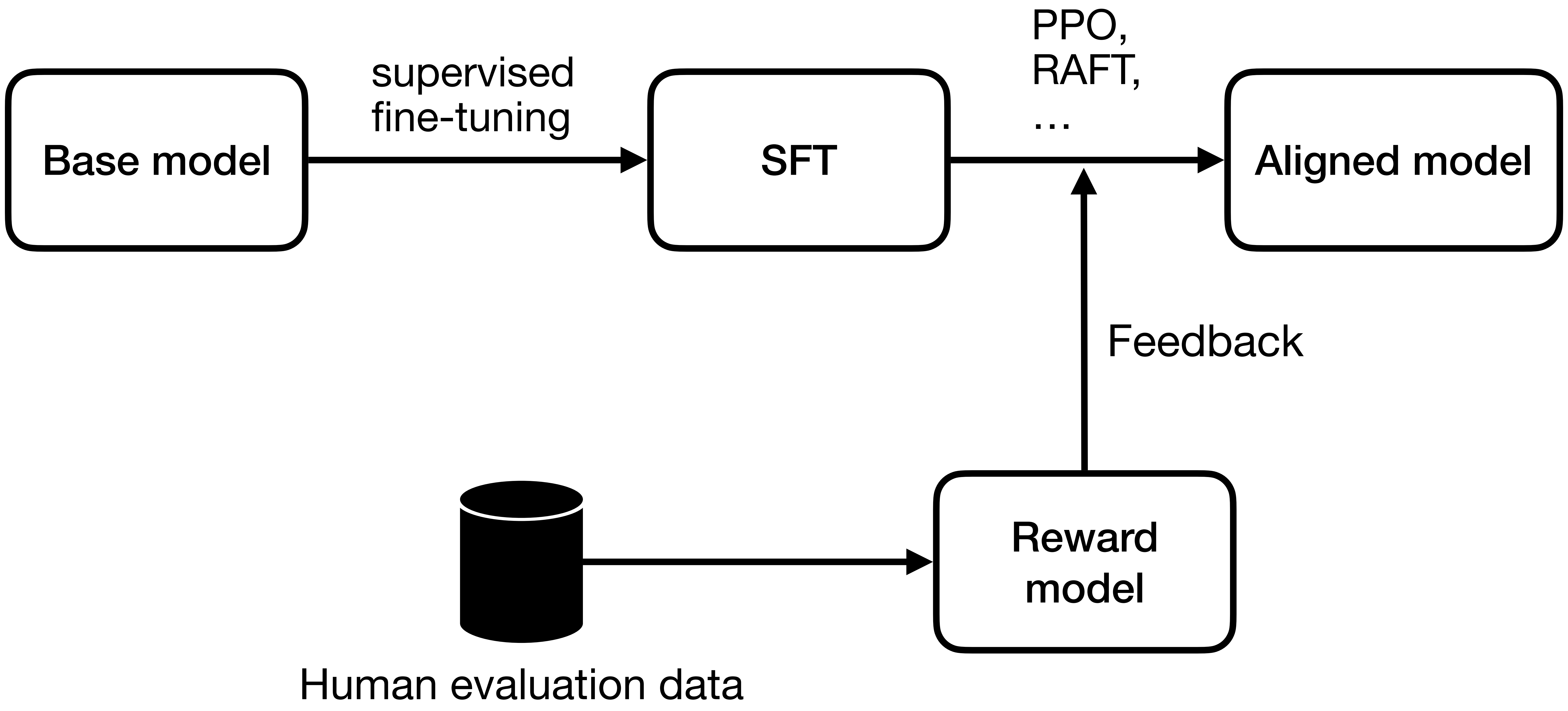
Improving Machine Translation with Human Feedback: An Exploration of Quality Estimation as a Reward Model

Zhiwei He

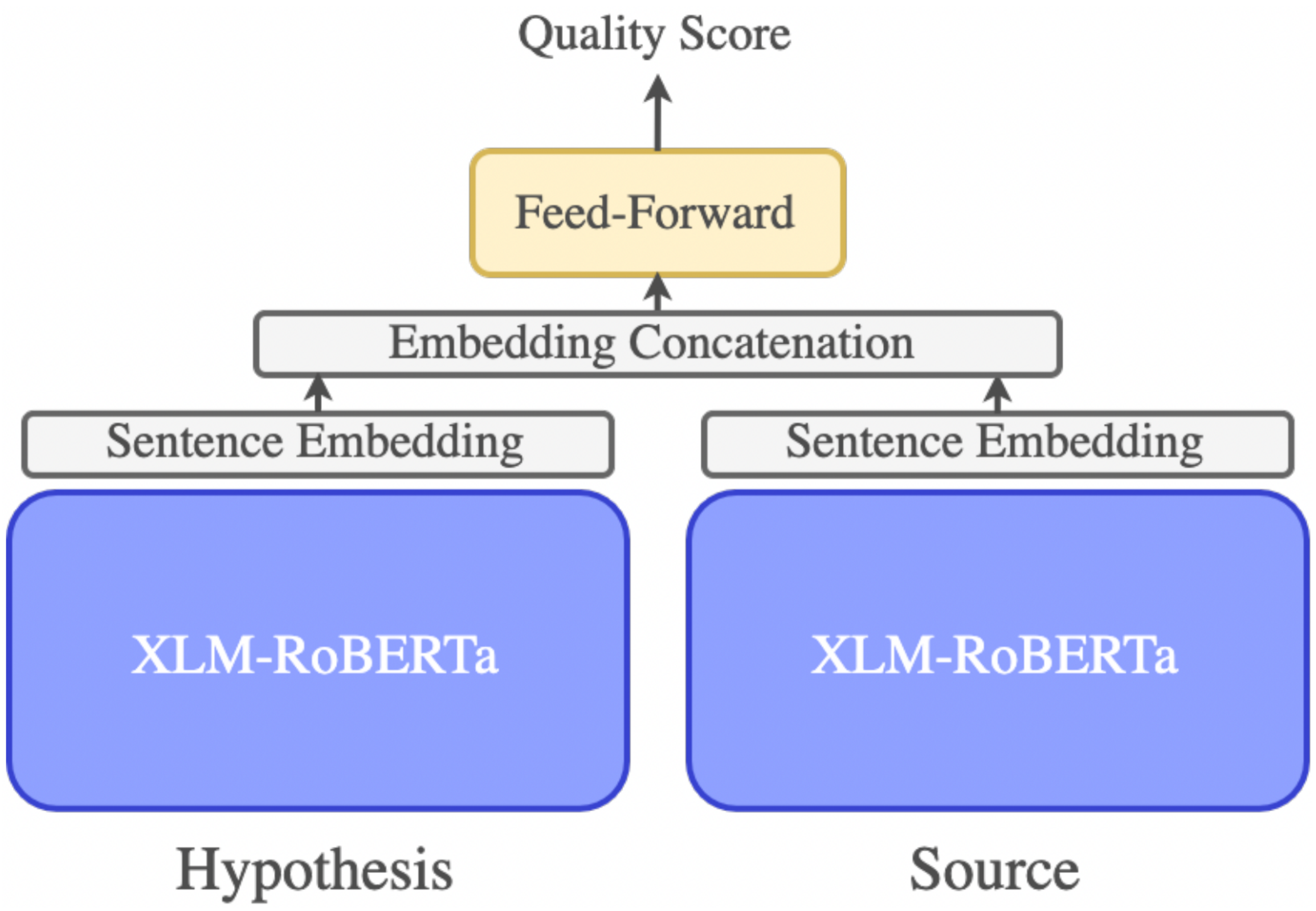
LLMs have already benefited from learning from human feedback



https://arxiv.org/abs/2203.02155

Can MT models learn from human feedback?

Modeling human preference in MT: Quality Estimation (QE)



- A sentence-level QE model can provide a numerical score to indicate the quality of the translation.
- Reference-free

https://aclanthology.org/2021.wmt-1.111/

Can MT models learn from human feedback?

Modeling human preference in MT: Quality Estimation (QE)

Metric	avg rank
METRICX XXL	1.20
COMET-22	1.32
UNITE	1.86
BLEURT-20	1.91
COMET-20	2.36
MATESE	2.57
COMETKIWI*	2.70
MS-COMET-22	2.84
UNITE-SRC*	3.03
YISI-1	3.27
COMET-QE*	3.33
MATESE-QE*	3.85
MEE4	3.87
BERTSCORE	3.88
MS-COMET-QE-22*	4.06
CHRF	4.70
F101SPBLEU	4.97
HWTSC-TEACHER-SIM*	5.17
BLEU	5.31
REUSE*	6.69

Table 1: Official ranking of all primary submissions of the WMT22 Metric Task. The final score is the weighted average ranking over 201 different scenarios. Metrics with * are reference-free metrics.

- Today's most advanced QE models closely match human preferences.
- Can we function them as **reward models** in feedback training?

https://www.statmt.org/wmt22/pdf/2022.wmt-1.2

Feedback Training in MT

Reward rAnked FineTuning (RAFT)

- MT model: $\boxed{M = P(y|x;\theta)}$
- QE-based reward model: $\boxed{r(x,y)}$
- Objective

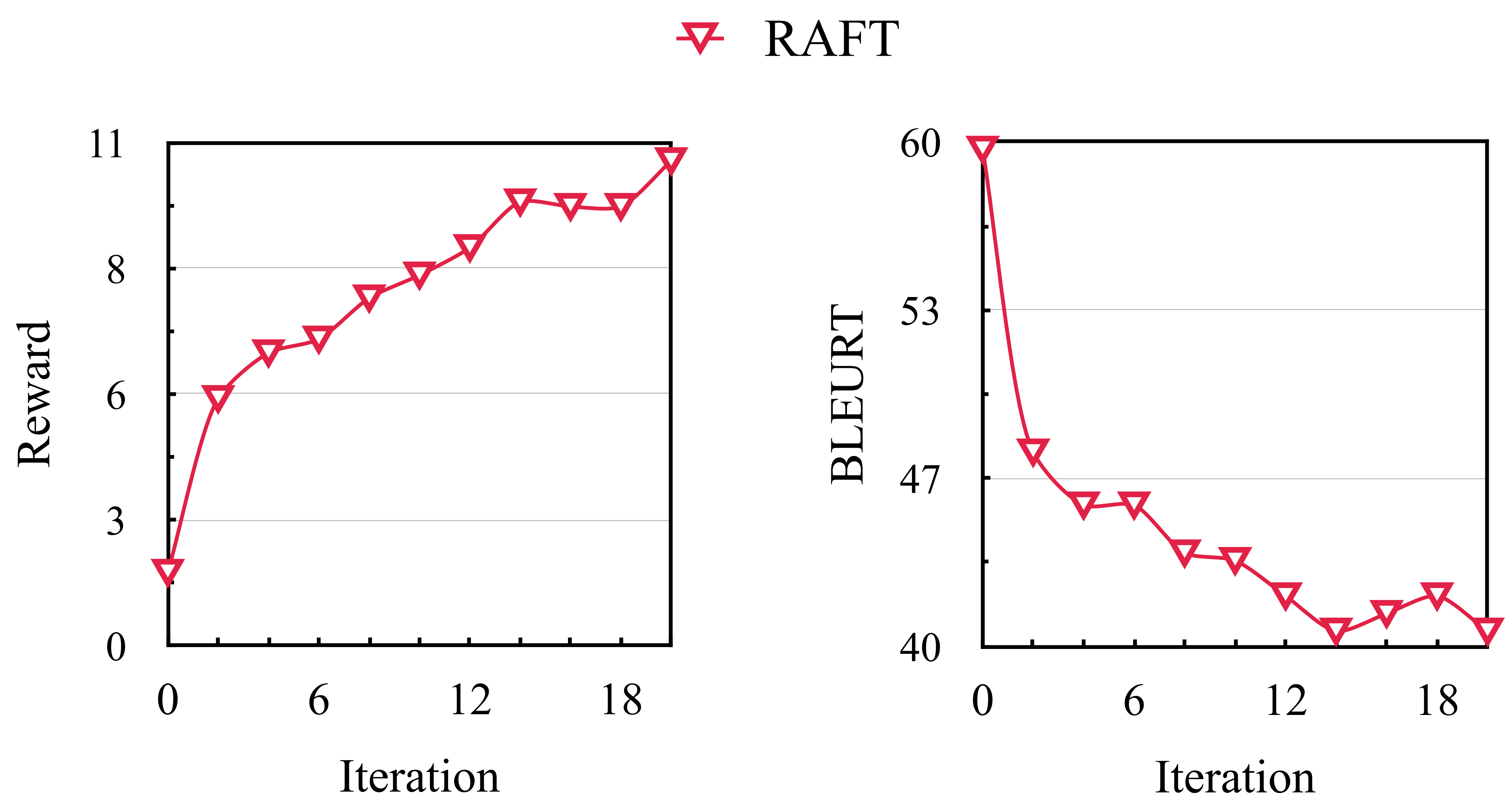
$$\boxed{\max_{\theta} \mathbb{E}_{x\sim\mathcal{D},y\sim P(y|x;\theta)} r(x,y)}$$

