Appendix A. Guidelines for human evaluation

1. General instructions

You will be assessing translations at the sentence level. Each translated sentence is aligned with its corresponding source text. You have the flexibility to revise previous annotations as needed.

There are two tasks for you to finish.

The first one is error-analysis-based. Your task is to identify errors within each translated sentence, with a maximum limit of five errors. If there are more than five errors, focus on marking the five most severe ones. In cases where the translation is severely distorted or unrelated to the source, mark a single *Non-translation* error that covers the entire segment.

To identify an error, highlight the relevant portion of the translation and choose a category/sub-category and severity level from the available options. When identifying errors, please be as fine-grained as possible. For instance, if a sentence contains two mistranslated words, record them as separate mistranslation errors. If a single section of text has multiple errors, indicate the most severe one.

Please pay particular attention to the context when annotating. If a translation may be questionable on its own but fits within the context, it should not be considered erroneous. Conversely, if a translation might be acceptable in some contexts, but not for the current sentence, mark it as incorrect.

There is a special error category called *Non-translation*, which can only be used once per sentence and should encompass the entire sentence. If Non-translation is selected, no other errors should be identified.

The second task is impression-based. You need to report your level of agreement with 7 rubrics on a 7-point Likert scale, in which 1 means “completely disagree”, 7 means “completely agree”, and 4 means “basically agree.” The rubrics are as follows:

(1) Cohesion: The translation flow is mostly smooth and cohesive. There is no logical disconnection or meaning inconsistency.
(2) Adherence to norms: The translation fulfills the common standards, requirements, and norms of diplomatic translation.
(3) Style, tone, and register appropriateness: The translation is consistent in style, tone, and register with the source text. For example, if the source text has a formal tone and sophisticated style, the translation also reflects that formality and sophistication.
(4) Cultural sensitivity: The translation demonstrates cultural sensitivity. It suitably conveys culture specific items (CSI), humor, and other cultural nuances in a way that is understandable and relatable to the target audience.
(5) Clarity: The translation is clear and easily understandable to the target audience. It does not contain ambiguities, excessive jargon, or overly complex language that may hinder comprehension.
(6) Practicality: The translation can be directly put to actual use. In this case, it can be put on the government website for people across the world to read.

2. Specifications

We select a portion of parameters from the ASTM F2575–14 Standard Guide for Quality Assurance in Translation cations to describe what is expected of the translation.

Source-content information

Source language: Chinese

Text type: Diplomatic remarks from the Chinese spokesmen.

Audience: Chinese readers who intend to know China's stance on a range of important foreign affairs (e.g., reporters, politicians, and diplomats).

Purpose: to deliver China's stance and attitudes on a range of important foreign affairs.

Complexity: usually written in a relatively complex and formal style, commonly found in official statements or diplomatic contexts.

Origin: official website of the Ministry of Foreign Affairs of The People's Republic of China (https://www.fmprc.gov.cn/fyrbt_673021/dhdw_673027/index_1.shtml)

Target content requirements

Target language: English

Audience: international readers who intend to know China's stance on a range of important foreign affairs (e.g., reporters, politicians, and diplomats).

Purpose: to deliver China's stance and attitudes on a range of important foreign affairs.

Format: written texts displayed on the government website, which are transcribed and carefully edited from spoken remarks.

Style: in a formal, official, and often assertive tone commonly found in official statements or diplomatic discourse.

3. Error typology, severity levels and penalty levels

a. Error typology

Error type	Subtype	Definition
Accuracy		The target text does not accurately reflect the source text, allowing for any differences authorized by specifications.
	Addition	The target text includes text not present in the source.
	Omission	Content is missing from the translation that is present in the source.
	Mistranslation	The target content does not accurately represent the source content.
	Over-translation	The target text is more specific than the source text.
	Under-translation	The target text is less specific than the source text.
Fluency		Issues related to the form or content of a text, irrespective of whether it is a translation or not.
	Punctuation	Punctuation is used incorrectly (for the locale or style).
	Spelling	Issues related to spelling of words.
	Grammar	Issues related to the grammar or syntax of the text, other than spelling and orthography.
	Inconsistency	The text shows internal inconsistency.
	Link/cross-reference	Links are inconsistent in the text.

