Universal Multi-Dimensional Text Evaluation Enhanced with Auxiliary Evaluation Aspects

VIRGINIA TECH.

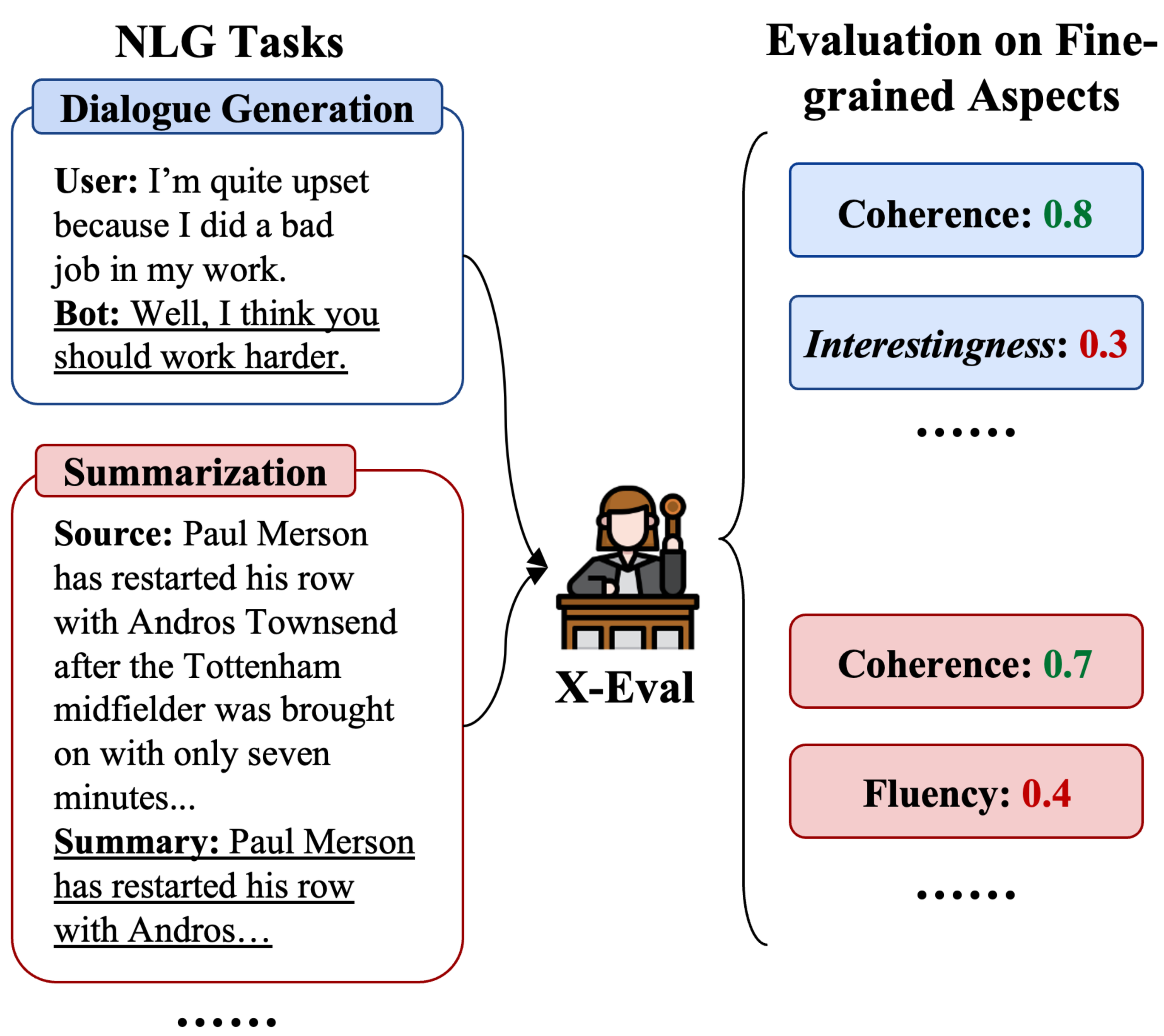
Minqian Liu Ying Shen Zhiyang Xu Lifu Huang

