1. **Prompt Template 1 for “Source + MT”**

Score the following translation from {source\_lang} to {target\_lang} by comparing the source and the translation on a continuous scale from 0 to 100, where score of zero means ‘no meaning preserved’ and score of one hundred means ‘perfect meaning and grammar’.

{source\_lang} source: "{source\_seg}"

{target\_lang} translation: "{target\_seg}"

Score:

1. **Prompt Template 2 for “MT + Ref”**

Score the following translation from {source\_lang} to {target\_lang} with respect to the human reference on a continuous scale from 0 to 100, where score of zero means "no meaning preserved" and score of one hundred means "perfect meaning and grammar".

{target\_lang} translation: "{target\_seg}"

{target\_lang} human reference: {reference\_seg}

Score:

1. **Prompt Template 3 for “Source + MT + Ref”—GEMBA**

Score the following translation from {source\_lang} to {target\_lang} with respect to the human reference on a continuous scale from 0 to 100, where score of zero means "no meaning preserved" and score of one hundred means "perfect meaning and grammar".

{source\_lang} source: "{source\_seg}"

{target\_lang} human reference: {reference\_seg}

{target\_lang} translation: "{target\_seg}"

Score:

1. **Prompt Template 4 for “Source + MT + Error words”**

Score the following translation from {source\_lang} to {target\_lang} by comparing the source and the translation and considering the error words in translation on a continuous scale from 0 to 100, where score of zero means "no meaning preserved" and score of one hundred means "perfect meaning and grammar".

{source\_lang} source: "{source\_seg}"

{target\_lang} translation: "{target\_seg}"

Error words: “{error\_words}”

Score:

1. **Prompt Template 5 for “Source + MT +Ref + Error words”**

Score the following translation from {source\_lang} to {target\_lang} with respect to the human reference and considering the error words in translation on a continuous scale from 0 to 100, where score of zero means "no meaning preserved" and score of one hundred means "perfect meaning and grammar".

{source\_lang} source: "{source\_seg}"

{target\_lang} human reference: {reference\_seg}

{target\_lang} translation: "{target\_seg}"

Error words: “{error\_words}”

Score:

1. **Prompt Template 6 for “Source + MT +Ref + annotation guidelines/instructions”**

Score the following translation from {source\_lang} to {target\_lang } with respect to the human reference on a continuous scale from 0 to 100, where score of zero means "no meaning preserved" in terms of accuracy, contextual understanding, grammar, syntax, overall readability and score of one hundred means "perfect meaning and grammar" in terms of accuracy, contextual understanding, grammar, syntax, overall readability

{source\_lang} source: "{source\_seg}"

{target\_lang} human reference: {reference\_seg}

{target\_lang} translation: "{target\_seg}"

Score:

1. **Prompt Template 7 for CoT**

Prompt 1:

You are going to evaluate the quality for {language\_pair} translation. You need to think step by step. First read the following source, machine translation and reference translation. Analyze where the machine translation is different from the reference translation.

Source: {source\_sentence}

Machine translation: {target\_sentence}

Reference translation: {reference\_translation}

Prompt 2:

A large language model did an evaluation of machine translation quality for the {source\_language} sentence, which is given as below: {output\_from\_Prompt1}

Based on your analysis, score the machine translation quality on a continuous scale from 0 to 100, where score of zero means "no meaning preserved" and

score of one hundred means "perfect meaning and grammar". Provide the score strictly in JSON format.

1. **Prompt Template 8 for few-shot learning**

You are going to evaluate the quality of machine translation given the source, machine translation and reference translation. The followings are examples of scoring translation quality.

{5 examples}

Now score the following translation from {source\_lang} to {target\_lang} with respect to the human reference and examples above on a continuous scale from 0 to 100, where score of zero means "no meaning preserved" and score of one hundred means "perfect meaning and grammar".

{source\_lang} source: "{source\_seg}"

{target\_lang} human reference: {reference\_seg}

{target\_lang} translation: "{target\_seg}"

Score: