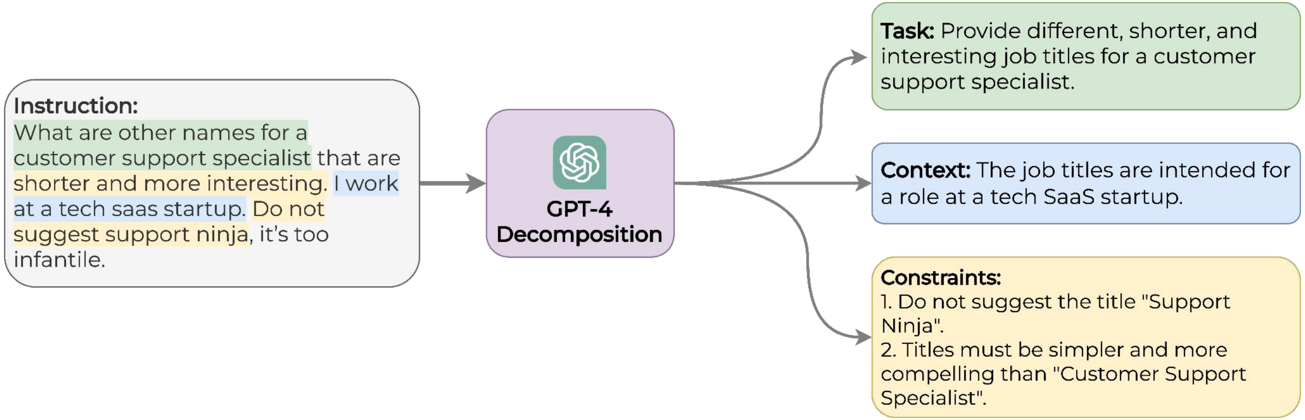
🎯 Our contributions

- 📊 **REALINSTRUCT:** The first benchmark using *real user requests* to evaluate LLMs on multi-constrained instruction following.
- 🔄 **DECRIM:** The first System-2 self-correction pipeline that improve LLMs to follow multi-constrained instructions, without making any assumptions about the constraints.
- 🧑‍⚖️ **LLM-as-a-Judge:** We analyse the success of LLMs as evaluators to benchmark other LLMs and to guide self-correction for multi-constrained instructions.

02 The REALINSTRUCT benchmark

🏗️ Dataset Construction

- 🔍 **Data Filtering:** Mining non-code, English user instructions with constraints from a pool of real user conversations with AI.
- 🧩 **Decomposition**