Department of Computer Science, Virginia Tech

{minqianliu, yings, zhiyangx, lifuh}@vt.edu

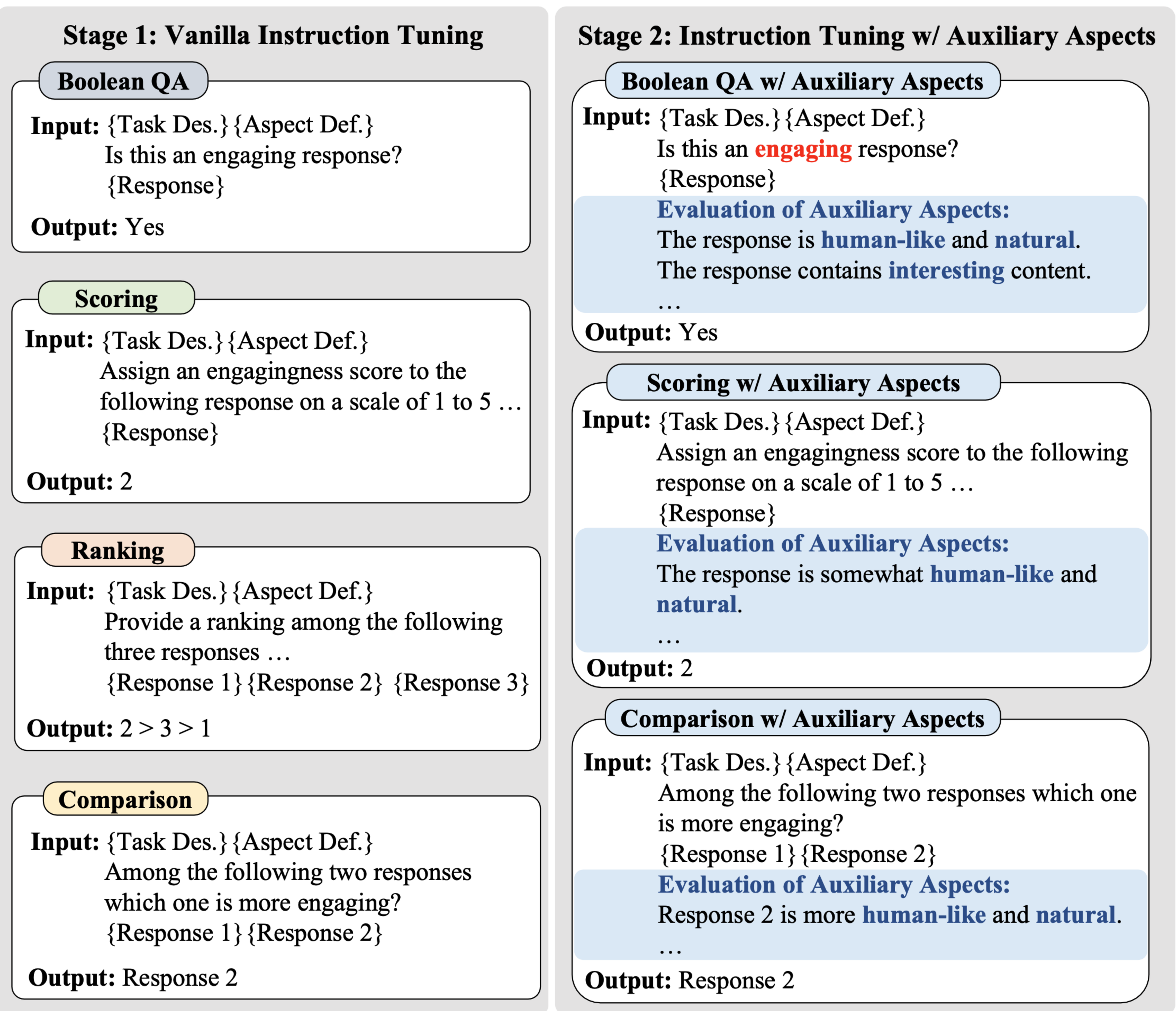
Overview

- Background: Multi-aspect evaluation is crucial to assess the performance of language models comprehensively.
- Problem: How to generalize to any *customized* evaluation aspects?
- Solution: Two-stage instruction tuning to (1) learn to follow customized evaluation instructions, and (2) exploit the connections between fine-grained evaluation aspects.
- Contributions: A universal evaluator with the following strengths:
- (1) Generalization ability: adapt to user-specified instructions in a zero-shot manner with a unified model.
- (2) Strong performance with high efficiency: achieves strong performance with only 780M parameters.
- (3) Reference-free and open-source.

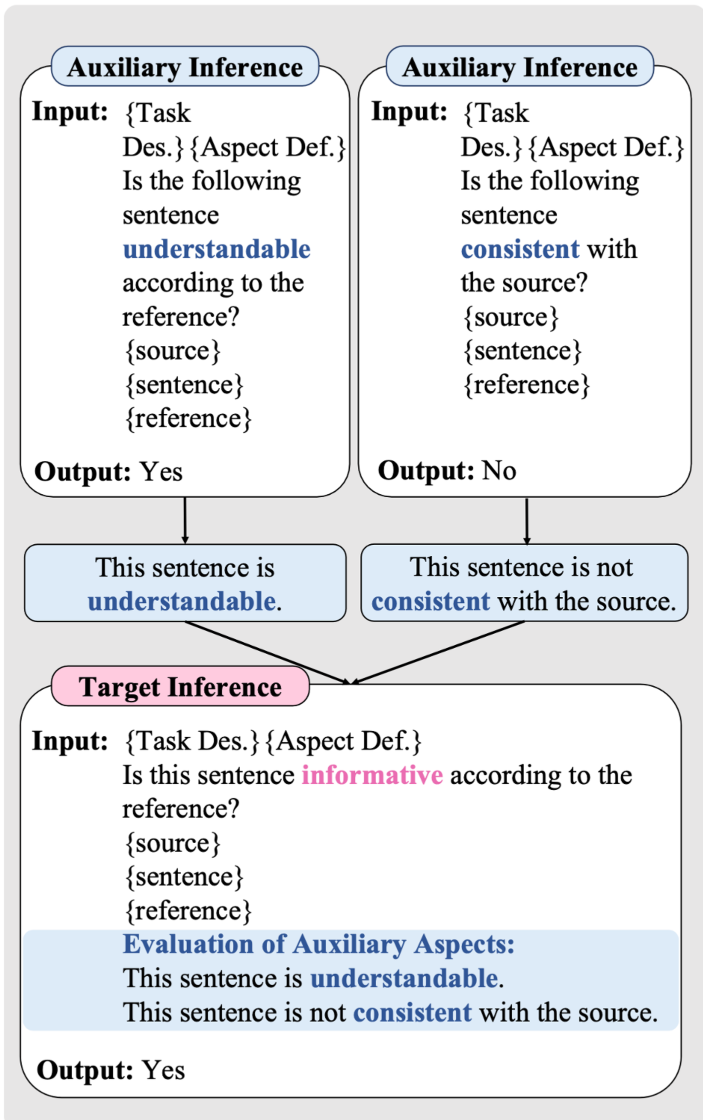


X-EVAL: Two-stage Instruction Tuning with Auxiliary Aspects

- Derive four types of evaluation tasks to increase the task diversity.
- Enhance instruction tuning with auxiliary aspects.
- Introduce AspectInstruct, the first multi-aspect evaluation instruction tuning dataset with 27 diverse aspects on three NLG tasks.



