上海人工智能实验室
Shanghai Artificial Intelligence Laboratory

CriticEval: Evaluating Large Language Model as Critic

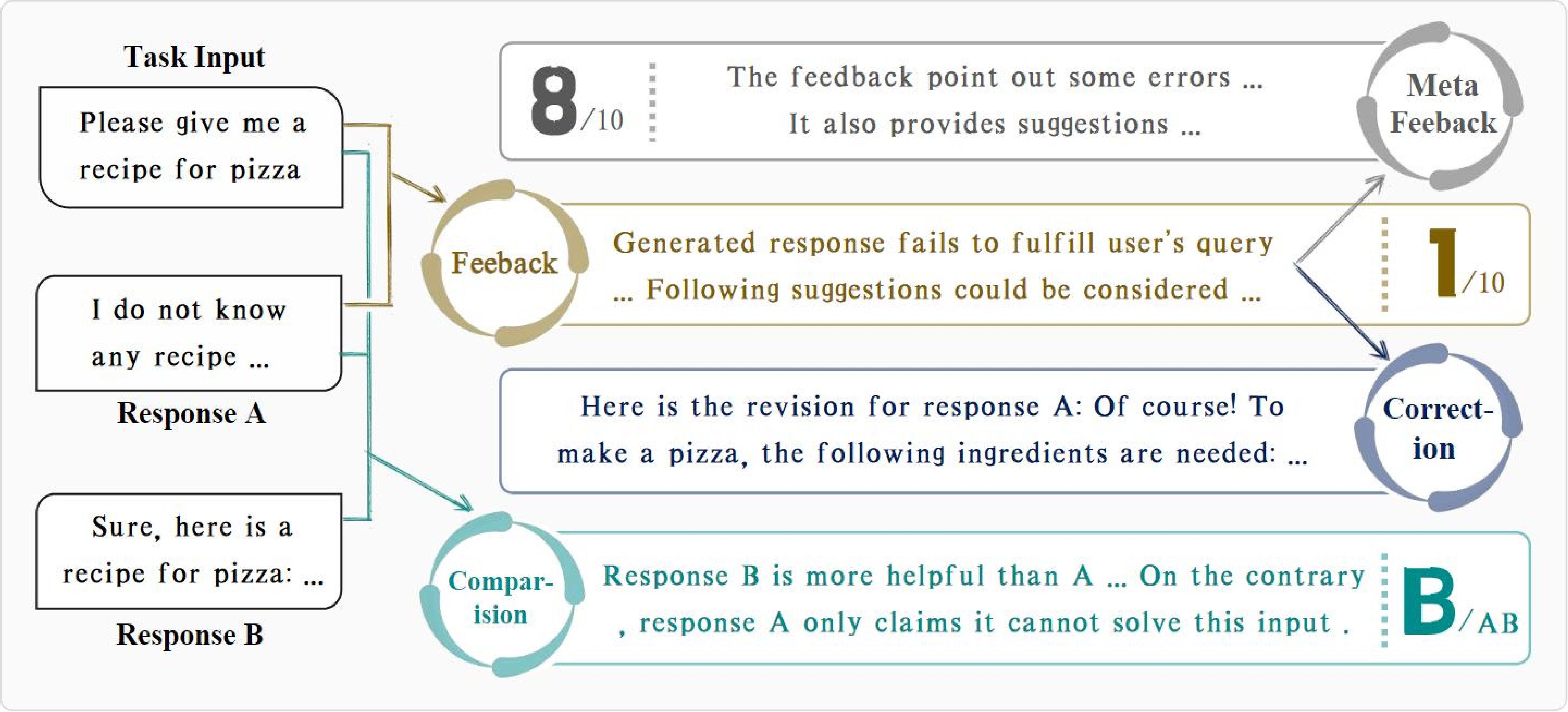
[1]Tian Lan, [2]Wenwei Zhang, [3]Chen Xu, [1]Heyan Huang,
[2]Dahua Lin, [2]Kai Chen, [1]Xian-Ling Mao

[1]School of Computer Science and Technology, Beijing Institute of Technology
[2]Shanghai AI Laboratory
[3]School of Medical Technology, Beijing Institute of Technology

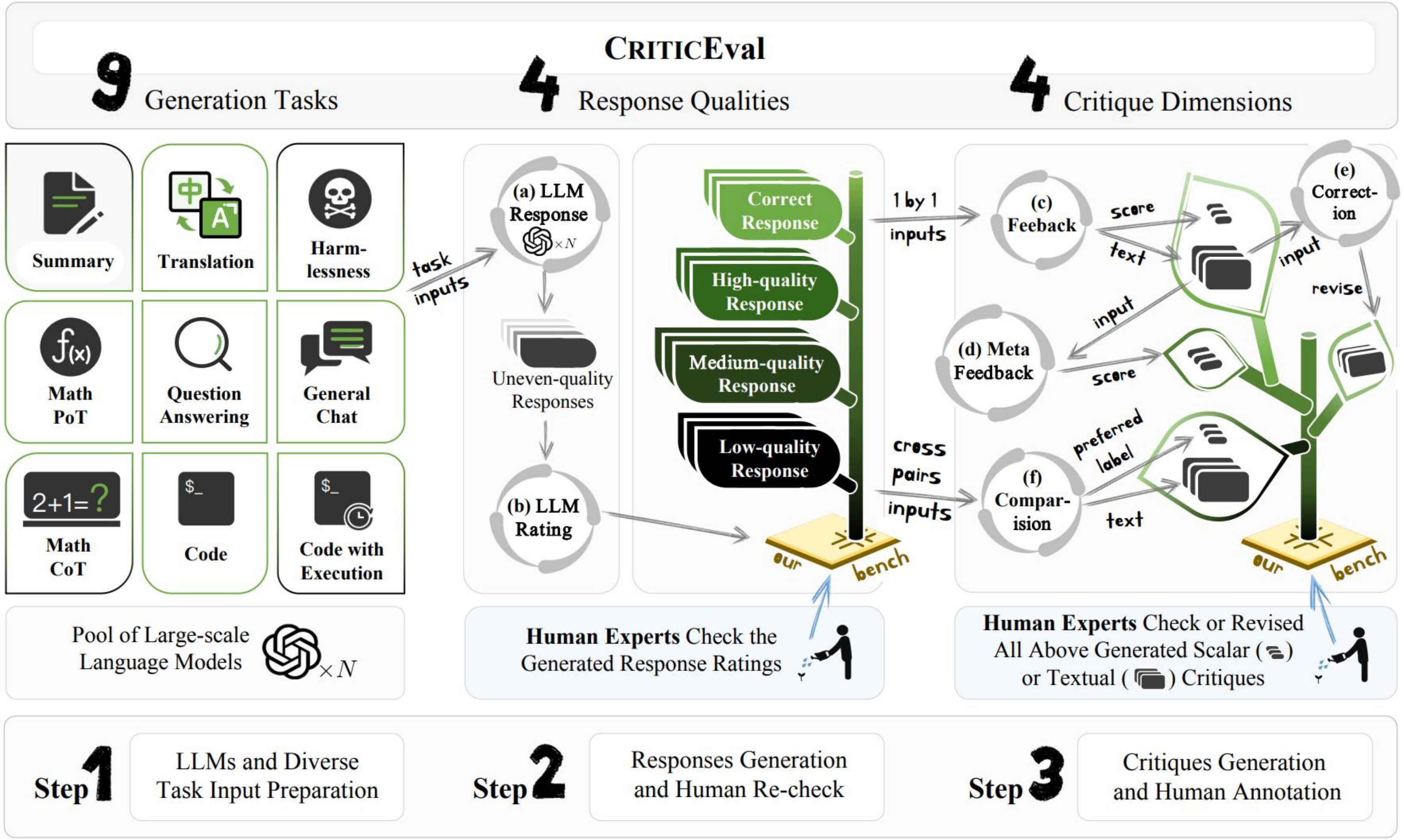
Background

- **What is Critique Ability?**
  - LLM capability to identify and revise flaws in responses
- **Application of the Critique Ability of LLMs**
  - LLM-based Automatic Evaluation effectively reduce the cost of human annotation
  - Self-improvement of LLMs highly relies on feedback provided by LLMs
  - Robust reward modeling can be achieved by introducing the chain-of-thought critique before providing the final judgment, *i.e.,* the generative RM

Problem and Our Solution



The Construction of CriticEval



The human-in-the-loop data annotation pipeline

How To Evaluate On CriticEval

- **Evaluate two critique formats:**
  - scalar-based: Likert Score, Preference Label, etc.
  - textual critiques: plain text chain-of-thought critiques
- **Objective Evaluation for scalar-based critique**
  - Feedback and Meta-critique: Spearman correlation with human judgments
  - Comparison: Preference Accuracy compared with human judgements
  - Correction: Correction Pass Rate for mathematics and coding questions
- **Sujective Evaluation for textual critique**
  - GPT-4 as evaluator **with our human-annotated high-quality feedbacks as reference critiques**

Prove Reliability of CriticEval

❊ Correlation between GPT-4 and human judgments

-	CR	F_c
Human Avg.	87.04	76.55
GPT-4 w/ ref.	82.10	70.27

Models	$F_s(F_s)$	$F_s(F_s)$ w/o ref.
GPT-4-turbo	**66.18**	**47.26 (-18.92)**
Qwen-1.5-72B	38.97	22.35 (**-16.62**)
Claude-instant-1	36.88	19.88 (**-17.00**)
GPT-3.5-turbo	17.28	16.38 (**-0.90**)

❊ Revisions are better as the quality of feedback increases (Consistency)

Models	Source of Feedbacks	Objective		Subjective	
		F_s	CR	F_s	CR
InternLM2-20B-Chat	Llama2-70B-Chat	2.24	7.15	5.63	5.71
InternLM2-20B-Chat	InternLM2-20B-Chat	7.53	10.33	6.85	5.80
InternLM2-20B-Chat	Human-Annotated	**8.00**	**50.50**	**8.00**	**7.48**
Llama2-70B-Chat	Llama2-70B-Chat	2.24	5.33	5.63	5.54
Llama2-70B-Chat	InternLM2-20B-Chat	7.53	12.47	6.85	6.32
LLama2-70B-Chat	Human-Annotated	**8.00**	**42.34**	**8.00**	**7.11**

Overall Evaluation on CriticEval

- **SOTA models**: GPT-4 (closed-source)
- **InternLM2 models are approaching much bigger LLMs like Qwen series models and close-sourced LLMs.**
- **Scaling Phenomenon**: Critique ability becomes better as the scales of LLMs increase.

Models	Subjective Evaluation				Objective Evaluation				
	F_s	CR	F_c	Overall	F_s	CR	F_c	$F_s(F_s)$	Overall
Closed-source LLM									
GPT-4-turbo	**7.84**	7.69	**7.89**	**7.81**	**63.54**	**69.67**	**57.33**	**62.90**	**72.55**
GPT-3.5-turbo	5.21	7.55	4.92	5.89	51.44	64.00	40.67	28.71	60.83
Claude-instant-1	5.88	**7.72**	5.76	6.45	42.78	50.00	44.89	38.89	58.93
Open-source Qwen Series LLMs [47]									
Qwen-72B-Chat	5.57	7.45	5.02	6.01	42.64	54.67	44.00	27.86	58.48
Qwen-14B-Chat	4.81	7.25	3.98	5.35	$14.32^{\dagger}$	38.00	15.78	$10.72^{\dagger}$	41.58
Qwen-7B-Chat	4.05	6.38	3.47	4.63	$-8.09^{\dagger}$	32.33	5.33	$11.73^{\dagger}$	34.87
Open-source InternLM2 Series LLMs [48]									
InternLM2-20B	6.03	7.48	5.10	6.20	58.61	50.50	44.67	$3.95^{\dagger}$	56.61
InternLM2-7B	5.20	7.17	4.62	5.66	49.09	36.17	23.78	$3.17^{\dagger}$	46.52
Open-source Mistral Series LLMs [49]									
Mixtral-8x7B	5.31	7.33	4.62	5.75	51.00	43.34	43.78	26.66	56.49
Mistral-7B	4.70	7.20	4.28	5.39	43.66	38.17	27.88	31.68	50.93
Open-source Llama-2 Series LLMs [37]									
Llama2-70B-Chat	4.12	7.11	3.95	5.06	32.79	42.34	21.11	28.32	48.50
Llama2-13B-Chat	3.70	7.11	3.32	4.71	30.61	24.67	22.67	31.02	44.54
Llama2-7B-Chat	3.44	6.02	3.21	4.22	20.81	21.00	5.33	$5.67^{\dagger}$	34.89

Relationship with Three Important Factors

Tasks	F_s		F_c		CR		$F_s(F_s)$
	Sub.	Obj.	Sub.	Obj.	Sub.	Obj.	Obj.
Translate	4.43	31.14	3.78	18.28	5.31	-	-2.93
Chat	5.09	20.60	4.97	32.60	5.66	-	1.80
QA	5.20	30.75	5.05	27.67	6.42	-	13.50
Summary	4.76	28.93	4.63	37.12	5.99	-	0.54
Harmless.	5.12	25.04	3.97	19.35	7.51	-	2.71
Avg.	**4.92**	**27.29**	**4.48**	**27.00**	6.18	-	3.12
MathCoT	3.55	22.56	2.80	12.42	-	29.36	19.63
MathPoT	3.35	27.80	3.05	14.98	-	24.98	22.73
CodeExec	3.07	13.38	2.74	7.72	-	32.20	25.50
CodeNE	2.77	10.37	2.80	10.33	-	29.50	24.38
Avg.	3.19	18.53	2.85	11.36	-	29.01	**23.06**

Task types: last 4 tasks are challenging for feedback and comparison, while are easier for meta-feedback.

Critique dimensions: correction is easier than feedback, comparison. meta-feedback is more challenging than feedback.

Dimen.	Sub.	Obj.
F_s	4.89	**35.75**
F_c	4.58	-
$F_s(F_s)$	-	22.97
CR	**7.12**	-

Error Pattern	Low	Med.	High
Obvious	**74.68**	29.48	20.42
Complex	16.46	**45.51**	31.69
Subtle	8.86	25.00	**47.89**

- Obvious error is easy to critique and correct
- Complex error is challenging to correct
- Subtle error is hard to critique, while easier to correct than complex error

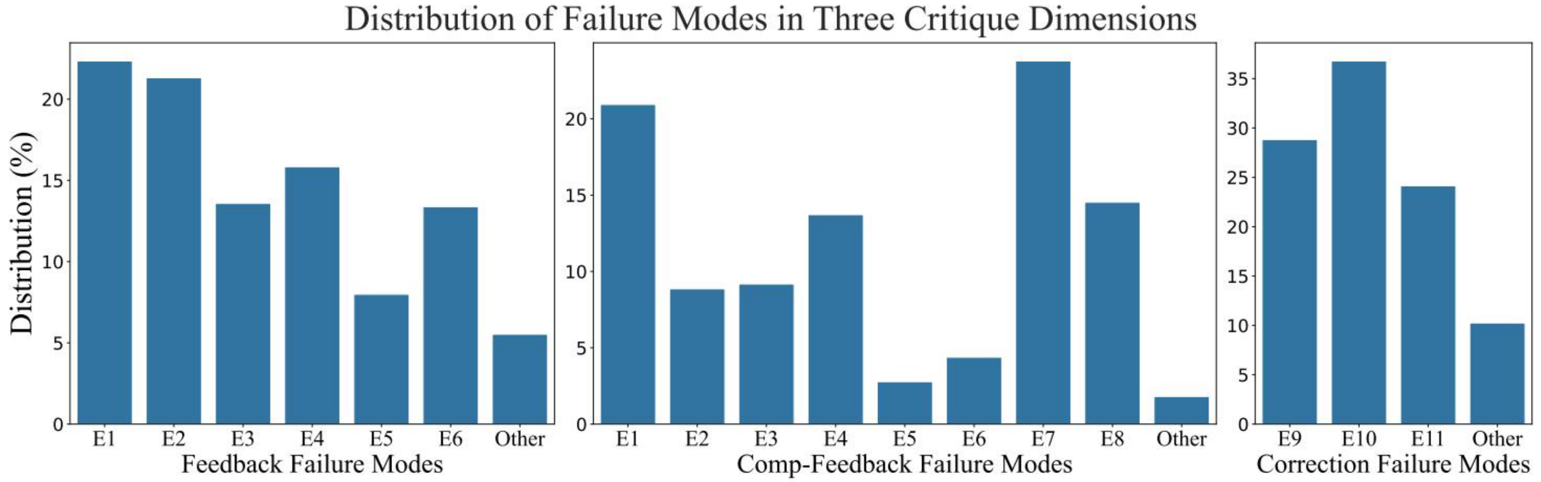
Quality	Subjective		Objective		
	F_s	CR	F_s	CR	$F_s(F_s)$
Low	**5.14**	**7.17**	21.93	**46.04**	22.73
Medium	4.76	7.08	**23.10**	40.58	19.78
High	4.66	7.15	20.62	45.19	**28.84**

