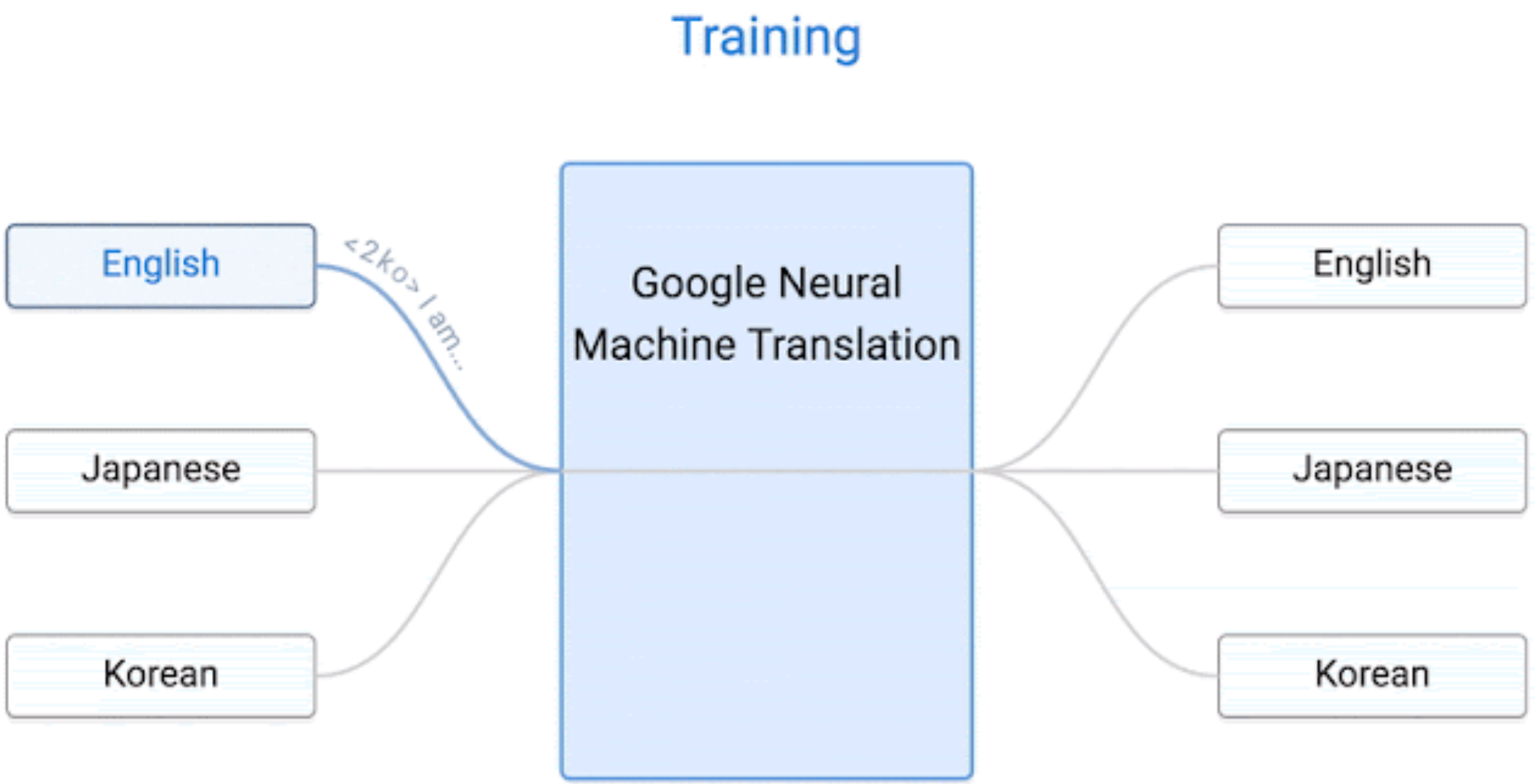
Improving Machine Translation with Human Strategy and Feedback

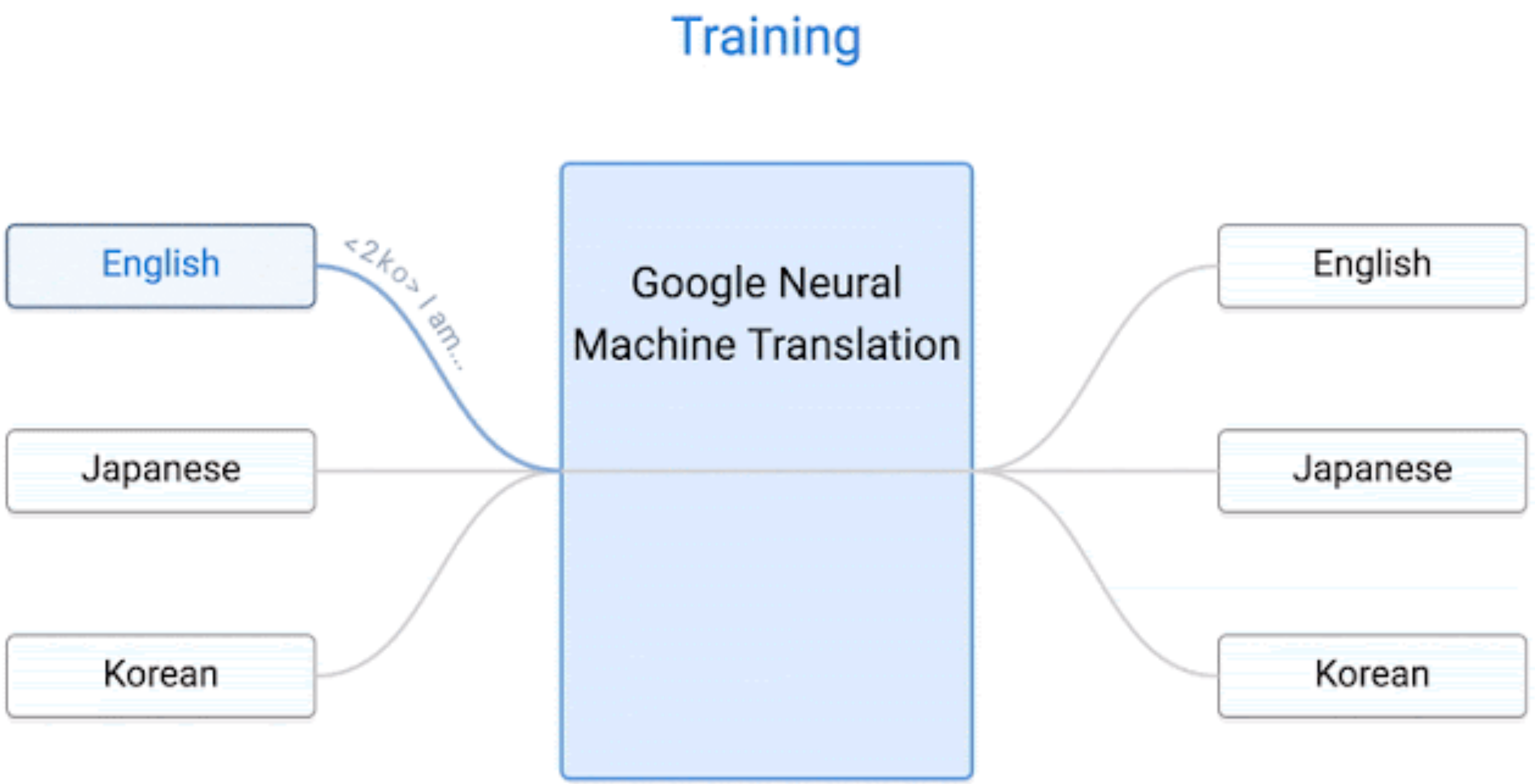
Zhiwei He & Rui Wang
Shanghai Jiao Tong University

The Neural Machine Translation Training Process



https://blog.research.google/2016/11/zero-shot-translation-with-googles.html?m=1

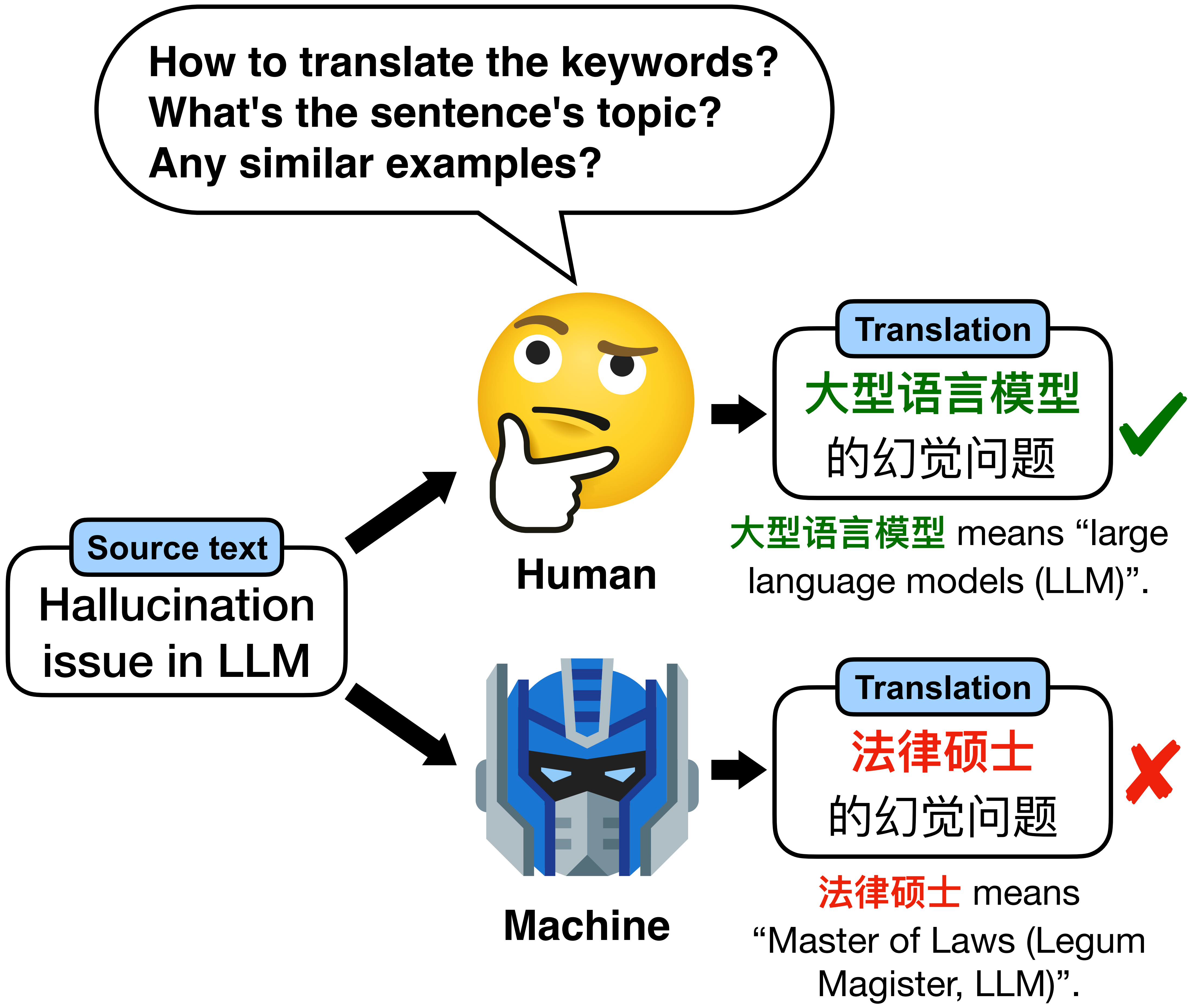
The Neural Machine Translation Training Process



https://blog.research.google/2016/11/zero-shot-translation-with-googles.html?m=1

Two Main Limitations of Current NMT Models

Limitation 1: Lacking Human Translation Strategies



- NMT models are trained to perform source-to-target mapping.
- A human translator can take preparatory steps to ensure high-quality translation.

Large language model (LLM) can adopt many human-like strategies in reasoning and planning tasks

Let's think step by step, …

Chain-of-Thought

Let me do a reflection and think about how to improve my strategy, …

Reflexion

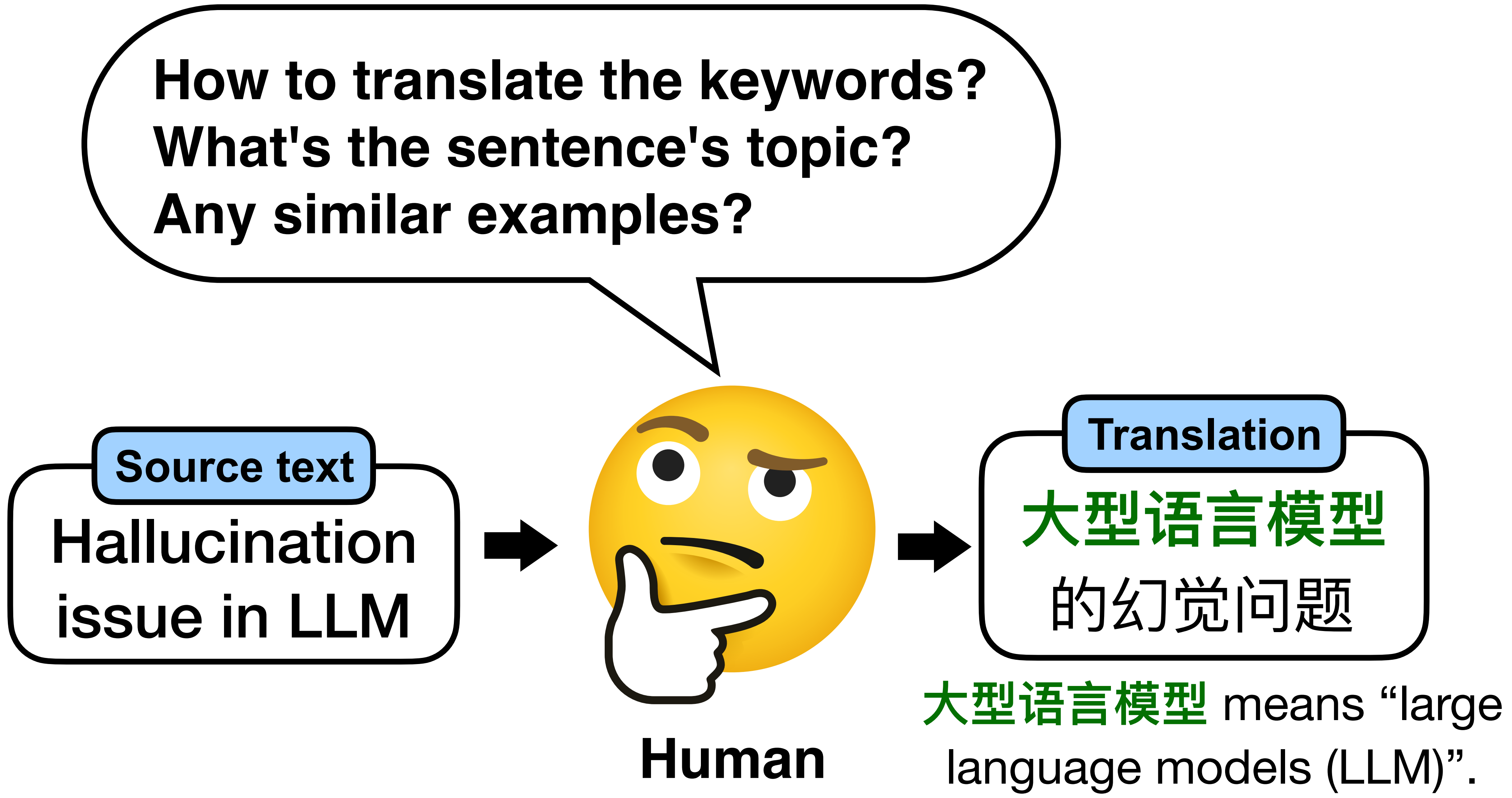
Let's take a step back and generate a more generic question, …

Step-Back prompting

https://arxiv.org/abs/2201.11903
https://arxiv.org/abs/2303.11366
https://arxiv.org/abs/2310.06117

Exploring Human-Like Translation Strategy with LLM

Preparatory steps that a human translator might take



✓ Identify **keywords** and consider how to translate them

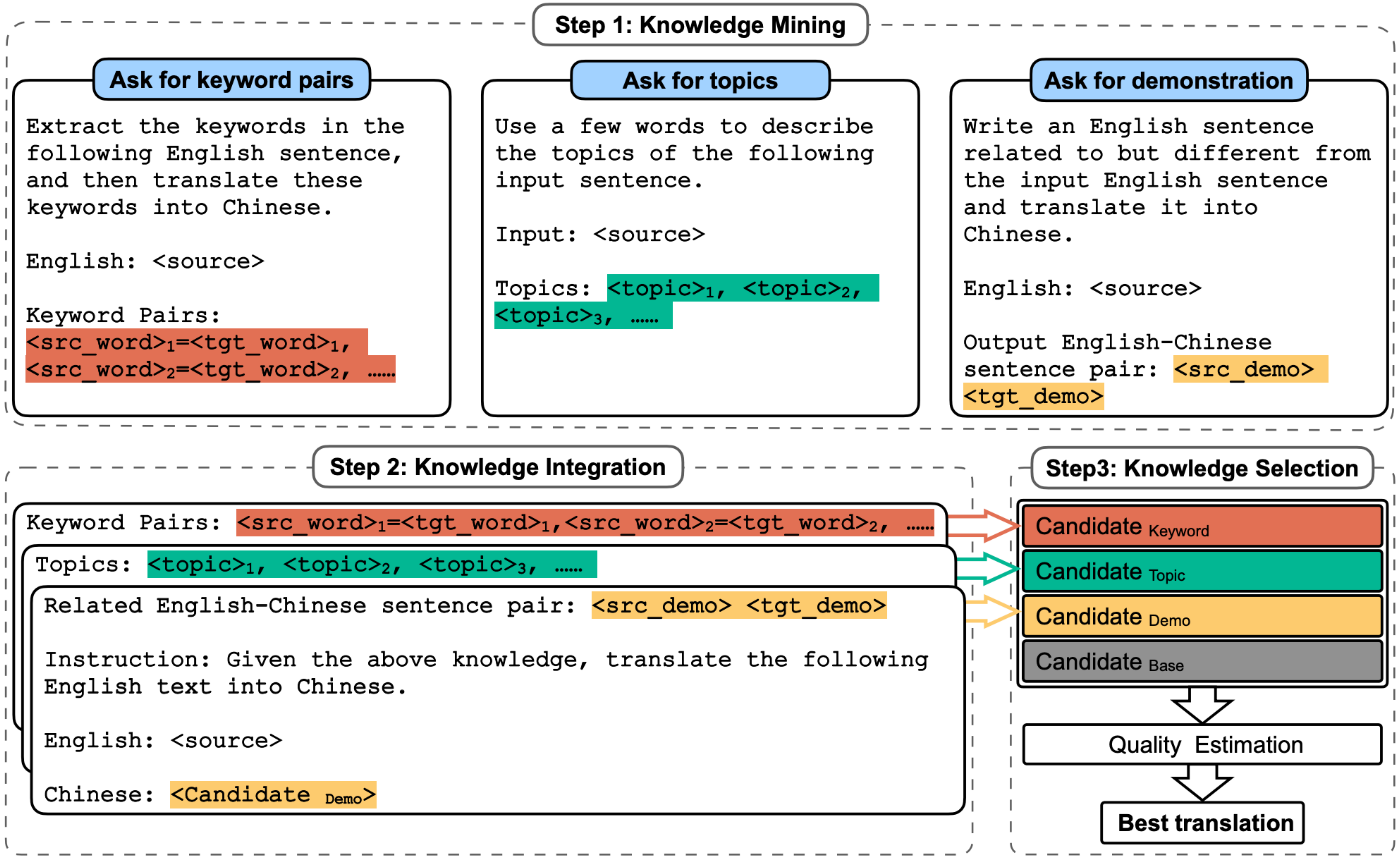
✓ Reflect on what the main **topic** of this text is

✓ Consider how **similar sentences (demonstrations)** are translated.

✓ ……

Exploring Human-Like Translation Strategy with LLM

MAPS: Multi-Aspect Prompting and Selection



Implementation of Knowledge Selection (Reranking Method)

- **LLM-SCQ**: Composing a single choice question (SCQ) that asks the LLM to choose the best candidate on its own.

- **COMET-QE**: A trained QE scorer that assigns a numerical score to each candidate. Selection is based on the highest score.

- **COMET** (oracle): A reference-based scorer that assigns a numerical score to each candidate. It can be considered as the oracle QE method, representing the **upper bound** of selection.

Main Results

Method	En-Zh	Zh-En	En-De	De-En	En-Ja	Ja-En	De-Fr	Fr-De	Cs-Uk	Uk-Cs	En-Hr
`WMT22 Best` | `COMET`											
WMT22 Best	86.8	81.0	87.4	85.0	89.3	81.6	85.7	89.5	91.6	92.2	88.4
`text-davinci-003` | `COMET`											
Baseline	86.2	81.6	85.8	85.2	87.9	81.8	82.8	86.3	88.0	89.2	85.9
5-Shot (Hendy et al.)	87.0	81.1	86.5	85.2	88.2	82.0	83.6	86.6	—	—	—
Rerank LLM-SCQ	86.4	81.7	86.0	85.2	88.0	82.0	83.0	86.4	88.3	89.4	86.3
MAPS LLM-SCQ	86.8	**82.0**	86.4	**85.4**	**88.5**	**82.4**	83.4	**86.9**	**88.8**	**89.9**	**86.5**
Rerank COMET-QE	86.9	82.1	86.4	85.5	88.8	82.3	83.4	86.8	89.4	90.1	87.1
MAPS COMET-QE	**87.6**	**82.6**	**87.2**	**85.7**	**89.5**	**82.9**	**84.1**	**87.5**	**90.1**	**91.1**	**88.1**
⇑ Rerank COMET	87.5	82.6	86.9	85.8	89.3	82.3	83.4	86.8	89.9	90.7	87.7
⇑ MAPS COMET	**88.5**	**83.8**	**88.0**	**86.7**	**90.3**	**82.9**	**84.1**	**87.5**	**90.9**	**92.0**	**89.0**
`text-davinci-003` | `BLEURT`											
Baseline	71.1	69.6	75.6	74.0	66.3	67.8	70.4	77.6	75.0	78.8	75.0
5-Shot (Hendy et al.)	72.2	69.2	76.3	74.5	67.1	68.0	70.9	78.0	—	—	—
Rerank LLM-SCQ	71.4	69.8	75.9	74.1	66.6	68.1	70.6	77.7	75.3	79.0	75.4
MAPS LLM-SCQ	72.1	**70.5**	76.3	74.4	**67.4**	**68.8**	**71.4**	**78.6**	**76.1**	**80.2**	**76.0**
Rerank COMET-QE	71.7	70.1	76.1	74.3	67.3	68.3	71.2	78.1	76.4	79.7	75.9
MAPS COMET-QE	**72.6**	**70.8**	**77.1**	**74.6**	**68.3**	**69.1**	**71.9**	**78.9**	**77.4**	**81.2**	**77.1**
⇑ Rerank COMET	72.4	70.6	76.5	74.6	68.0	68.8	71.8	78.6	76.8	80.2	76.4
⇑ MAPS COMET	**74.0**	**72.1**	**77.8**	**75.7**	**69.4**	**70.9**	**73.6**	**80.2**	**78.3**	**82.1**	**77.9**
`Alpaca` | `COMET`											
Baseline	58.9	73.1	75.5	81.9	56.6	71.8	71.7	75.4	74.1	71.1	65.9
Rerank COMET-QE	66.2	74.9	78.5	82.6	64.7	73.7	74.5	78.2	78.1	76.3	70.5
MAPS COMET-QE	**69.0**	**76.0**	**79.7**	**83.3**	**66.9**	**74.7**	**75.9**	**79.1**	**80.8**	**78.5**	**72.3**
`Alpaca` | `BLEURT`											
Baseline	42.3	58.0	62.2	69.8	31.4	55.4	52.2	63.4	52.4	54.3	53.2
Rerank COMET-QE	47.5	59.5	64.7	70.4	36.2	56.7	55.0	66.0	55.2	59.0	56.0
MAPS COMET-QE	**50.6**	**60.6**	**66.3**	**71.1**	**38.2**	**57.7**	**56.6**	**66.8**	**59.5**	**61.2**	**57.2**
`Vicuna` | `COMET`											
Baseline	81.3	78.4	79.8	82.9	82.3	77.3	75.5	77.1	74.9	72.7	69.3
Rerank COMET-QE	83.6	79.3	81.8	83.6	85.2	78.8	77.8	79.6	79.9	77.7	74.2
MAPS COMET-QE	**84.5**	**80.2**	**82.7**	**84.1**	**86.5**	**79.7**	**79.2**	**81.1**	**81.8**	**80.1**	**76.0**
`Vicuna` | `BLEURT`											
Baseline	64.9	65.3	67.4	71.0	58.7	62.8	58.8	66.0	57.8	56.6	57.7
Rerank COMET-QE	66.7	66.0	69.2	71.8	61.6	64.0	61.2	68.2	61.8	61.2	60.5
MAPS COMET-QE	**67.8**	**66.9**	**70.0**	**72.4**	**63.0**	**64.8**	**62.5**	**69.3**	**64.0**	**64.3**	**63.4**

- The effectiveness of MAPS has been validated across a wide range of settings.

