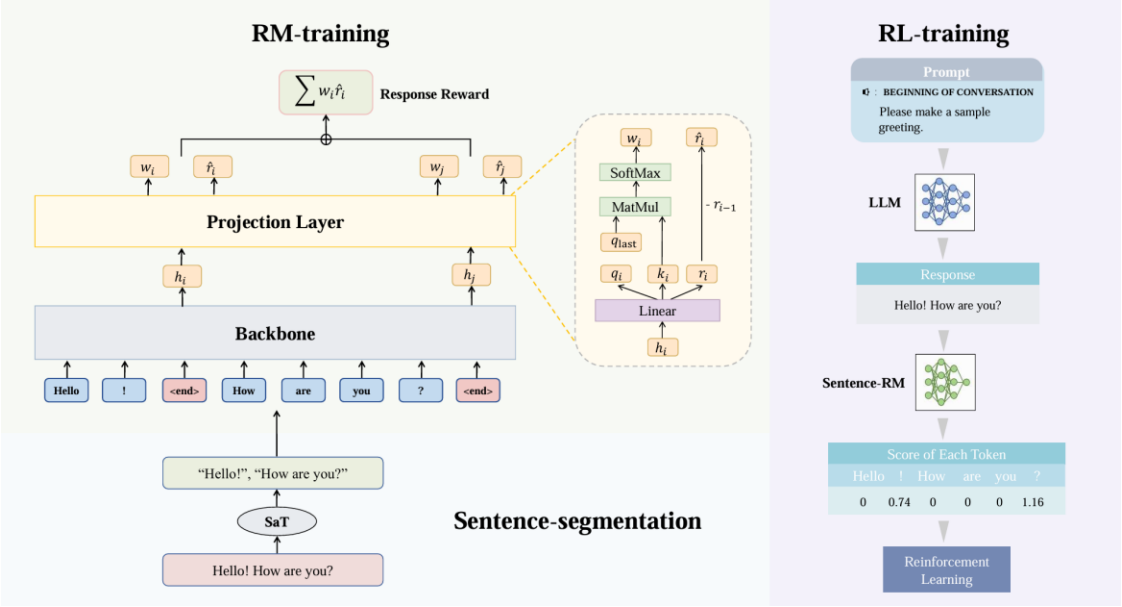
Sentence-level Reward Model can Generalize Better for Aligning LLM from Human Preference

Wenjie Qiu, Yi-Chen Li, Xuqin Zhang, Tianyi Zhang, Yihang Zhang, Zongzhang Zhang, Yang Yu

Frontiers of Computer Science, DOI: 10.1007/s11704-026-51483-4

Problems & Ideas

- Problems of conventional reward modeling approaches:
  - Response-level reward models provide only sparse single reward at the end of generation.
  - Token-level reward models lack explicit complete semantic information.
- Ideas: Propose a sentence-level reward model as an intermediate grained solution, which takes complete sentence as the basic reward unit.



Given an input response, a segmenter first identifies sentence boundaries. The projection layer then transforms the hidden states of boundary tokens into sentence-level rewards and weights. The overall response-level reward is their weighted sum. During RL training, the sentence-level reward model provides non-zero rewards at sentence boundaries.

Main Contributions

- Contributions:
  - A novel sentence-level reward model for LLM alignment, computing fine-grained sentence rewards via differential operations and aggregating them through attention;
  - A sentence boundary-guided reward assignment scheme to tackle sparse reward in RL, providing stable and semantically complete fine-grained optimization signals;
  - A sentence reward-driven LLM alignment method to mitigate length hacking and semantic bias, enhancing LLM performance and quality.

Method	RewardBench	AlpacaEval 2.0			Arena-Hard	
		LC	WR	Len	WR	Len
Response	78.00	63.78	67.88	1587	61.20	2044
Token	76.37	60.82	69.61	2105	69.20	2278
Segment	77.96	65.01	69.94	1576	65.40	2054
GenRM	63.50	62.35	70.71	1423	63.60	1879
DPO	–	58.02	63.42	1045	58.00	1733
Res2Sent	–	56.94	60.66	1437	62.60	1936
Sent2Res	–	63.33	70.99	1550	67.10	2063
Sentence (Reinforce++)	**80.72**	**69.83**	72.33	1447	**71.00**	2021
Sentence (PPO)	**80.72**	65.90	**76.92**	1381	69.80	1974

The win rates (WR) and Length-Control win rates (LC) compared to SFT model on AlpacaEval2 and Arena-Hard benchmarks. Results demonstrate the superiority of our method and its favorable trade-off between the quality and response length.