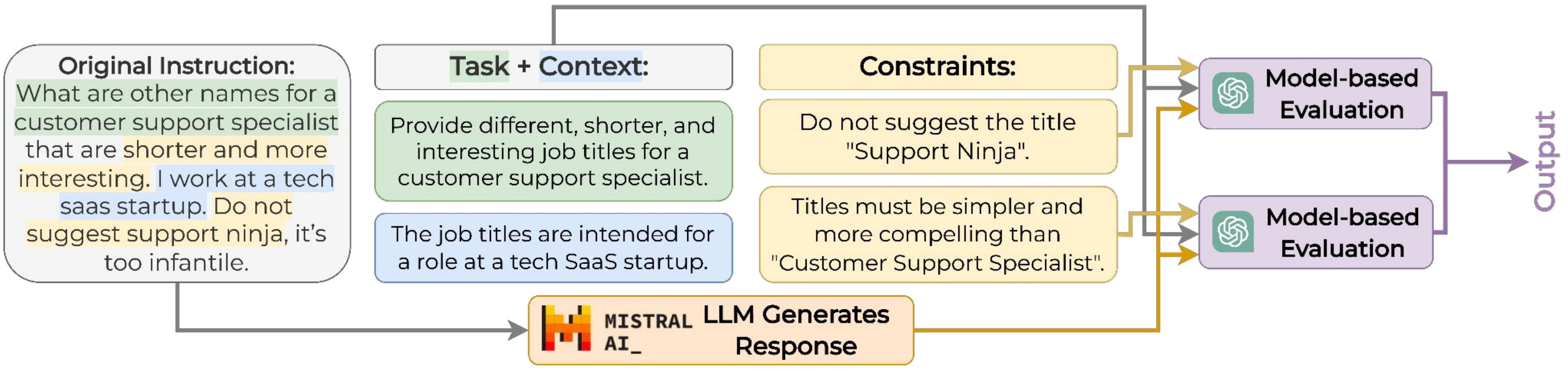
🏗️ Dataset Construction

- 🔍 **Data Filtering:** Mining non-code, English user instructions with constraints from a pool of real user conversations with AI.
- 🧩 **Decomposition:** Use GPT-4 to break down user requests into Task+Context and Constraints.
- 🕵️ **Human Validation:** Manual validation ensures accuracy of the decomposed data.

⚖️ Comparison with representative works

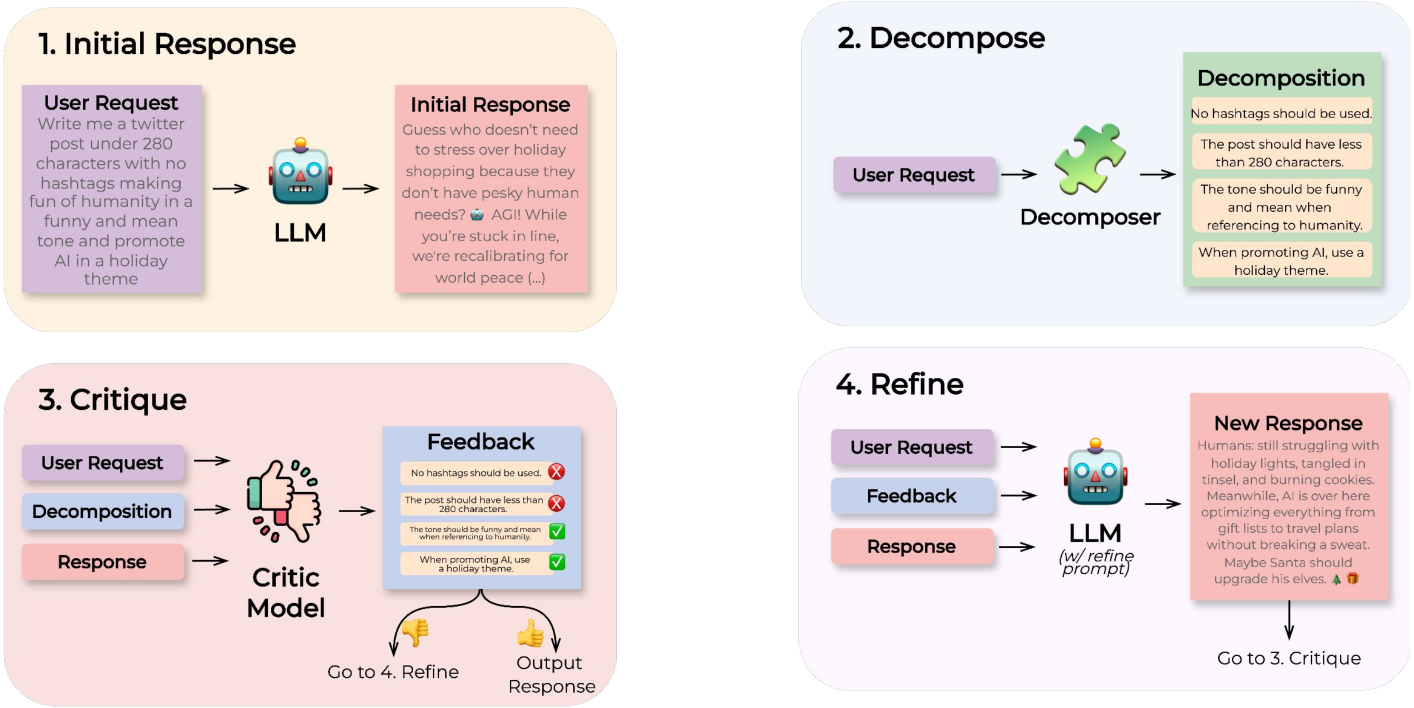
Benchmark	Instruction source	Constraints source	Evaluation	Size (Instructions)	Constraint types	Avg.Constraints per Instruction
COLLIE (Yao et al., 2024a)	Synthetic	Synthetic	Rule-based	2,080	13	N/A
IFEval (Zhou et al., 2023a)	Synthetic	Synthetic	Rule-based	541	25	1.4
FollowBench (Jiang et al., 2024)	Crowdsourced + Synthetic	Synthetic	Model-based + Rule-based	795	6	5
InfoBench (Qin et al., 2024)	Crowdsourced	Crowdsourced	Model-based + Rule-based	500	5	4.5
REALINSTRUCT (ours)	**Real Users**	**Real Users**	**Model-based**	**302 (test) + 842 (val)**	**20+**	**3.5 (test)**