✓ Across **11** language pairs, **3** LLMs, and **2** metrics, MAPS consistently boost translation.

✓ Equipped with MAPS, `text-davinci-003` surpasses the best submissions in WMT22 in 5 out of the 11 translation directions.

Main Results

Method	En-Zh	Zh-En	En-De	De-En	En-Ja	Ja-En	De-Fr	Fr-De	Cs-Uk	Uk-Cs	En-Hr
WMT22 Best | COMET											
WMT22 Best	86.8	81.0	87.4	85.0	89.3	81.6	85.7	89.5	91.6	92.2	88.4
text-davinci-003 | COMET											
Baseline	86.2	81.6	85.8	85.2	87.9	81.8	82.8	86.3	88.0	89.2	85.9
5-Shot (Hendy et al.)	87.0	81.1	86.5	85.2	88.2	82.0	83.6	86.6	—	—	—
Rerank LLM-SCQ	86.4	81.7	86.0	85.2	88.0	82.0	83.0	86.4	88.3	89.4	86.3
MAPS LLM-SCQ	86.8	**82.0**	86.4	**85.4**	**88.5**	**82.4**	83.4	**86.9**	**88.8**	**89.9**	**86.5**
Rerank COMET-QE	86.9	82.1	86.4	85.5	88.8	82.3	83.4	86.8	89.4	90.1	87.1
MAPS COMET-QE	**87.6**	**82.6**	**87.2**	**85.7**	**89.5**	**82.9**	**84.1**	**87.5**	**90.1**	**91.1**	**88.1**
⇑ Rerank COMET	87.5	82.6	86.9	85.8	89.3	82.3	83.4	86.8	89.9	90.7	87.7
⇑ MAPS COMET	**88.5**	**83.8**	**88.0**	**86.7**	**90.3**	**82.9**	**84.1**	**87.5**	**90.9**	**92.0**	**89.0**
text-davinci-003 | BLEURT											
Baseline	71.1	69.6	75.6	74.0	66.3	67.8	70.4	77.6	75.0	78.8	75.0
5-Shot (Hendy et al.)	72.2	69.2	76.3	74.5	67.1	68.0	70.9	78.0	—	—	—
Rerank LLM-SCQ	71.4	69.8	75.9	74.1	66.6	68.1	70.6	77.7	75.3	79.0	75.4
MAPS LLM-SCQ	72.1	**70.5**	76.3	74.4	**67.4**	**68.8**	**71.4**	**78.6**	**76.1**	**80.2**	**76.0**
Rerank COMET-QE	71.7	70.1	76.1	74.3	67.3	68.3	71.2	78.1	76.4	79.7	75.9
MAPS COMET-QE	**72.6**	**70.8**	**77.1**	**74.6**	**68.3**	**69.1**	**71.9**	**78.9**	**77.4**	**81.2**	**77.1**
⇑ Rerank COMET	72.4	70.6	76.5	74.6	68.0	68.8	71.8	78.6	76.8	80.2	76.4
⇑ MAPS COMET	**74.0**	**72.1**	**77.8**	**75.7**	**69.4**	**70.9**	**73.6**	**80.2**	**78.3**	**82.1**	**77.9**
Alpaca | COMET											
Baseline	58.9	73.1	75.5	81.9	56.6	71.8	71.7	75.4	74.1	71.1	65.9
Rerank COMET-QE	66.2	74.9	78.5	82.6	64.7	73.7	74.5	78.2	78.1	76.3	70.5
MAPS COMET-QE	**69.0**	**76.0**	**79.7**	**83.3**	**66.9**	**74.7**	**75.9**	**79.1**	**80.8**	**78.5**	**72.3**
Alpaca | BLEURT											
Baseline	42.3	58.0	62.2	69.8	31.4	55.4	52.2	63.4	52.4	54.3	53.2
Rerank COMET-QE	47.5	59.5	64.7	70.4	36.2	56.7	55.0	66.0	55.2	59.0	56.0
MAPS COMET-QE	**50.6**	**60.6**	**66.3**	**71.1**	**38.2**	**57.7**	**56.6**	**66.8**	**59.5**	**61.2**	**57.2**
Vicuna | COMET											
Baseline	81.3	78.4	79.8	82.9	82.3	77.3	75.5	77.1	74.9	72.7	69.3
Rerank COMET-QE	83.6	79.3	81.8	83.6	85.2	78.8	77.8	79.6	79.9	77.7	74.2
MAPS COMET-QE	**84.5**	**80.2**	**82.7**	**84.1**	**86.5**	**79.7**	**79.2**	**81.1**	**81.8**	**80.1**	**76.0**
Vicuna | BLEURT											
Baseline	64.9	65.3	67.4	71.0	58.7	62.8	58.8	66.0	57.8	56.6	57.7
Rerank COMET-QE	66.7	66.0	69.2	71.8	61.6	64.0	61.2	68.2	61.8	61.2	60.5
MAPS COMET-QE	**67.8**	**66.9**	**70.0**	**72.4**	**63.0**	**64.8**	**62.5**	**69.3**	**64.0**	**64.3**	**63.4**

- The effectiveness of MAPS has been validated across a wide range of settings.

✓ Across **11** language pairs, **3** LLMs, and **2** metrics, MAPS consistently boost translation.

✓ Equipped with MAPS, `text-davinci-003` surpasses the best submissions in WMT22 in 5 out of the 11 translation directions.

Main Results

Method	En-Zh	Zh-En	En-De	De-En	En-Ja	Ja-En	De-Fr	Fr-De	Cs-Uk	Uk-Cs	En-Hr
`WMT22 Best | COMET`											
WMT22 Best	86.8	81.0	87.4	85.0	89.3	81.6	85.7	89.5	91.6	92.2	88.4
`text-davinci-003 | COMET`											
Baseline	86.2	81.6	85.8	85.2	87.9	81.8	82.8	86.3	88.0	89.2	85.9
5-Shot (Hendy et al.)	87.0	81.1	86.5	85.2	88.2	82.0	83.6	86.6	—	—	—
Rerank LLM-SCQ	86.4	81.7	86.0	85.2	88.0	82.0	83.0	86.4	88.3	89.4	86.3
MAPS LLM-SCQ	86.8	**82.0**	86.4	**85.4**	**88.5**	**82.4**	83.4	**86.9**	**88.8**	**89.9**	**86.5**
Rerank COMET-QE	86.9	82.1	86.4	85.5	88.8	82.3	83.4	86.8	89.4	90.1	87.1
MAPS COMET-QE	**87.6**	**82.6**	**87.2**	**85.7**	**89.5**	**82.9**	**84.1**	**87.5**	**90.1**	**91.1**	**88.1**
⇑ Rerank COMET	87.5	82.6	86.9	85.8	89.3	82.3	83.4	86.8	89.9	90.7	87.7
⇑ MAPS COMET	**88.5**	**83.8**	**88.0**	**86.7**	**90.3**	**82.9**	**84.1**	**87.5**	**90.9**	**92.0**	**89.0**
`text-davinci-003 | BLEURT`											
Baseline	71.1	69.6	75.6	74.0	66.3	67.8	70.4	77.6	75.0	78.8	75.0
5-Shot (Hendy et al.)	72.2	69.2	76.3	74.5	67.1	68.0	70.9	78.0	—	—	—
Rerank LLM-SCQ	71.4	69.8	75.9	74.1	66.6	68.1	70.6	77.7	75.3	79.0	75.4
MAPS LLM-SCQ	72.1	**70.5**	76.3	74.4	**67.4**	**68.8**	**71.4**	**78.6**	**76.1**	**80.2**	**76.0**
Rerank COMET-QE	71.7	70.1	76.1	74.3	67.3	68.3	71.2	78.1	76.4	79.7	75.9
MAPS COMET-QE	**72.6**	**70.8**	**77.1**	**74.6**	**68.3**	**69.1**	**71.9**	**78.9**	**77.4**	**81.2**	**77.1**
⇑ Rerank COMET	72.4	70.6	76.5	74.6	68.0	68.8	71.8	78.6	76.8	80.2	76.4
⇑ MAPS COMET	**74.0**	**72.1**	**77.8**	**75.7**	**69.4**	**70.9**	**73.6**	**80.2**	**78.3**	**82.1**	**77.9**
`Alpaca | COMET`											
Baseline	58.9	73.1	75.5	81.9	56.6	71.8	71.7	75.4	74.1	71.1	65.9
Rerank COMET-QE	66.2	74.9	78.5	82.6	64.7	73.7	74.5	78.2	78.1	76.3	70.5
MAPS COMET-QE	**69.0**	**76.0**	**79.7**	**83.3**	**66.9**	**74.7**	**75.9**	**79.1**	**80.8**	**78.5**	**72.3**
`Alpaca | BLEURT`											
Baseline	42.3	58.0	62.2	69.8	31.4	55.4	52.2	63.4	52.4	54.3	53.2
Rerank COMET-QE	47.5	59.5	64.7	70.4	36.2	56.7	55.0	66.0	55.2	59.0	56.0
MAPS COMET-QE	**50.6**	**60.6**	**66.3**	**71.1**	**38.2**	**57.7**	**56.6**	**66.8**	**59.5**	**61.2**	**57.2**
`Vicuna | COMET`											
Baseline	81.3	78.4	79.8	82.9	82.3	77.3	75.5	77.1	74.9	72.7	69.3
Rerank COMET-QE	83.6	79.3	81.8	83.6	85.2	78.8	77.8	79.6	79.9	77.7	74.2
MAPS COMET-QE	**84.5**	**80.2**	**82.7**	**84.1**	**86.5**	**79.7**	**79.2**	**81.1**	**81.8**	**80.1**	**76.0**
`Vicuna | BLEURT`											
Baseline	64.9	65.3	67.4	71.0	58.7	62.8	58.8	66.0	57.8	56.6	57.7
Rerank COMET-QE	66.7	66.0	69.2	71.8	61.6	64.0	61.2	68.2	61.8	61.2	60.5
MAPS COMET-QE	**67.8**	**66.9**	**70.0**	**72.4**	**63.0**	**64.8**	**62.5**	**69.3**	**64.0**	**64.3**	**63.4**

- Using the same knowledge selection method, **MAPS** outperforms **Rerank** consistently.
- This indicates that the improvements brought by MAPS stem from three types of translation-related knowledge:
  - ✓ keywords
  - ✓ topics
  - ✓ relevant demonstrations.

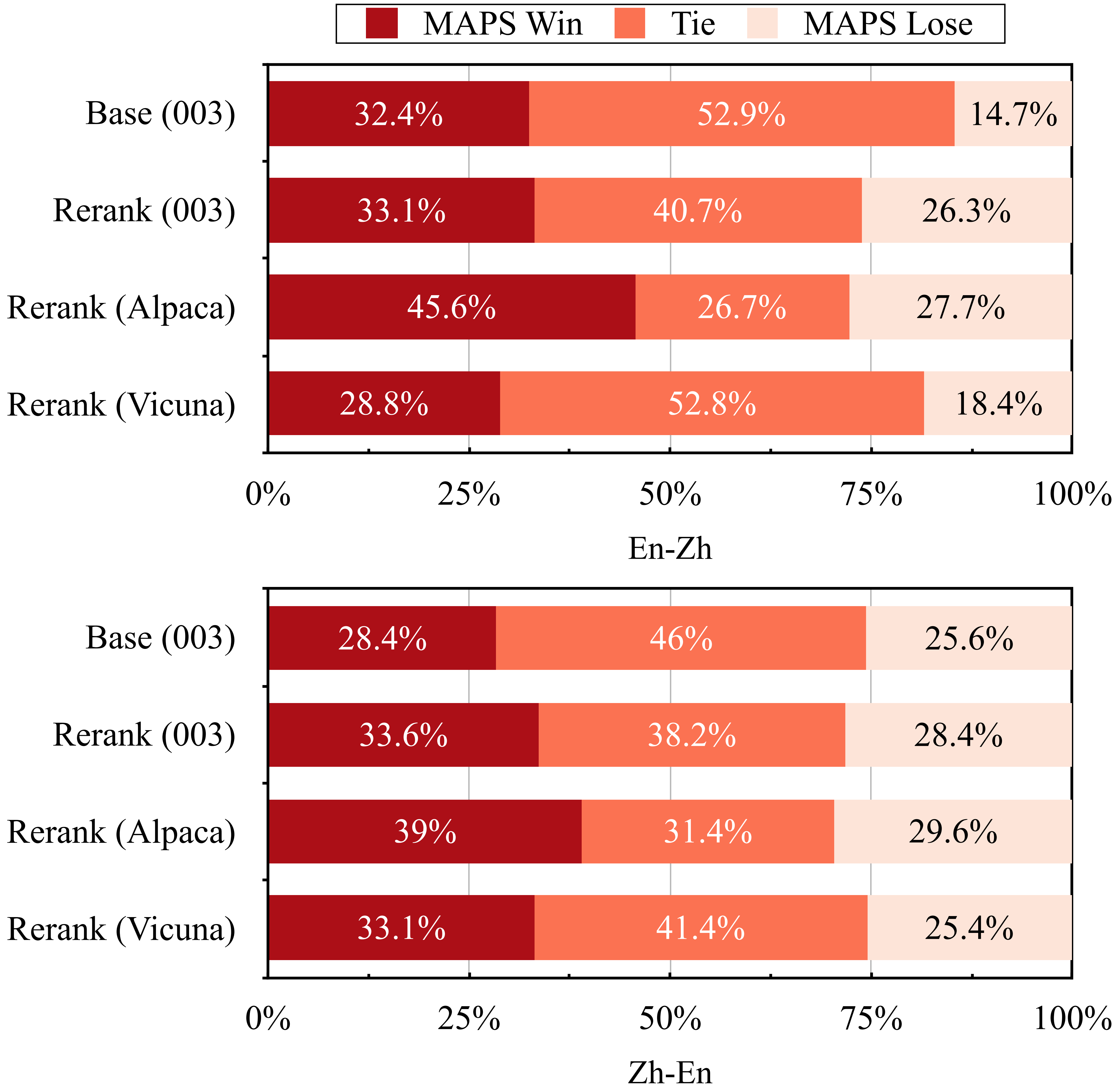
Main Results

Method	En-Zh	Zh-En	En-De	De-En	En-Ja	Ja-En	De-Fr	Fr-De	Cs-Uk	Uk-Cs	En-Hr
`WMT22 Best | COMET`											
WMT22 Best	86.8	81.0	87.4	85.0	89.3	81.6	85.7	89.5	91.6	92.2	88.4
`text-davinci-003 | COMET`											
Baseline	86.2	81.6	85.8	85.2	87.9	81.8	82.8	86.3	88.0	89.2	85.9
5-Shot (Hendy et al.)	87.0	81.1	86.5	85.2	88.2	82.0	83.6	86.6	—	—	—
Rerank LLM-SCQ	86.4	81.7	86.0	85.2	88.0	82.0	83.0	86.4	88.3	89.4	86.3
MAPS LLM-SCQ	86.8	**82.0**	86.4	**85.4**	**88.5**	**82.4**	83.4	**86.9**	**88.8**	**89.9**	**86.5**
Rerank COMET-QE	86.9	82.1	86.4	85.5	88.8	82.3	83.4	86.8	89.4	90.1	87.1
MAPS COMET-QE	**87.6**	**82.6**	**87.2**	**85.7**	**89.5**	**82.9**	**84.1**	**87.5**	**90.1**	**91.1**	**88.1**
⇑ **Rerank** COMET	87.5	82.6	86.9	85.8	89.3	82.3	83.4	86.8	89.9	90.7	87.7
⇑ **MAPS** COMET	**88.5**	**83.8**	**88.0**	**86.7**	**90.3**	**82.9**	**84.1**	**87.5**	**90.9**	**92.0**	**89.0**
`text-davinci-003 | BLEURT`											
Baseline	71.1	69.6	75.6	74.0	66.3	67.8	70.4	77.6	75.0	78.8	75.0
5-Shot (Hendy et al.)	72.2	69.2	76.3	74.5	67.1	68.0	70.9	78.0	—	—	—
Rerank LLM-SCQ	71.4	69.8	75.9	74.1	66.6	68.1	70.6	77.7	75.3	79.0	75.4
MAPS LLM-SCQ	72.1	**70.5**	76.3	74.4	**67.4**	**68.8**	**71.4**	**78.6**	**76.1**	**80.2**	**76.0**
Rerank COMET-QE	71.7	70.1	76.1	74.3	67.3	68.3	71.2	78.1	76.4	79.7	75.9
MAPS COMET-QE	**72.6**	**70.8**	**77.1**	**74.6**	**68.3**	**69.1**	**71.9**	**78.9**	**77.4**	**81.2**	**77.1**
⇑ **Rerank** COMET	72.4	70.6	76.5	74.6	68.0	68.8	71.8	78.6	76.8	80.2	76.4
⇑ **MAPS** COMET	**74.0**	**72.1**	**77.8**	**75.7**	**69.4**	**70.9**	**73.6**	**80.2**	**78.3**	**82.1**	**77.9**
`Alpaca | COMET`											
Baseline	58.9	73.1	75.5	81.9	56.6	71.8	71.7	75.4	74.1	71.1	65.9
Rerank COMET-QE	66.2	74.9	78.5	82.6	64.7	73.7	74.5	78.2	78.1	76.3	70.5
MAPS COMET-QE	**69.0**	**76.0**	**79.7**	**83.3**	**66.9**	**74.7**	**75.9**	**79.1**	**80.8**	**78.5**	**72.3**
`Alpaca | BLEURT`											
Baseline	42.3	58.0	62.2	69.8	31.4	55.4	52.2	63.4	52.4	54.3	53.2
Rerank COMET-QE	47.5	59.5	64.7	70.4	36.2	56.7	55.0	66.0	55.2	59.0	56.0
MAPS COMET-QE	**50.6**	**60.6**	**66.3**	**71.1**	**38.2**	**57.7**	**56.6**	**66.8**	**59.5**	**61.2**	**57.2**
`Vicuna | COMET`											
Baseline	81.3	78.4	79.8	82.9	82.3	77.3	75.5	77.1	74.9	72.7	69.3
Rerank COMET-QE	83.6	79.3	81.8	83.6	85.2	78.8	77.8	79.6	79.9	77.7	74.2
MAPS COMET-QE	**84.5**	**80.2**	**82.7**	**84.1**	**86.5**	**79.7**	**79.2**	**81.1**	**81.8**	**80.1**	**76.0**
`Vicuna | BLEURT`											
Baseline	64.9	65.3	67.4	71.0	58.7	62.8	58.8	66.0	57.8	56.6	57.7
Rerank COMET-QE	66.7	66.0	69.2	71.8	61.6	64.0	61.2	68.2	61.8	61.2	60.5
MAPS COMET-QE	**67.8**	**66.9**	**70.0**	**72.4**	**63.0**	**64.8**	**62.5**	**69.3**	**64.0**	**64.3**	**63.4**

- MAPS exhibits a higher upper bound for selection.
  - COMET: MAPS > Rerank

Human Evaluation

Preference study



☑ MAPS is generally more preferred by humans.

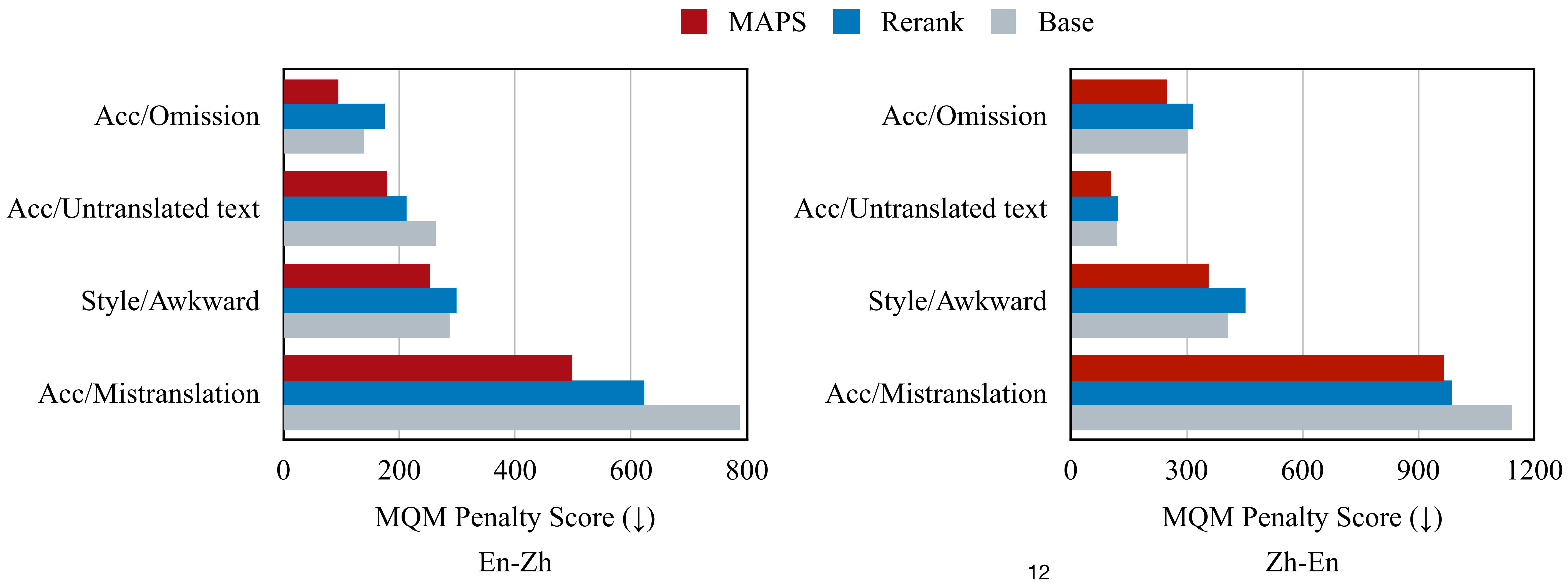
Human Evaluation

Multidimensional quality metrics (MQM)

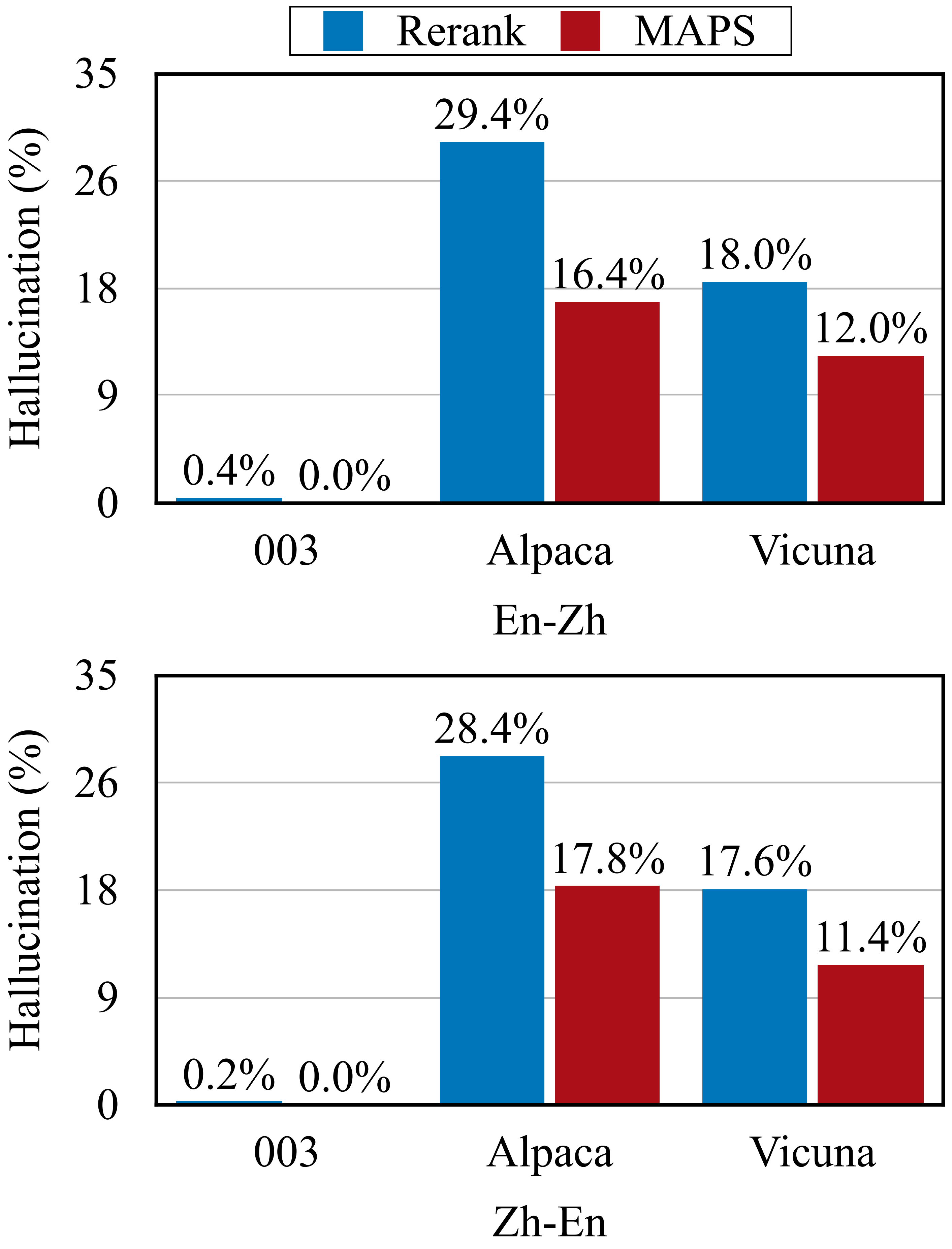
Method	En-Zh	Zh-En
Base	1.94	2.96
Rerank	1.79	2.84
MAPS	**1.59**	**2.60**

Table 2: Averaged MQM Score (↓).

☑ MAPS reduces mistranslation, awkward style, untranslated text, and omission errors.



Hallucination and Ambiguity



Human-annotated hallucination errors

- ☑ MAPS reduces LLM's hallucinations
- ☑ MAPS helps ambiguity resolution

Method	COMET	BLEURT	Accuracy
Rerank	81.5	70.2	61.5
MAPS	**82.2**	**70.6**	**65.5**

Accuracy of ambiguity resolution

Using single type of knowledge does not result in consistent improvement

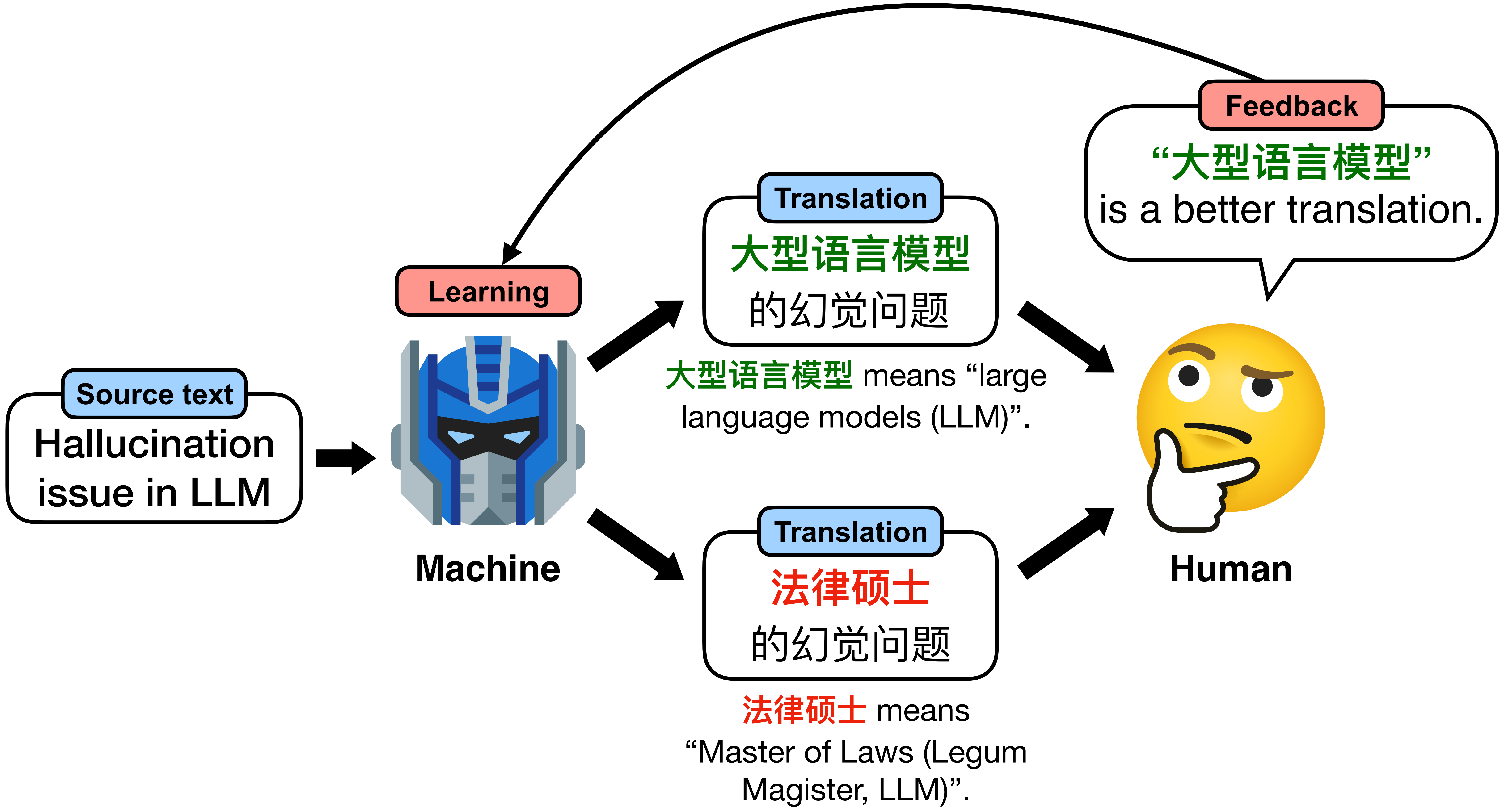
Method	En-Zh	Zh-En	En-De	De-En	En-Ja	Ja-En	De-Fr	Fr-De
			`text-davinci-003` | COMET					
Baseline	86.2	81.6	85.8	85.2	87.9	81.8	82.8	86.3
+Keyword	86.2	81.5	85.5	84.9	88.0	81.5	82.6	86.2
+Topic	86.4	81.7	85.6	85.2	88.1	81.9	83.1	86.3
+Demo	86.9	81.8	86.6	85.2	88.5	81.8	83.4	86.7

- ☑ Self-generated knowledge from LLM can be noisy.
- ☑ Using multiple knowledge and knowledge selection are important.
- ☑ Please refer to the paper for further discussion.

https://arxiv.org/pdf/2305.04118.pdf

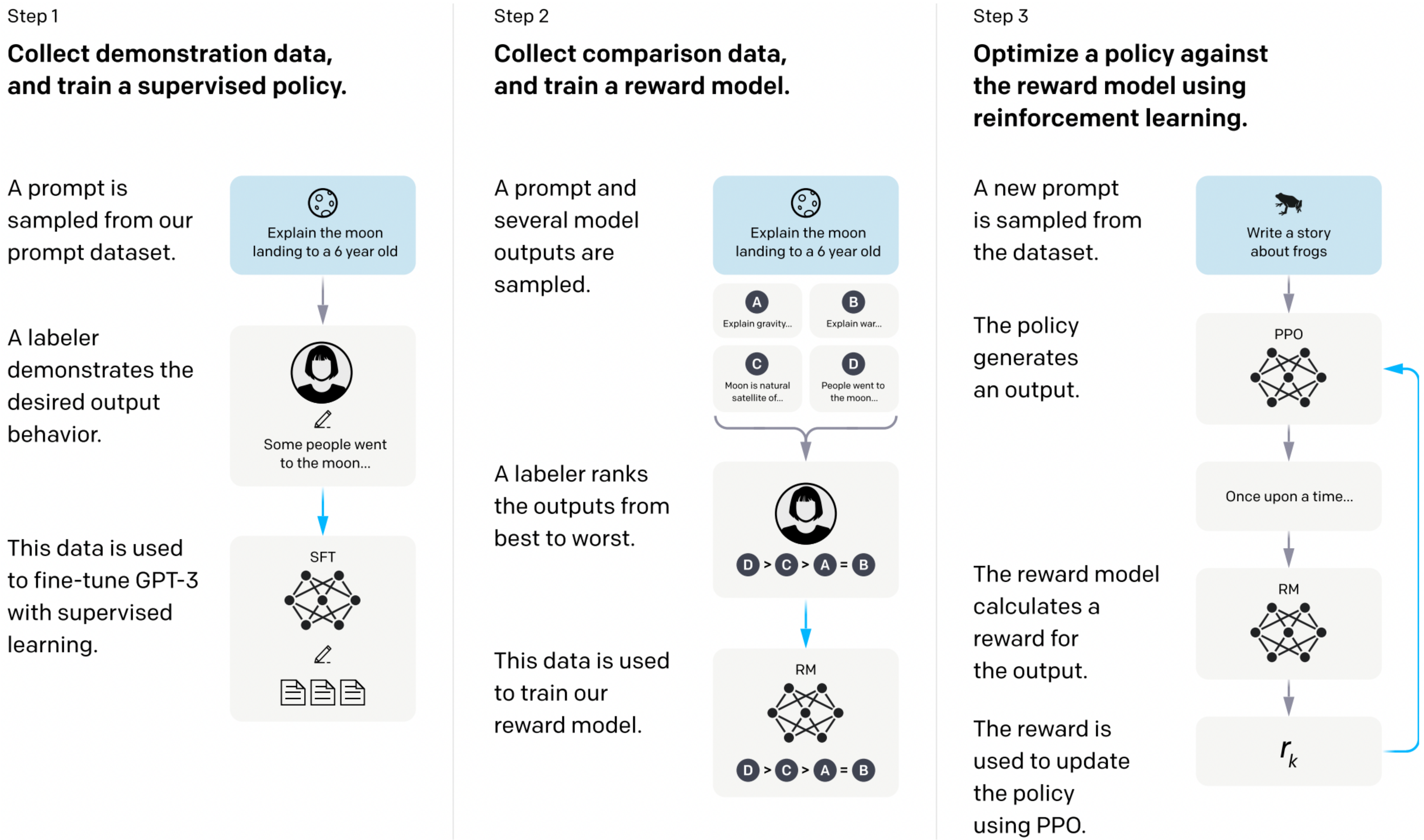
Two Main Limitations of Current NMT Models

Limitation 2: Lacking Human Feedback



- Trained on vast amounts of crawled data, models do not understanding what makes a good translation.
- Incapable of improving translations based on human feedback.

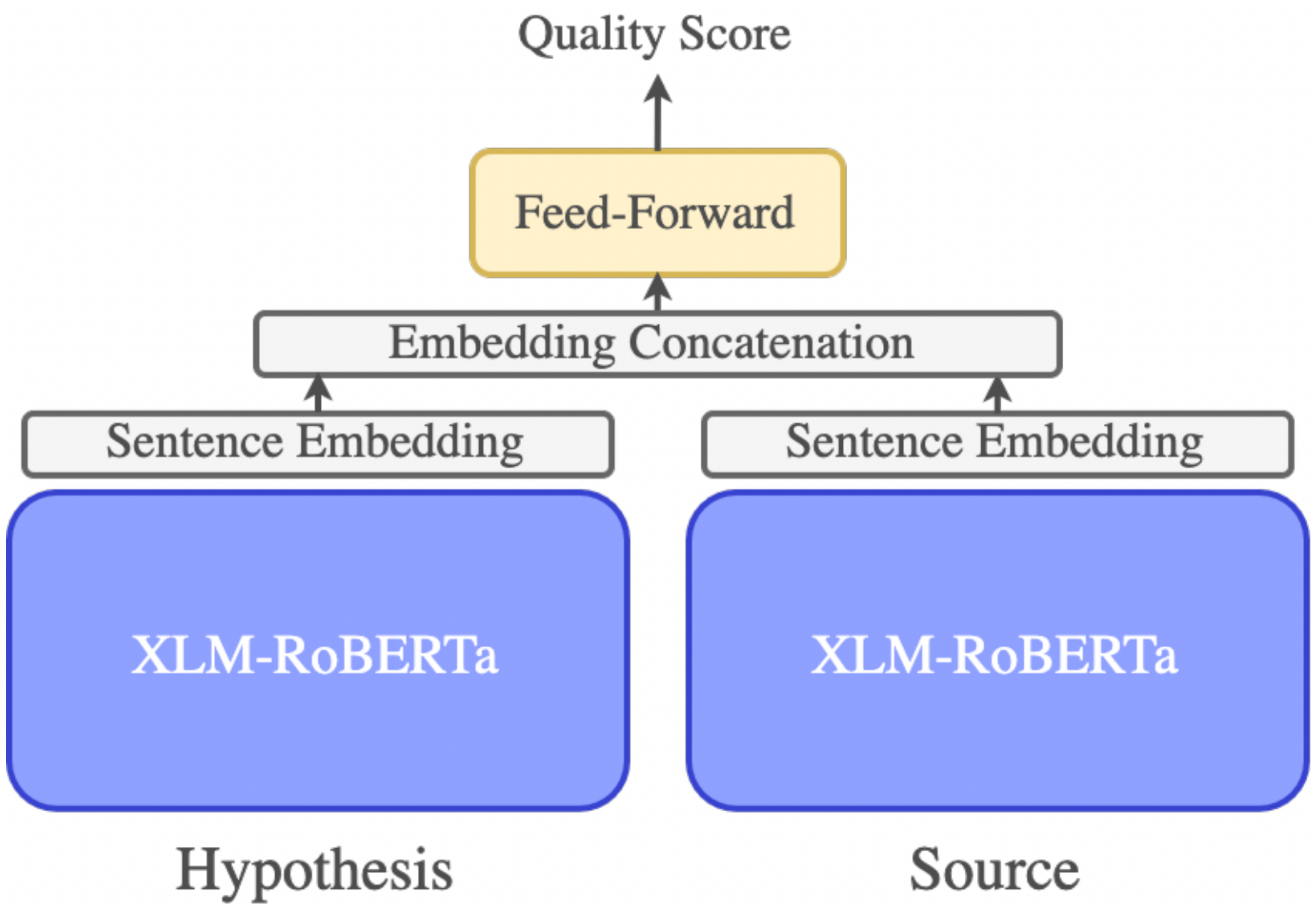
LLMs have already benefited from learning from human feedback



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

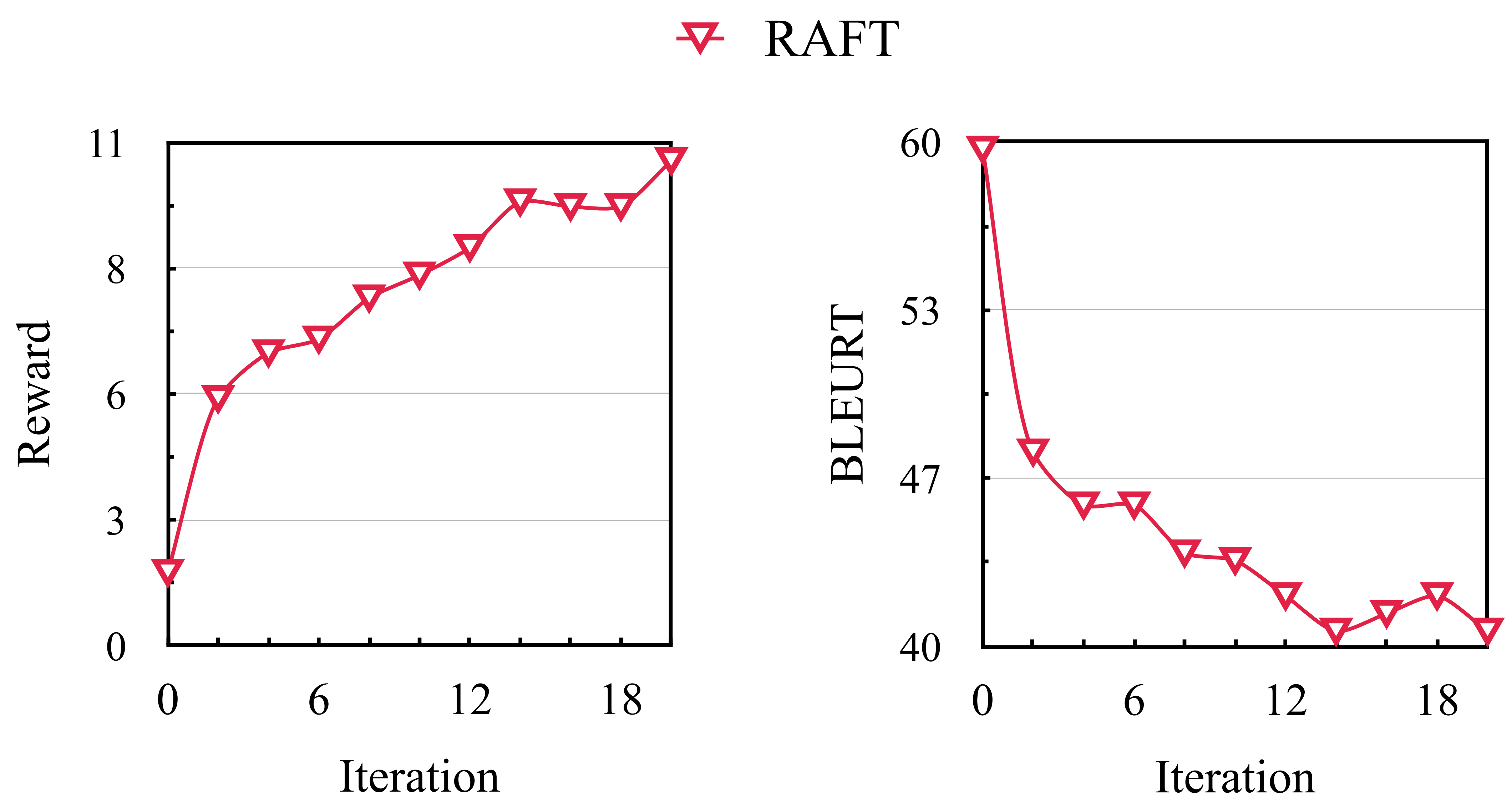
$$\boxed{\max_{\theta} \mathbb{E}_{x\sim\mathcal{D}, y\sim P(y|x;\theta)} r(x,y)}$$

Algorithm 1 RAFT

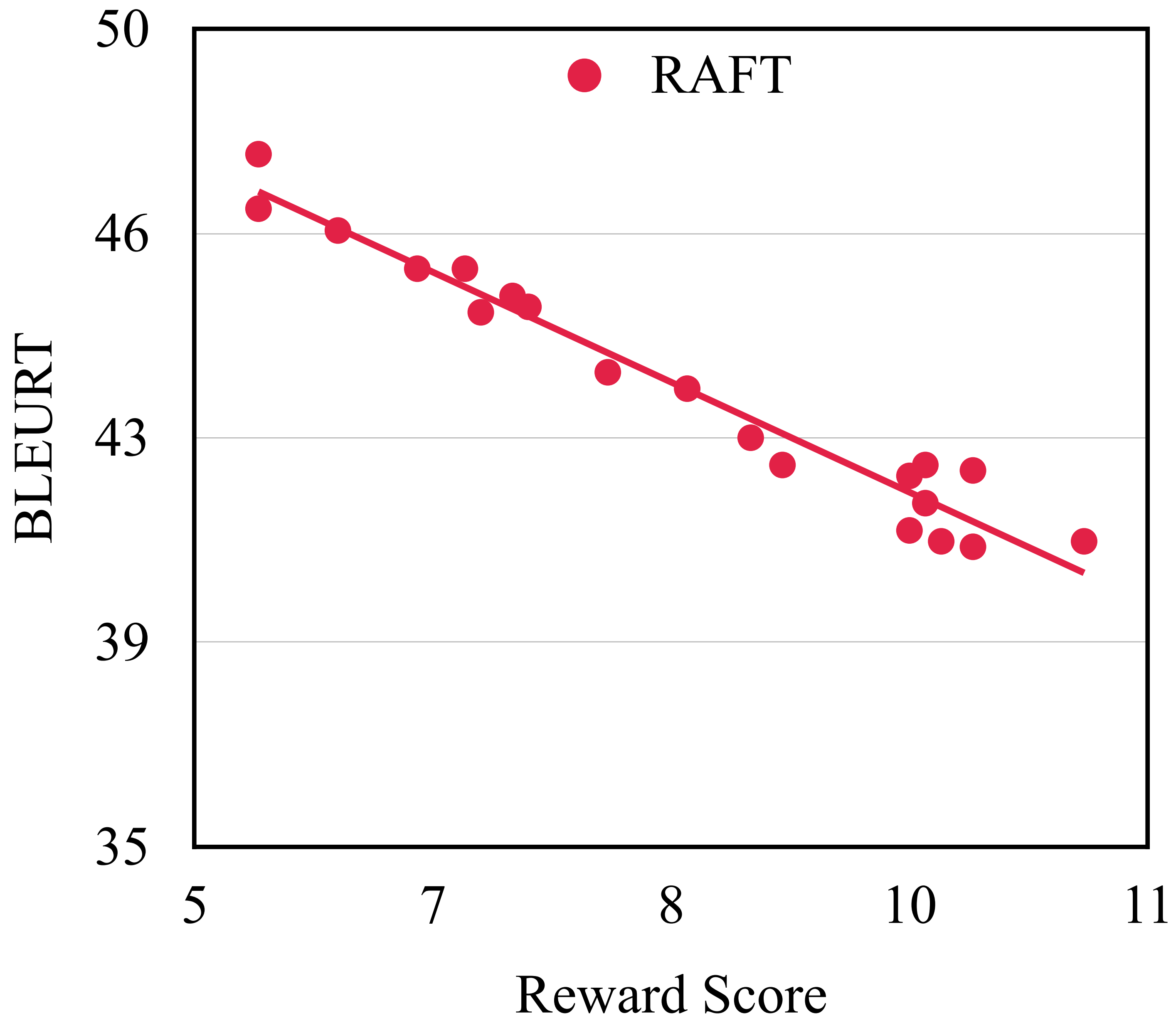
Require: Training set $\mathcal{X}$, reward function $r(x,y)$, initial model $M_0 = P(y|x;\theta_0)$, batch size b, temperature T, the number of candidate k

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

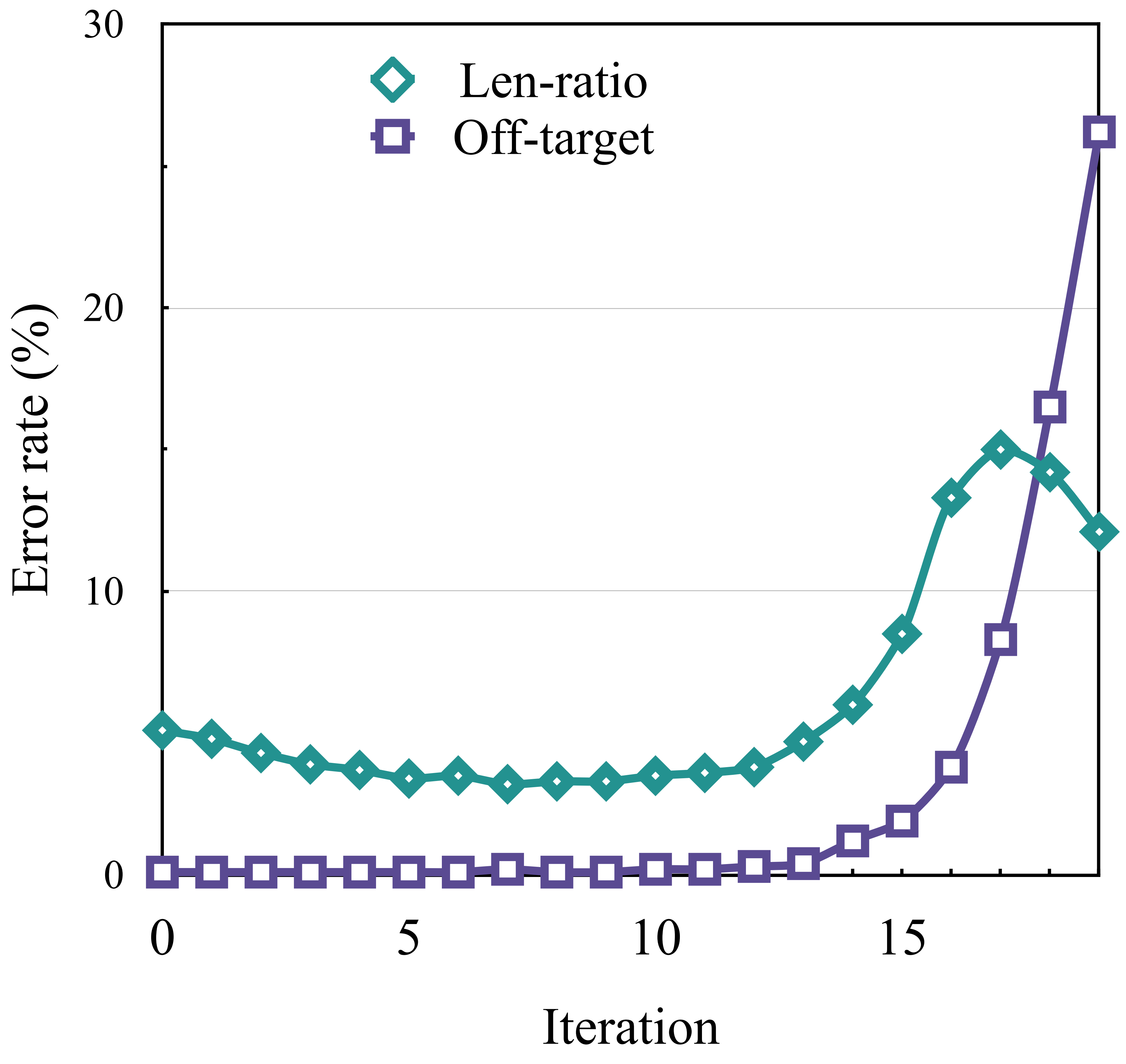
Table 1: A case of Chinese⇒English translation where the QE model (COMET-QE-DA) assigns higher scores to length-ratio and off-target errors than an error-free translation. Error spans are highlighted.

📍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

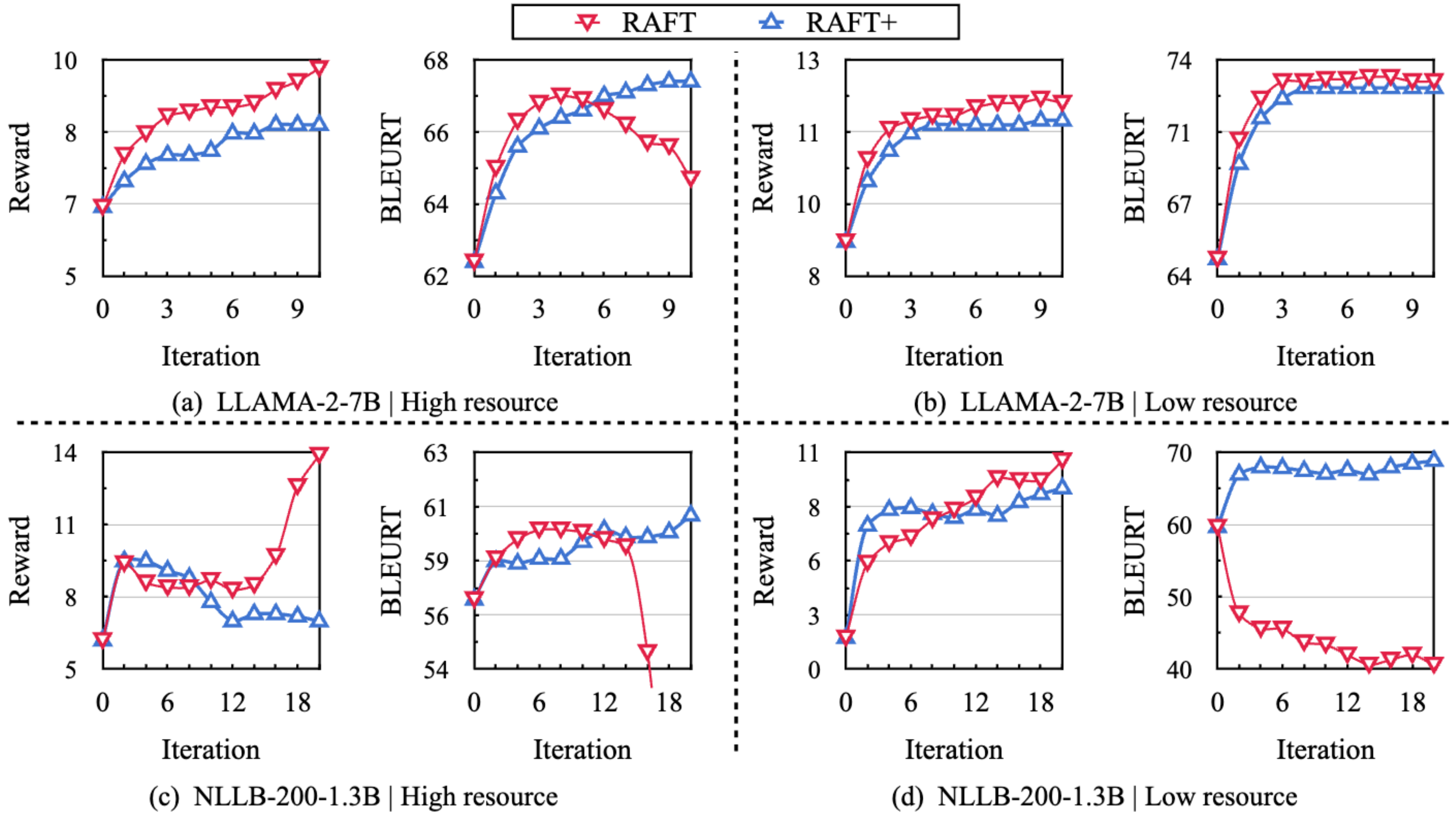
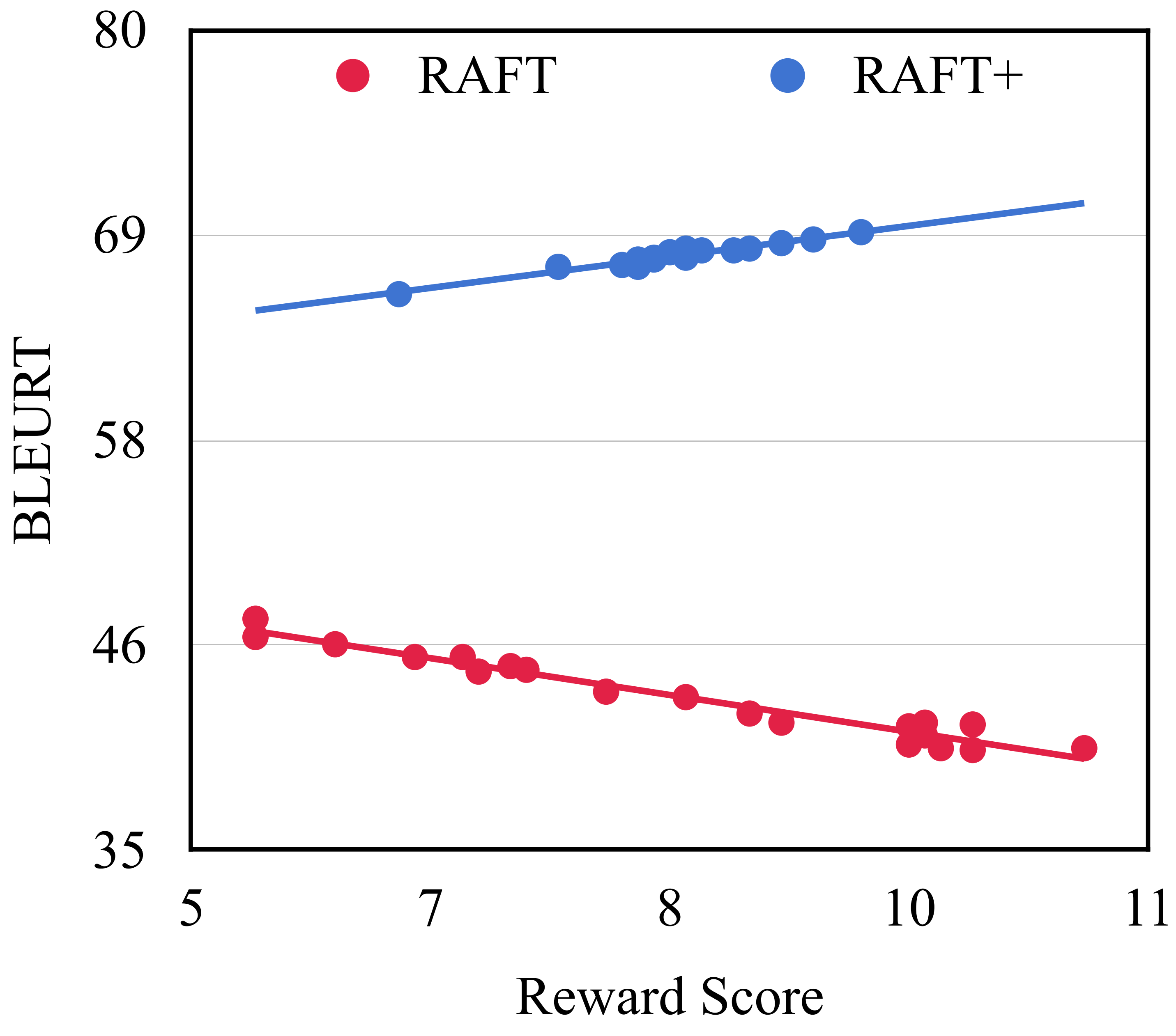


Figure 3: Training curves under various settings. The metrics are average values for all language pairs on the development set. The QE-based reward model is COMET-QE-DA.