Algorithm 1 RAFT

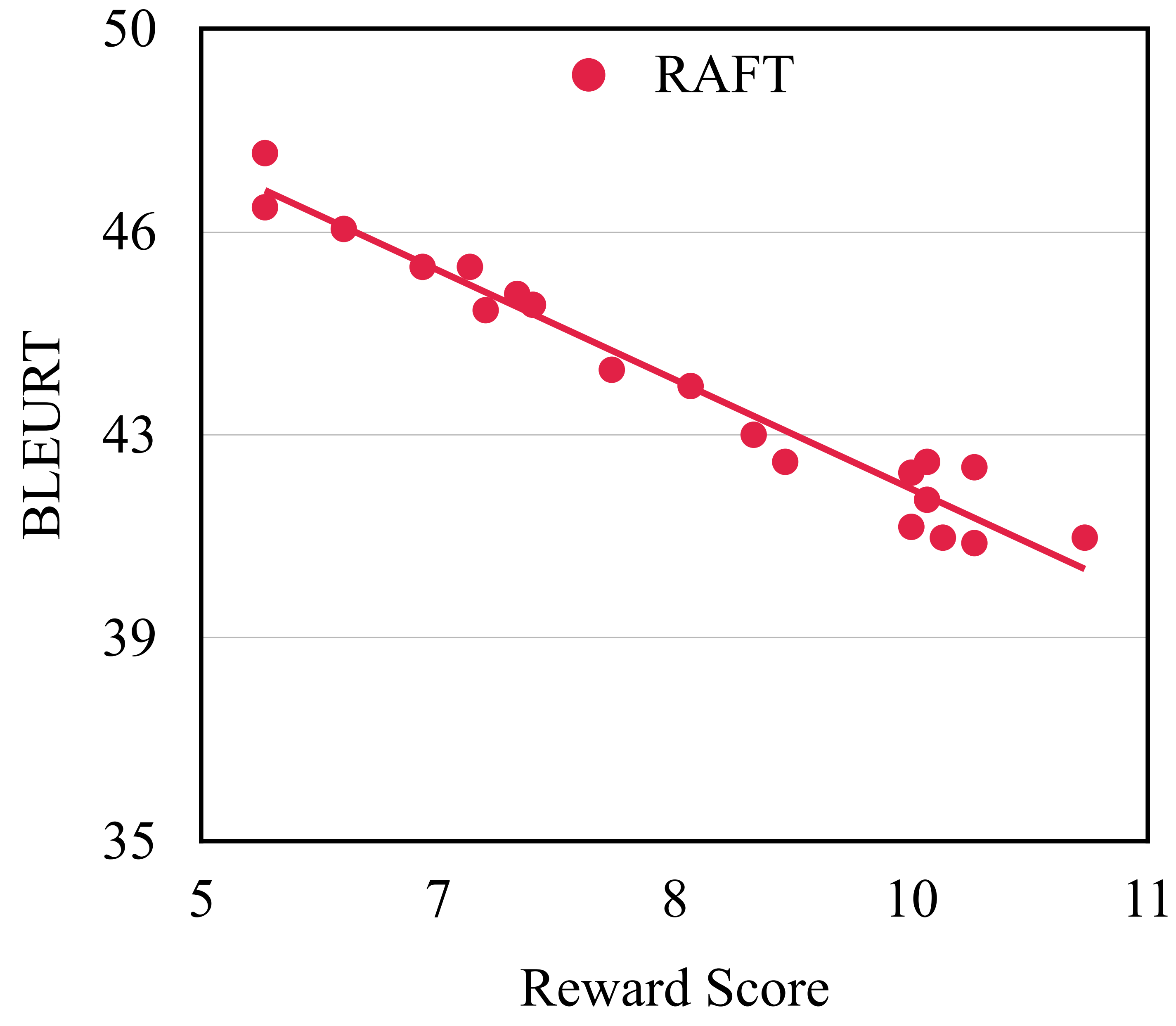
Require: Training set $\mathcal{X}$, reward function $r(x,y)$, initial model $M_0 = P(y|x;\theta_0)$, batch size b, temperature T, the number of candidate k

1: **for** iteration i in $0, 1, \ldots, N-1$ **do**
2: $D_i \leftarrow \text{SampleBatch}(\mathcal{X}, b)$
3: $\mathcal{B} = \emptyset$
4: **for** $x \in D_i$ **do**
5: $y_1, \ldots, y_k \sim P_T(y|x;\theta_i)$
6: $y^* = \arg\max_{y_j \in \{y_1,\ldots,y_k\}} r(x, y_j)$
7: $\mathcal{B} = \mathcal{B} \cup \{(x, y^*)\}$
8: Fine-tune θ_i on $\mathcal{B}$ to obtain $M_{i+1} = P(y|x;\theta_{i+1})$.

Results Not as Expected



As training progresses, reward goes up,
but translation quality goes down.



The two show a negative linear correlation

Why? Overoptimization!

QE (reward) model is not perfect

Error type	Translation	Reward
None	The rule of drinking Red Label Whisky:	2.84
Len-ratio (too long/short translation)	The rule of drinking Red Label Whisky: 1. Always drink responsibly. 2. Never drink alone. 3. Avoid drinking on an empty stomach. 4. Set limits and stick to them. 5. Drink in moderation.	5.60
Off-target (wrong target language)	So trinkt man Red-Label-Whisky:	4.58

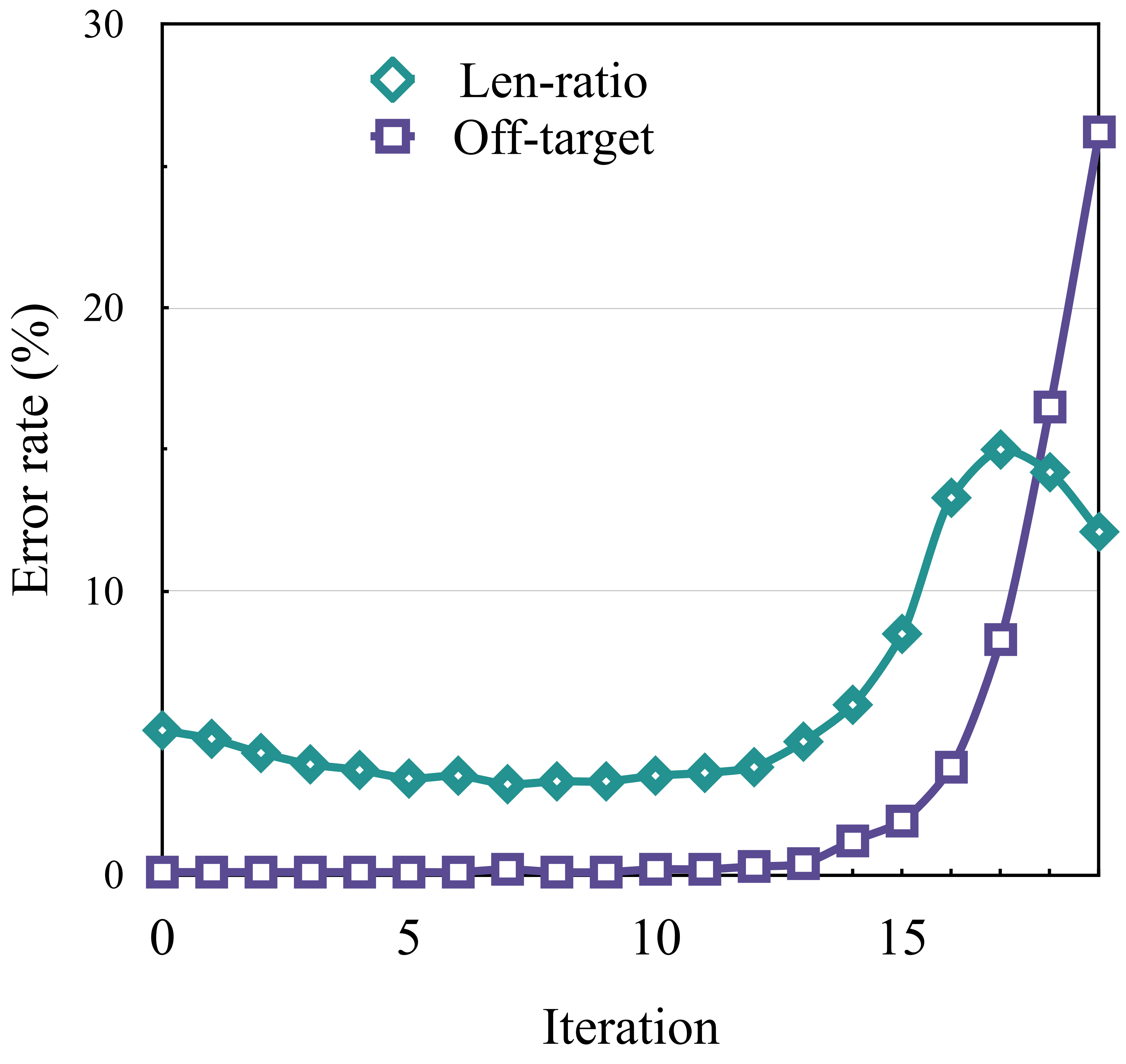
Table 1: A case of Chinese⇒English translation where the QE model (COMET-QE-DA) assigns higher scores to length-ratio and off-target errors than an error-free translation. Error spans are highlighted.

QE model may assign high scores to erroneous translations in some cases.

- The two most common errors
  - Len-ratio error
  - Off-target error

Why? Overoptimization!

Models can quickly capture and learn from these error patterns



☑ Overoptimizing against an imperfect reward model can lead to systems that receive good feedback from the reward model, but not humans.

How to mitigate overoptimization?

Add penalty term in reward

$$r^{+}(x,y) = \begin{cases} r(x,y) - P & \text{if } C(x,y) \\ r(x,y) & \text{otherwise} \end{cases}$$

- C(x, y) = True if (x, y) is a len-ratio or off-target error.
- We refer to this method as RAFT+.

RAFT+ versus RAFT

RAFT+ significantly mitigates overoptimization

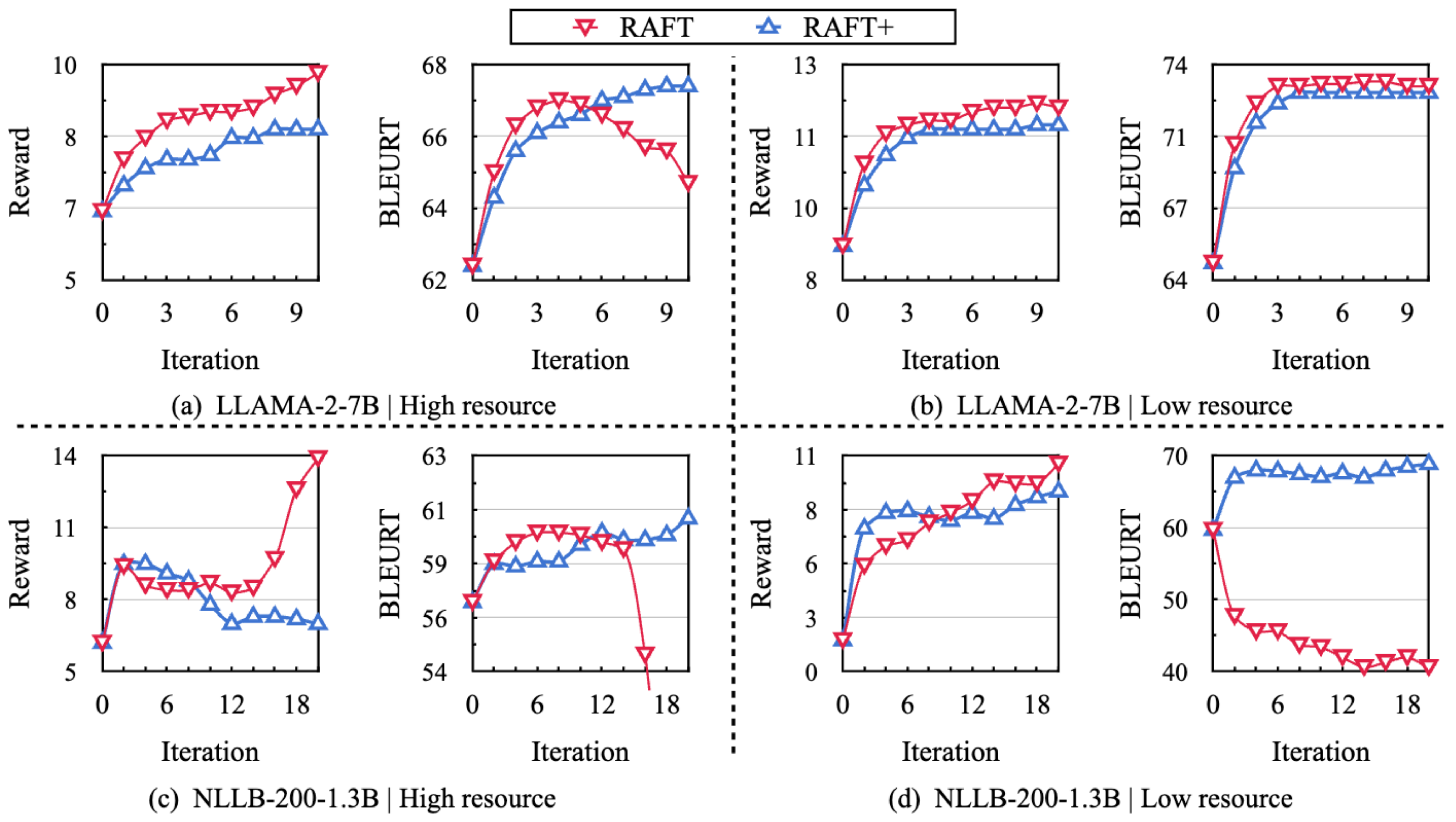
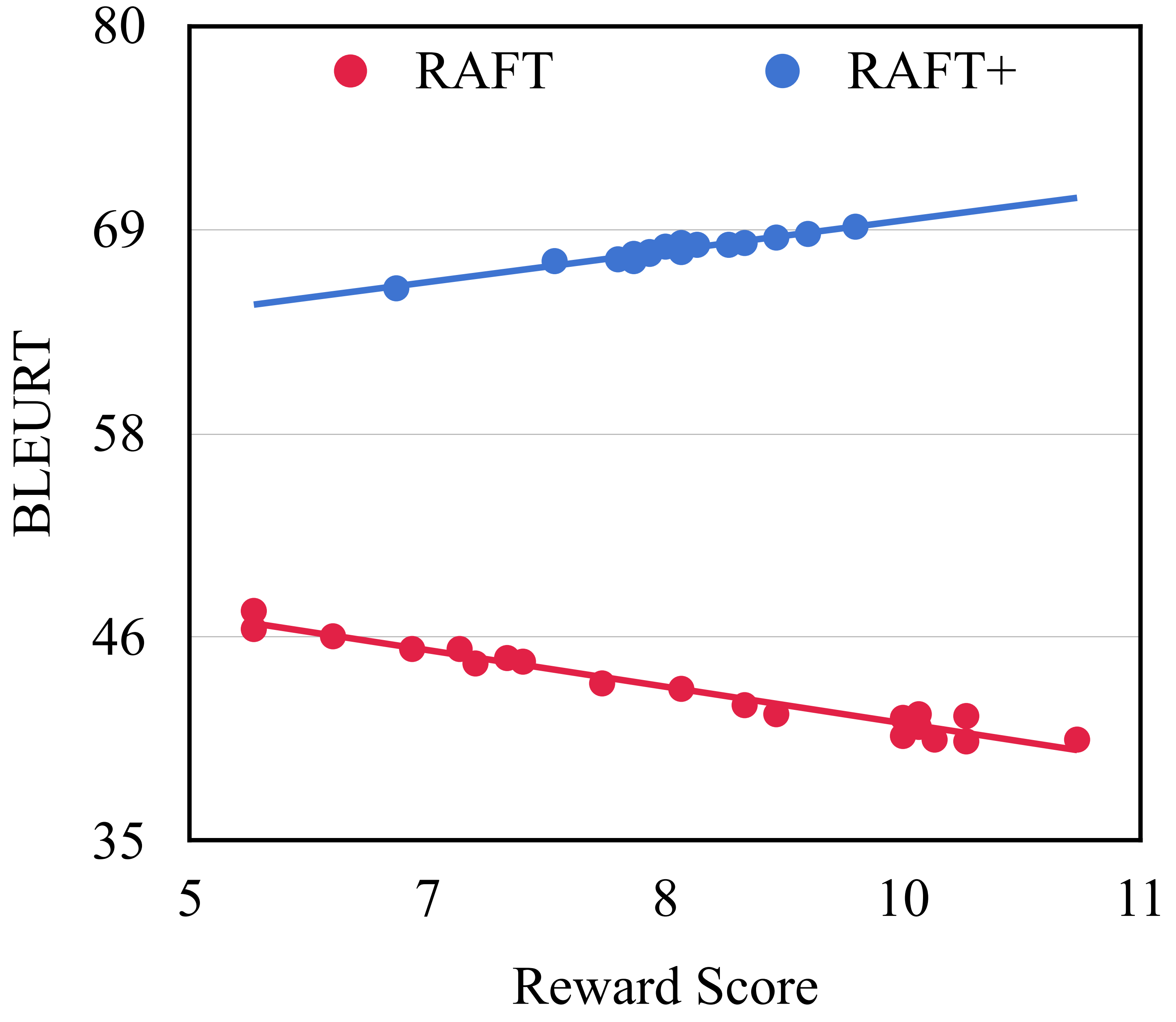


Figure 3: Training curves under various settings. The metrics are average values for all language pairs on the development set. The QE-based reward model is COMET-QE-DA.



Under the RAFT+ algorithm,
the reward score and translation quality
show positive linear correlation.

After addressing overoptimization

Feedback training is very effective, especially in low-resource languages

Method	De⇒En		En⇒De		Zh⇒En		En⇒Zh		**Average**	
	COMET	BLEURT	COMET	BLEURT	COMET	BLEURT	COMET	BLEURT	**COMET**	**BLEURT**
LLAMA-2-7B										
SFT	82.5	70.5	80.7	68.2	76.1	62.3	84.9	69.3	81.0	67.6
REWARD MODEL: COMET-QE-DA										
RAFT	83.7	72.1	82.8	71.1	78.7	65.3	85.9	70.1	82.8$_{\uparrow 1.7}$	69.7$_{\uparrow 2.1}$
RAFT+	83.6	72.1	84.4	73.9	79.0	66.1	85.4	69.3	**83.1**$_{\uparrow 2.1}$	**70.3**$_{\uparrow 2.7}$
REWARD MODEL: COMET-QE-MQM										
RAFT	83.3	72.0	84.8	75.1	77.8	64.3	86.1	70.4	83.0$_{\uparrow 2.0}$	70.5$_{\uparrow 2.9}$
RAFT+	83.7	72.4	85.6	75.7	78.6	65.6	85.8	70.0	**83.4**$_{\uparrow 2.4}$	**70.9**$_{\uparrow 3.3}$
NLLB-200-1.3B										
SFT	70.9	52.5	85.3	74.8	66.0	48.4	83.7	69.1	76.5	61.2
REWARD MODEL: COMET-QE-DA										
RAFT	73.2	52.2	85.8	75.1	67.9	50.5	84.2	68.9	77.8$_{\uparrow 1.3}$	61.7$_{\uparrow 0.5}$
RAFT+	74.2	56.7	85.8	75.2	69.0	52.6	84.0	67.9	**78.2**$_{\uparrow 1.7}$	**63.1**$_{\uparrow 1.9}$
REWARD MODEL: COMET-QE-MQM										
RAFT	82.8	71.3	83.9	73.4	76.1	62.3	84.6	68.6	81.8$_{\uparrow 5.3}$	68.9$_{\uparrow 7.7}$
RAFT+	83.3	71.8	84.6	74.4	76.7	62.9	84.6	68.4	**82.3**$_{\uparrow 5.8}$	**69.4**$_{\uparrow 8.2}$

(a) High-resource language pairs

Method	En⇒Uk		Uk⇒En		Uk⇒Cs		Cs⇒Uk		**Average**	
	COMET	BLEURT	COMET	BLEURT	COMET	BLEURT	COMET	BLEURT	**COMET**	**BLEURT**
LLAMA-2-7B										
SFT	79.2	64.0	76.7	66.0	70.0	53.2	71.2	51.3	74.3	58.6
REWARD MODEL: COMET-QE-DA										
RAFT	82.3	68.0	81.4	71.1	82.5	69.5	84.3	69.9	**82.6**$_{\uparrow 8.3}$	**69.6**$_{\uparrow 11.0}$
RAFT+	82.0	67.8	81.5	71.2	82.2	68.8	84.5	70.1	**82.6**$_{\uparrow 8.3}$	69.5$_{\uparrow 10.9}$
REWARD MODEL: COMET-QE-MQM										
RAFT	80.7	65.5	76.7	66.0	75.7	59.9	75.2	54.8	77.1$_{\uparrow 2.8}$	61.5$_{\uparrow 2.9}$
RAFT+	81.2	67.0	79.2	68.9	77.3	62.3	78.8	60.7	**79.1**$_{\uparrow 4.8}$	**64.8**$_{\uparrow 6.2}$
NLLB-200-1.3B										
SFT	83.1	70.2	71.1	62.7	73.2	61.5	57.3	43.4	71.2	59.4
REWARD MODEL: COMET-QE-DA										
RAFT	85.2	72.5	64.7	33.2	70.5	29.7	73.8	30.1	73.6$_{\uparrow 2.4}$	41.4$_{\downarrow 18.0}$
RAFT+	84.5	71.3	77.7	67.0	83.1	70.3	72.0	55.1	**79.3**$_{\uparrow 8.1}$	**65.9**$_{\uparrow 6.6}$
REWARD MODEL: COMET-QE-MQM										
RAFT	85.8	73.2	67.5	50.0	71.1	41.6	71.1	42.7	73.9$_{\uparrow 2.7}$	51.9$_{\downarrow 7.5}$
RAFT+	84.5	71.8	76.4	66.1	82.1	69.9	71.4	54.5	**78.6**$_{\uparrow 7.4}$	**65.6**$_{\uparrow 6.2}$

(b) Low-resource language pairs

Human Preference Study

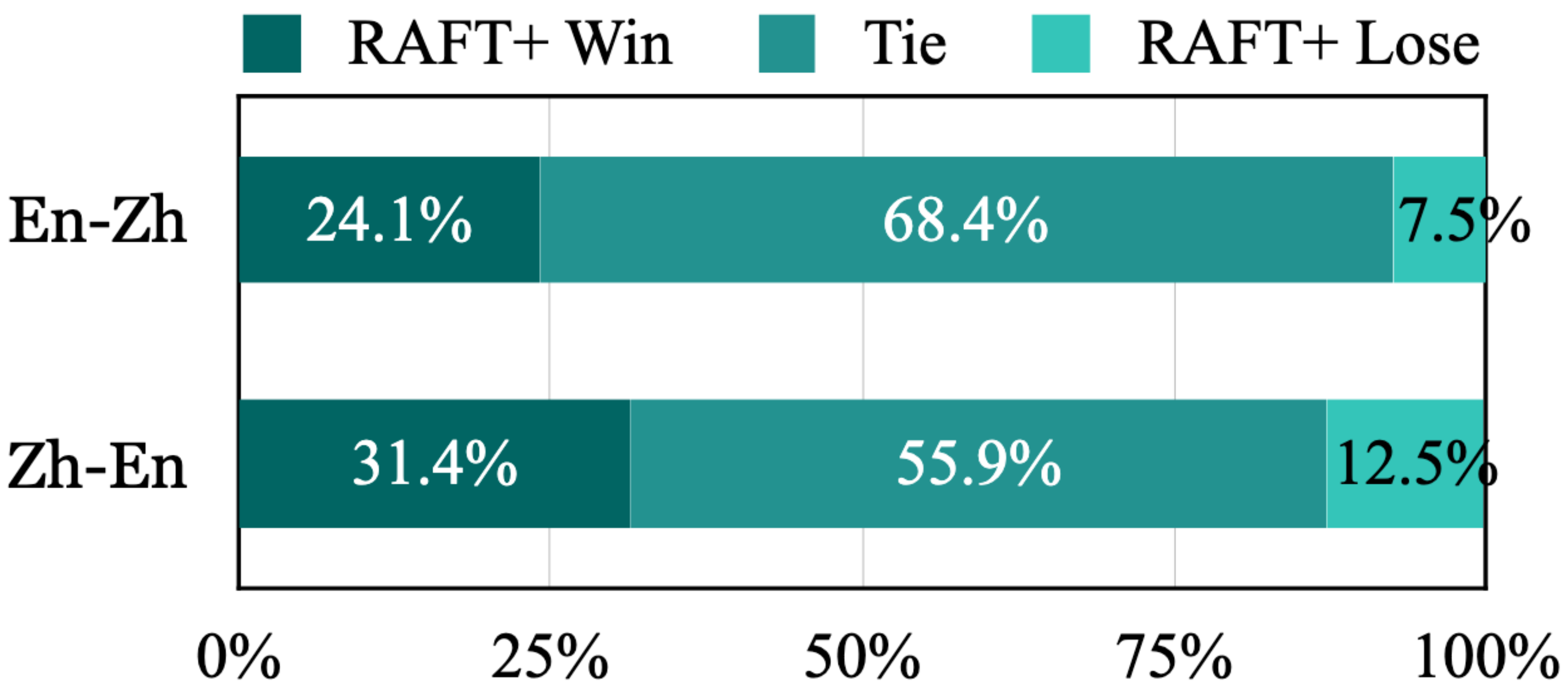


Figure 4: Human preference evaluation, comparing RAFT+ to SFT model on En⇔Zh test sets.

☑ Humans prefer models trained with feedback.