Inference with Auxiliary Aspects



Algorithm 1: Inference Pipeline

Input: Set of evaluation aspects $\mathcal{A}$, Target aspect a_t, NLG system's output x, Additional set of texts $\mathcal{S}$, Scoring function $f(\cdot)$, Evaluation verbalizer $v(\cdot)$, Similarity measure $sim(\cdot)$, Sentence encoder $\mathcal{E}$

Output: Target score c_t

```
// Determine top-k auxiliary aspects
L ← {(sim(E(a), E(a_t)), a) | a ∈ A \ {a_t}}
Sort L in descending order based on similarity
A^R ← first k aspects from sorted L
// Generate verbalized evaluation
   results for auxiliary aspects
Initialize an empty auxiliary evaluation set H
for a_r ∈ A^R do
    // Score for auxiliary aspect
    c_r ← f(x, S_r, a_r)
    // Add verbalized evaluation to the
       auxiliary evaluation set
    H ← [H; v(c_r, a_r)]
S_t ← [S_t; H]
// Evaluate the target aspect
c_t ← f(x, S_t, a_t)
return c_t
```

Experiments & Discussions

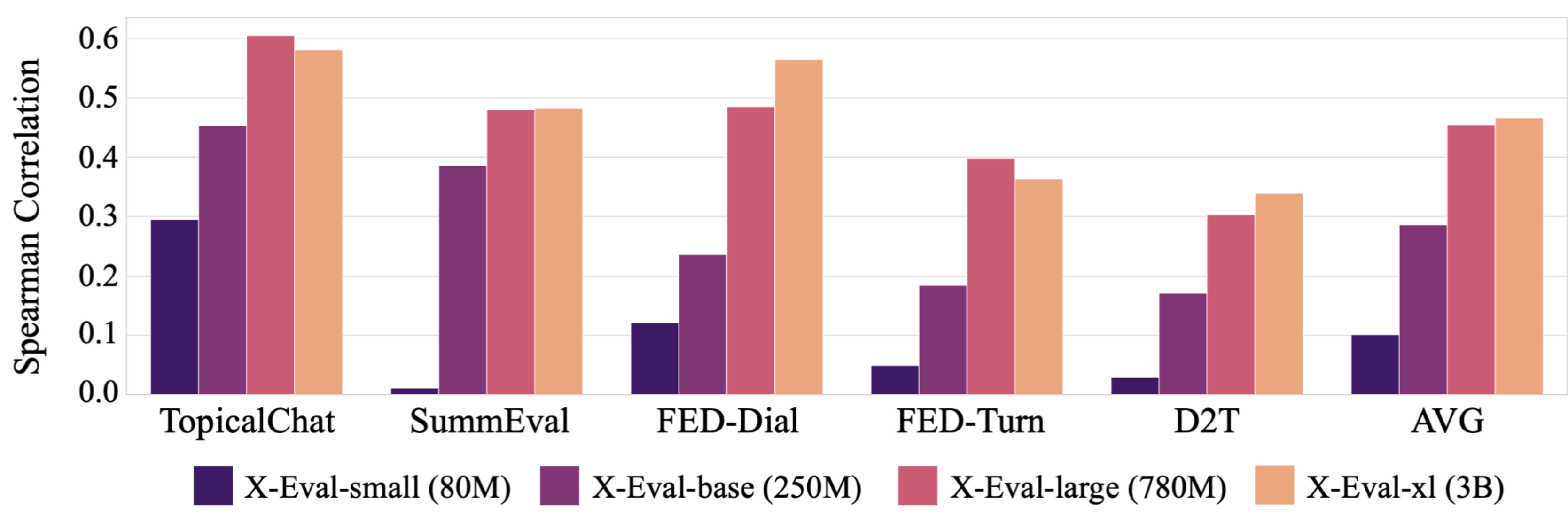
Meta-evaluation on dialogue based on unseen aspects on FED:

Metrics	Dialogue-level DEP	LIK	UND	FLE	INF	INQ	AVG	Turn-level INT	SPE	COR	SEM	UND	AVG
BARTScore (Yuan et al., 2021)	0.082	0.099	-0.115	0.093	0.092	0.062	0.052	0.159	0.083	0.076	0.100	0.120	0.128
DynaEval (Zhang et al., 2021)	0.498	0.416	0.365	0.383	0.426	0.410	0.416	0.327	0.346	0.242	0.202	0.200	0.263
UniEval (Zhong et al., 2022)	0.046	0.009	-0.024	-0.003	-0.070	0.085	0.030	0.435	**0.381**	0.125	0.051	0.082	0.215
GPTScore (GPT-3-d01) (Fu et al., 2023)	**0.669**	**0.634**	0.524	0.515	**0.602**	0.503	**0.574**	0.501	0.214	0.434	**0.444**	0.365	0.392
GPTScore (GPT-3-d03) (Fu et al., 2023)	0.341	0.184	0.196	0.072	0.317	-0.101	0.168	0.224	0.151	0.428	0.405	0.311	0.304
G-Eval (GPT-3.5)† (Liu et al., 2023)	0.339	0.392	0.123	0.344	0.232	0.101	0.259	0.30	0.280	0.430	0.390	0.274	0.335
G-Eval (GPT-4)† (Liu et al., 2023)	0.583	0.614	**0.602**	**0.587**	0.510	**0.551**	0.573	**0.506**	0.368	**0.522**	0.443	**0.438**	**0.455**
X-EVAL (Ours)	0.583	0.436	0.588	0.324	0.480	0.497	0.485	0.421	0.370	0.492	0.376	0.332	0.398
- w/o Training	0.377	0.387	0.394	0.424	0.370	0.417	0.395	0.250	0.175	0.296	0.289	0.225	0.247
- w/o Instructions	0.350	0.333	0.495	0.355	0.425	0.435	0.399	0.477	0.353	0.203	0.255	0.211	0.300
- w/o Stage-Two Tuning	0.388	0.324	0.555	0.384	0.582	0.437	0.445	0.372	0.282	0.418	0.329	0.311	0.342

Meta-evaluation on summarization on SummEval:

Metrics	Coherence ρ	Coherence τ	Consistency ρ	Consistency τ	Fluency ρ	Fluency τ	Relevance ρ	Relevance τ	AVG ρ	AVG τ
ROUGE-L (Lin, 2004)	0.128	0.099	0.115	0.092	0.105	0.084	0.311	0.237	0.165	0.128
MOVERSscore (Zhao et al., 2019)	0.159	0.118	0.157	0.127	0.129	0.105	0.318	0.244	0.191	0.148
BERTScore (Zhang* et al., 2020)	0.284	0.211	0.110	0.090	0.193	0.158	0.312	0.243	0.225	0.175
BARTScore (Yuan et al., 2021)	0.448	0.342	0.382	0.315	0.356	0.292	0.356	0.273	0.385	0.305
UniEval (Zhong et al., 2022)	0.495	0.374	0.435	0.365	0.419	0.346	0.424	0.327	0.443	0.353
GPTScore (Fu et al., 2023)	0.434	–	0.449	–	0.403	–	0.381	–	0.417	–
G-Eval (GPT-3.5) (Liu et al., 2023)	0.440	0.335	0.386	0.318	0.424	0.347	0.385	0.293	0.401	0.320
G-Eval (GPT-4) (Liu et al., 2023)	**0.582**	**0.457**	**0.507**	**0.425**	0.455	**0.378**	**0.547**	**0.433**	**0.514**	**0.418**
X-EVAL (Ours)	0.530	0.382	0.428	0.340	**0.461**	0.365	0.500	0.361	0.480	0.362
- w/o Training	0.187	0.131	0.193	0.152	0.135	0.104	0.444	0.325	0.240	0.178
- w/o Instructions	0.458	0.333	0.414	0.328	0.395	0.309	0.496	0.359	0.441	0.333
- w/o Stage-Two Tuning	0.536	0.385	0.413	0.326	0.455	0.360	0.503	0.363	0.476	0.359

Effect of the scale of language model backbones:



Conclusion

- Present **X-Eval**, a novel two-stage instruction-tuning framework for text evaluation across seen and unseen aspects.
- Collect **AspectInstruct**, the first instruction-tuning dataset for multi-aspect evaluation.
- Our method achieves a **comparable if not higher correlation with human judgments** compared to the state-of-the-art NLG evaluators.