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 COMET	De⇒En BLEURT	En⇒De COMET	En⇒De BLEURT	Zh⇒En COMET	Zh⇒En BLEURT	En⇒Zh COMET	En⇒Zh BLEURT	**Average COMET**	**Average 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 COMET	En⇒Uk BLEURT	Uk⇒En COMET	Uk⇒En BLEURT	Uk⇒Cs COMET	Uk⇒Cs BLEURT	Cs⇒Uk COMET	Cs⇒Uk BLEURT	**Average COMET**	**Average 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

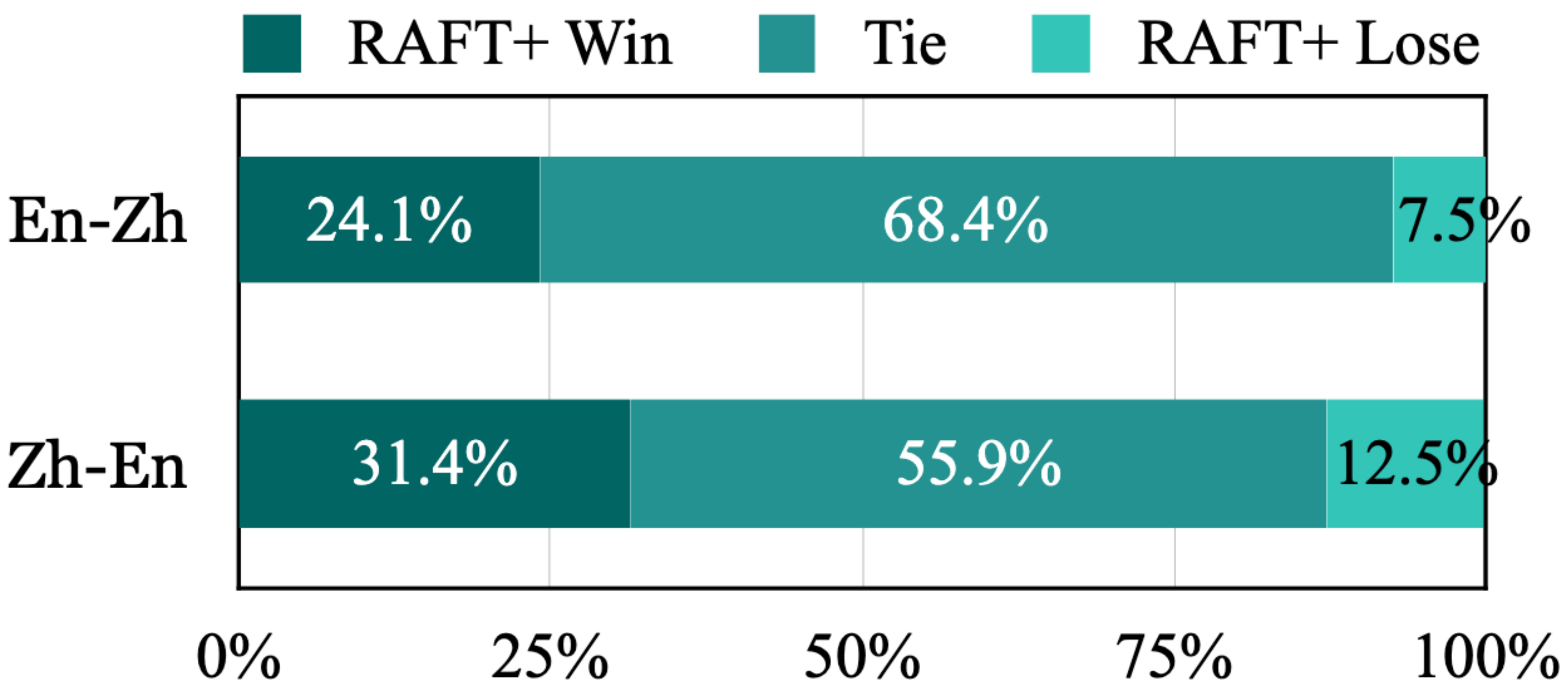


Figure 4: Human preference evaluation, comparing RAFT+ to SFT model on En⇔Zh test sets.

☑ Humans prefer models trained with feedback.

Data Efficiency of Feedback Training

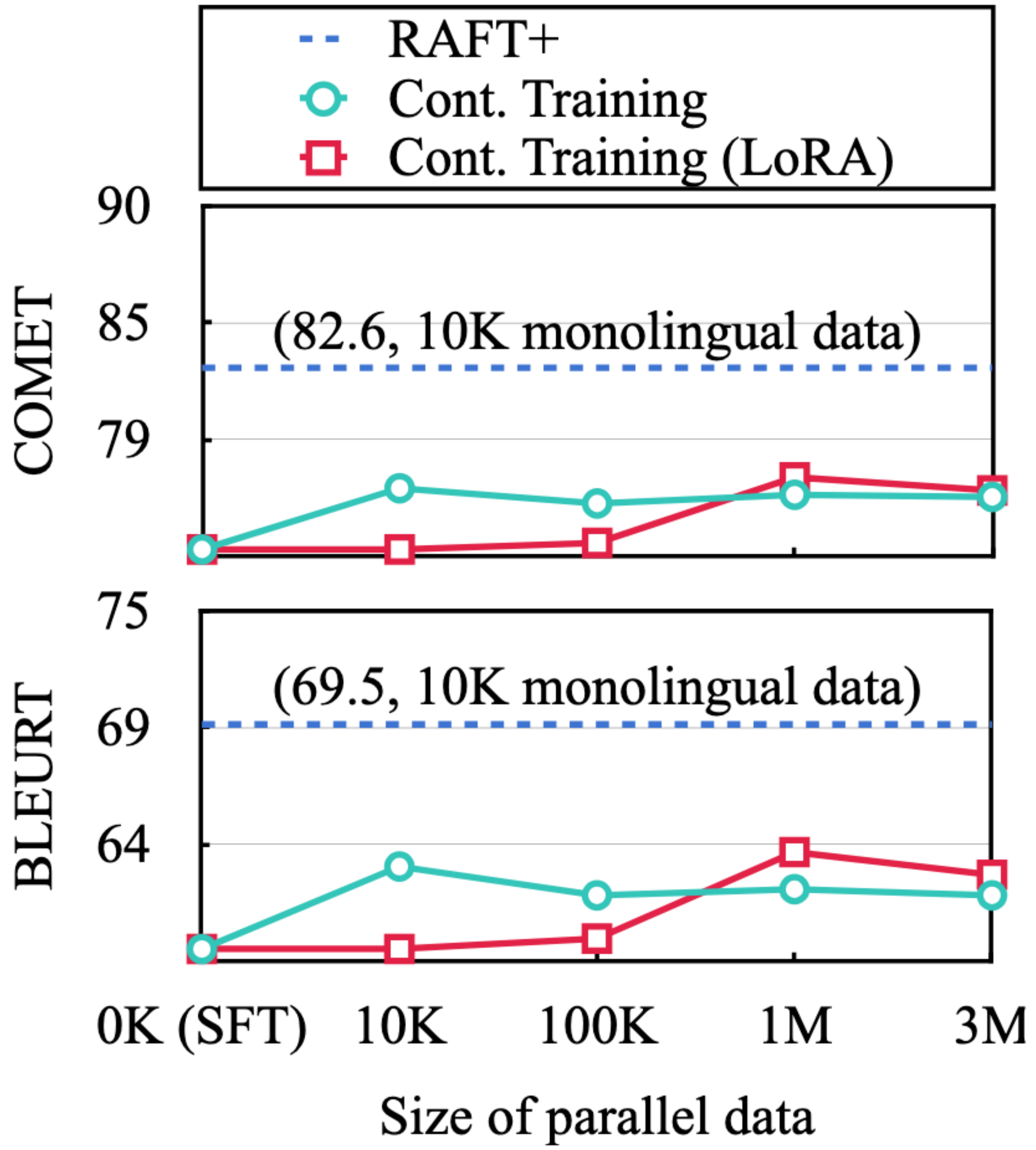


Figure 5: Comparison between RAFT+ and continuous training in the low-resource setting.

☑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

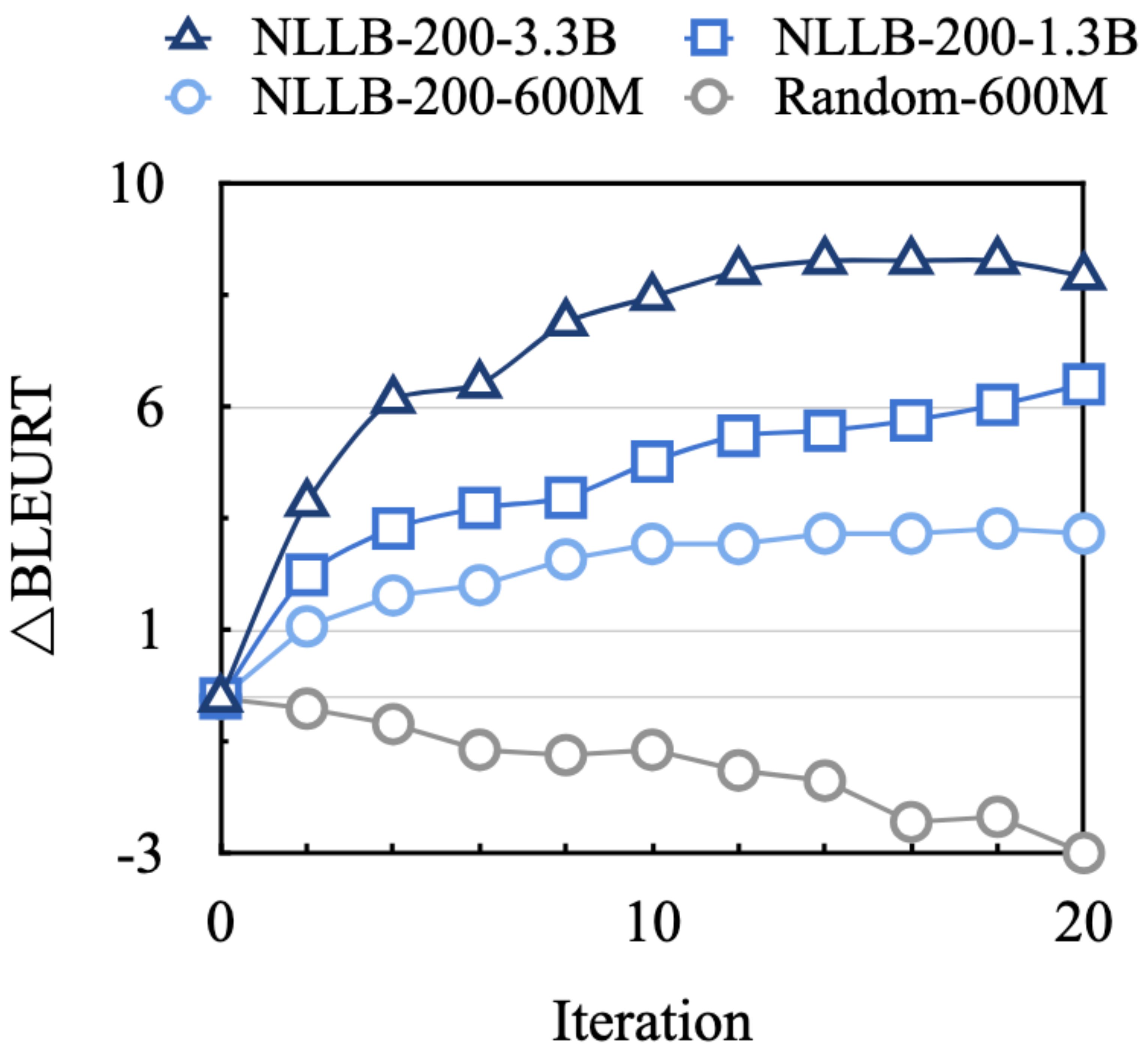


Figure 6: Training curves of RAFT+ (high-resource COMET-QE-MQM) under different base models. We report the change in BLEURT score for each checkpoint relative to the SFT model.

☑ 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.

Summary

☑ LLM can improve translation quality by mimicking human translation strategies.

☑ MT model can learn from human feedback (modeled by QE) after addressing overoptimization.