Data Efficiency of Feedback Training

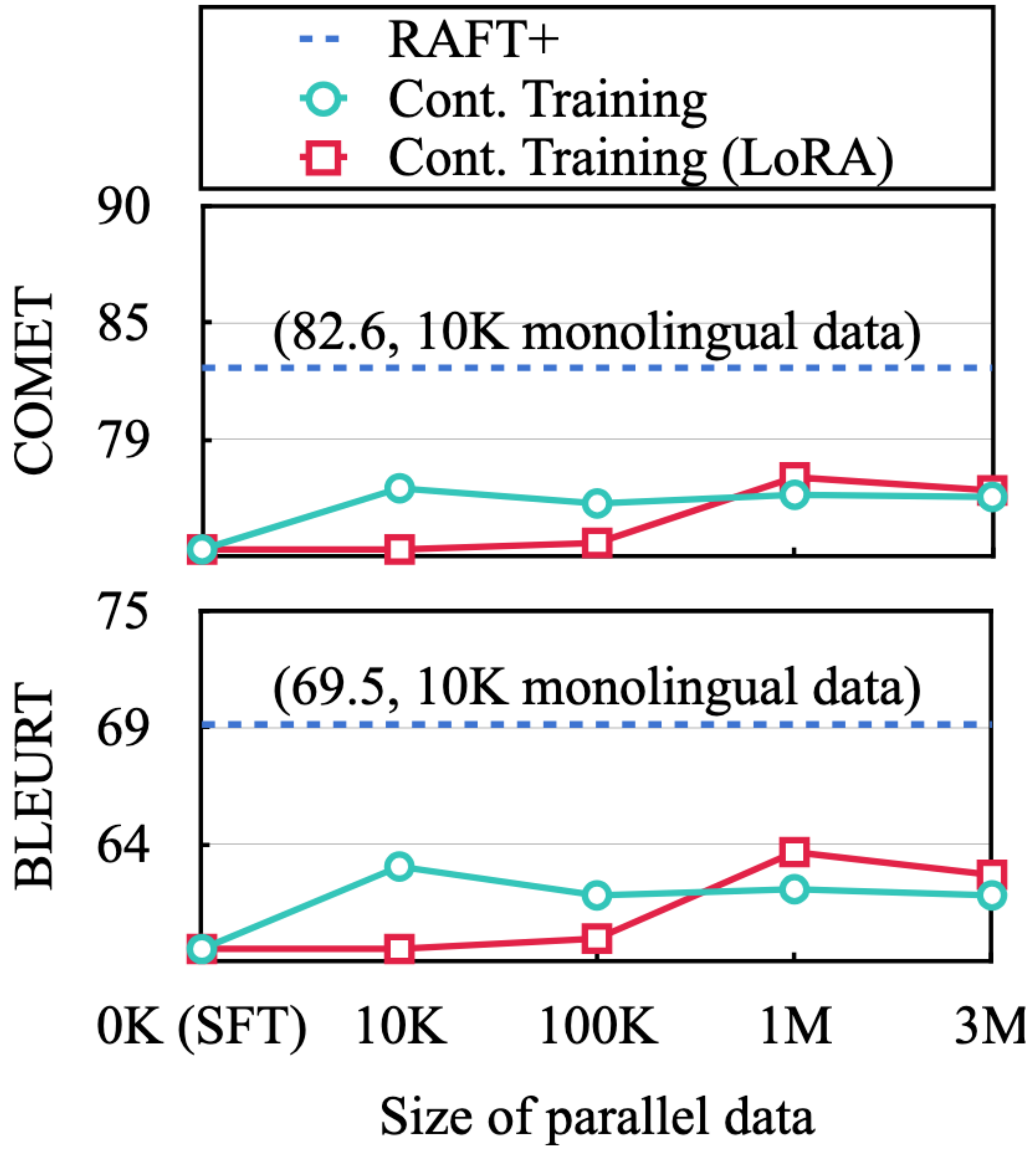


Figure 5: Comparison between RAFT+ and continuous training in the low-resource setting.

☑ Feedback training is data efficient.

- Continuous training with increasing amounts of parallel data fails to yield consistent improvements.
- RAFT+ performs markedly better using merely 10K monolingual data。

Effects of Scaling Model Size and Pretraining

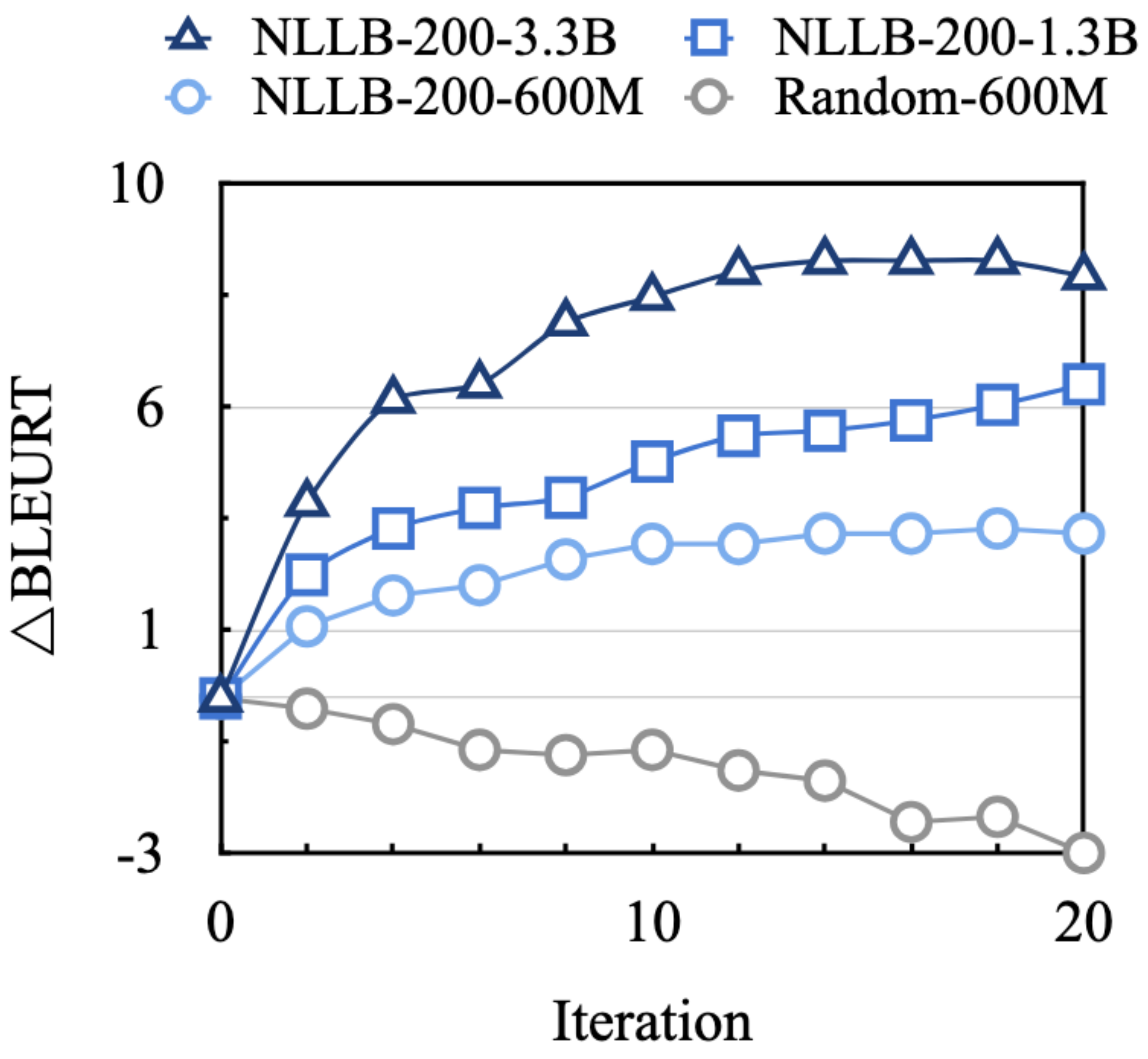


Figure 6: Training curves of RAFT+ (high-resource COMET-QE-MQM) under different base models. We report the change in BLEURT score for each checkpoint relative to the SFT model.

☑ Feedback training performs better on strong base models.

- Feedback training exhibits a more pronounced enhancement with a larger base model size.
- Feedback training is effective only when the base model has undergone pretraining.

Check our paper & code for more details

Paper

Code