🧑‍💻 RealInstruct Benchmark Flow

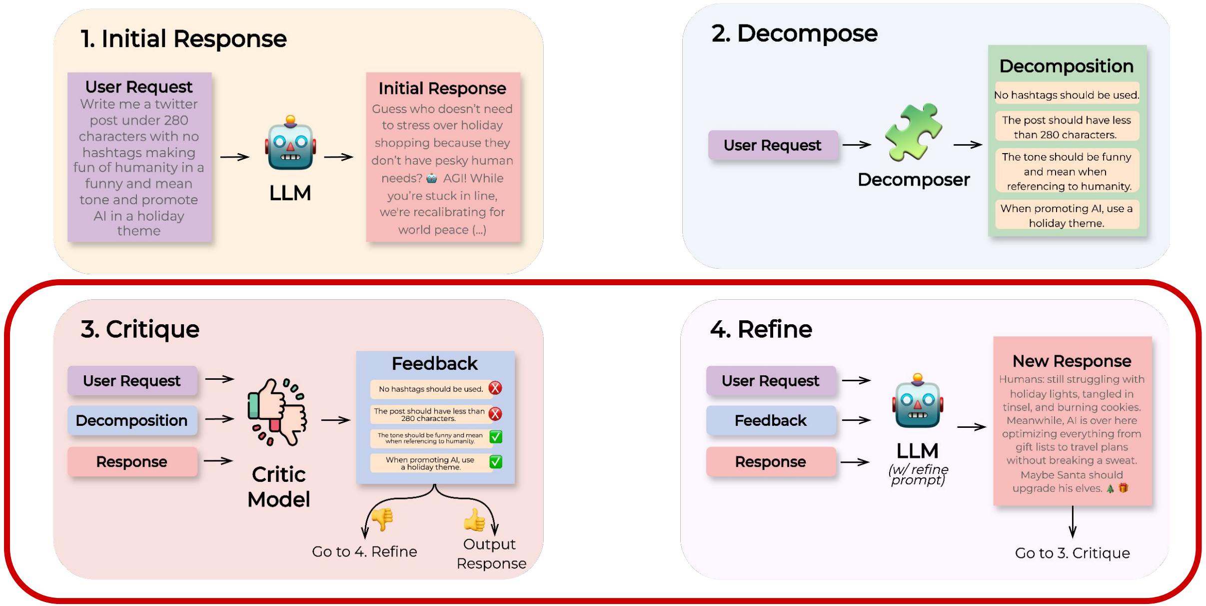


03 Decompose, Critique and Refine

DECRIM Pipeline



🔄 DECRIM Pipeline



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

⚖️ Comparison with previous works

- Most self-correction methods require Critic and Refining training.
- Recent prompt-based methods still struggle in real scenarios:
  - Lack specific constraint modeling (e.g., Self-Refine).
  - Make assumptions about constraint, like independence (e.g., BSM and other System 2 methods).
- **DECRIM**
  - Does not require LLM training for generation/refining.
  - **Works with any constraint:** does not make assumptions.

04 Experiments and Results

Part I

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

- Are LLMs enough reliable? Or as reliable as humans would be?
  - For Benchmarking on **REALINSTRUCT**.
  - For Criticizing on DeCRIM pipeline.
- Study proprietary and Open-source LLMs
  - Compare performance on **REALINSTRUCT** responses from Mistral and Vicuna
- Different adaptation approaches:
  - Prompt-based approaches (with and without CoT)
  - Mistral Weakly Supervised Fine-tuning

🧑‍⚖️ 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**

- GPT-4-Turbo + CoT prompt offers a more performant and cheaper alternative to GPT-4.
  - comparable to human performance.

Corr. GPT-4-Turbo vs. Expert: 0.58
Corr. Human 2 vs. Expert: 0.60

🧑‍⚖️ 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**

- Open-source LLMs offer lower costs but are unreliable judges.
- Even Weakly Supervised Mistral falls short.

We adopt GPT-4-Turbo + CoT for RealInstruct Benchmark

Part II

🎛️ LLMs' ability to follow multi-constrained instructions

Model	Instruction Accuracy	Constraint 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%

Results on REALINSTRUCT - Test Set

- 🔬We tested representative models, top performant on LLM Leaderboards at the time of study
  - Open LLM Leaderboard, Chatbot Arena

🎛️ LLMs’ ability to follow multi-constrained instructions

Model	Instruction Accuracy	Constraint 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%

Results on **REALINSTRUCT** - Test Set

- 📉 Even the best LLM of the study (GPT-4) fails to meet at least one constraint on **over 21%** of instructions.

🎛️ LLMs' ability to follow multi-constrained instructions

Model	Instruction Accuracy	Constraint 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%

Results on **REALINSTRUCT** - Test Set

- 🚧 Open-source Mistral outperforms proprietary GPT-3.5 but falls short of GPT-4's performance.

🎛️ LLMs' ability to follow multi-constrained instructions

- 🔢 LLMs often struggle with constraints involving **numbers**, **negations**, or **long instructions with large number of constraints**.

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**

Part III
💡 Effectiveness of our **DECRIM** pipeline

- Using Mistral as underlying model for DECRIM
- Datasets
  - REALINSTRUCT and IFEval (Zhou et al., 2023) (standard on LLM Leaderboards)
- Strong baselines
  - GPT-4, *"Make Sure"* prompt, Self-Refine (Madaan et al., 2023)
  - DECRIM with first generation *"Make Sure"* prompt and $N_{max} = 10$
- Decomposer and Critic:
  - LLM itself (Self-Decomposer and Self-Critic)
  - Mistral Weakly Supervised as Critic
  - Oracle Critic and Oracle Decomposer

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 due to low-quality feedback
    - Leads to over-refining good responses while ignoring bad ones.

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)**

Weak Critic + Ideal Decomp.

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

GPT-4 is Weak Critic for IFEval Macro F1: 62.9%

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

- 💪 **DECRIM is Effective even with Weak Critic**
  - Weak but minimally reliable Critic yields performance gains.
  - A Better Decomposer also enhances results.
  - Combining Better Decomposer + Weak Critic leads to significant improvements.
  - **Takeaway:** LLMs benefit from even minimally reliable feedback.

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

- 🏆 **Open LLMs can correct its outputs when given high-quality feedback**
  - With an Oracle Critic and Decomposer, Mistral outperforms GPT-4 on both datasets.
  - Better the feedback -> Better the performance.
  - Not following constraints is also a matter of alignment.

💡 Results on the Effectiveness of **DECRIM**

- 🚀 **DECRIM boosts Response Quality**
  - Response quality mostly stayed the same, but when changes occurred, the revised versions were often preferred.
  - Strong correlation between successful revision and the response quality.
  - However, too many revisions can reduce quality.
- ⏳ **Computation Overhead**
  - Mitigation: Refinement triggered only when Critic detects unmet constraints, with ~25% of responses revised after the first pass.
  - Need for revision drops exponentially, leading to a sublinear time growth as N_{max} increases.

05 Final Remarks

Summary of our Findings

- **Problem is still relevant:** Best LLM (GPT-4) missed at least one constraint on **over 21% of instructions.**
- **LLM-as-a-Judge:** Proprietary models match human reliability, while open models still lag.
- **DeCRIM:** Achieves up to 8% improvement with minimally reliable feedback and up to 34% with high-quality feedback, outperforming proprietary models in all datasets
    - System 2 approaches push LLM capabilities to the limit.
    - Strategies gaining momentum with Sys-2 reasoning models like GPT-o1

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

Thank you!

Questions? Suggestions?
Want to learn more?
Scan the QR code!