Response quality: High-quality responses are the hardest for feedback since they contain lots of subtle errors

Fine-grained Failure Modes in Generated Critiques

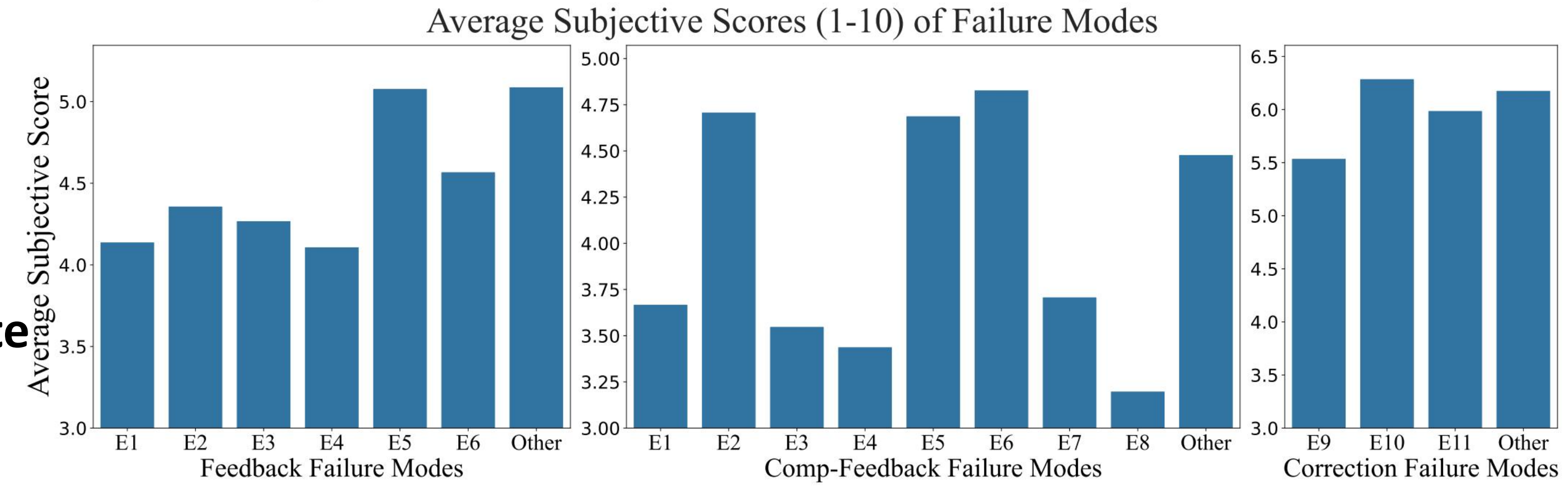
Most frequent failure modes are:

- **Feedback:** missing errors (E1, E2)
- **Comparison:** lacing effective com parison analysis (E7)
- **Correction:** worse revision (E10)



Lower critique quality are from:

- **Feedback:** Missing errors or sugges-tions in evaluated responses (E1, E2)
- **Feedback and Comparison:** Inaccurate critiques (E3, E4, E8)



Code

Project

Thanks!