Terminology	Wrong terms	A term (domain-specific word) is translated with a term other than the one expected for the domain or otherwise specified. Use of term that it is not the term a domain expert would use or because it gives rise to a conceptual mismatch.
	Inconsistent use of terminology	Terminology is used in an inconsistent manner within the text.
Style		The text has stylistic problems.
	Inconsistent style	Style is inconsistent within a text.
	Awkward	A text is written with an awkward style.
	Unidiomatic	The content is grammatical, but not idiomatic.
Other		Any other issues.

b. Severity levels

Non-translation	The translation is severely distorted or unrelated to the source.
Critical	Errors that may carry health, safety, legal or financial implications, violate geopolitical usage guidelines, damage the entities' reputation, or which could be seen as offensive.
Major	Errors that may confuse or mislead the readers due to significant changes in meaning or because errors appear in a visible or important part of the content.
Minor	Errors that don't lead to loss of meaning and wouldn't confuse or mislead the readers but would be noticed, would decrease stylistic quality, fluency or clarity, or would make the content less appealing.
Neutral	Used to log additional information, problems or changes to be made that don´t count as errors, e.g., they reflect a reviewer's choice or preferred style, they are repeated errors or instruction/glossary changes not yet implemented, a change to be made that the translator is not aware of.

c. Penalty levels

Non-translation	100	insert the penalty points for non-translation errors
Critical errors	25	insert the penalty points for critical errors
Major errors	10	insert the penalty points for major errors
Minor errors	1	insert the penalty points for minor errors
Neutral errors	0	insert the penalty points for neutral errors

Appendix B. Descriptive statistics of human-assigned scores

a. Descriptive statistics of ChatGPT 0 shot

	Error	Cohesion	Norms	Style, tone, and register	Cultural sensitivity	Clarity	Practicality
Mean	-5.62	5.18	4.92	4.88	5.06	4.90	4.23
Standard deviation	-6.30	1.44	1.4	1.35	1.61	1.51	1.57
Max	10.00	2.0	2.0	1.00	1.00	2.00	2.00
Min	-35.00	7.0	7.0	7.00	7.00	7.00	7.00
Kurtosis	-4.37	-0.37	0.33	-1.31	-0.92	-0.95	-0.93
Skewness	-1.23	-0.7	0.82	1.44	-0.34	-0.01	-0.19

b. Descriptive statistics of ChatGPT 1 shot

	Error	Cohesion	Norms	Style, tone, and register	Cultural sensitivity	Clarity	Practicality
Mean	-3.81	5.99	5.83	6.04	6.06	5.88	5.32
Standard deviation	-7.68	1.21	1.17	1.39	1.46	1.34	1.35
Max	10.00	2.00	3.00	1.00	1.00	2.00	2.00
Min	-25.00	7.00	7.00	7.00	7.00	7.00	7.00
Kurtosis	-1.27	-1.18	-0.65	-0.79	-0.33	-0.41	-0.54
Skewness	-1.22	0.76	-0.57	-0.56	-0.60	-0.59	-0.70

c. Descriptive statistics of ChatGPT context

	Error	Cohesion	Norms	Style, tone, and register	Cultural sensitivity	Clarity	Practicality
Mean	-4.11	5.68	5.51	5.62	5.58	5.47	5.08
Standard deviation	-9.44	1.51	1.34	1.47	1.63	1.25	1.74
Max	10.00	2.00	2.00	1.00	1.00	1.00	1.00
Min	-35.00	7.00	7.00	7.00	7.00	7.00	7.00
Kurtosis	-1.26	-0.95	-0.43	0.21	-0.55	-1.03	-1.03
Skewness	-1.41	-0.01	-0.45	-0.75	-0.42	0.31	0.31

d. Descriptive statistics of Google Translate

	Error	Cohesion	Norms	Style, tone, and register	Cultural sensitivity	Clarity	Practicality
Mean	-5.72	5.30	5.48	5.57	5.54	5.46	4.99
Standard deviation	-6.31	1.35	1.39	1.21	1.37	1.17	1.21
Max	10.00	1.00	1.00	3.00	1.00	2.00	2.00
Min	-35.00	7.00	7.00	7.00	7.00	7.00	7.00
Kurtosis	-4.37	-1.31	-0.79	-0.49	-0.75	-0.73	-0.99
Skewness	-1.23	1.43	-0.02	-0.68	-0.36	-0.56	0.42

e. Descriptive statistics of MS Translator

	Error	Cohesion	Norms	Style, tone, and register	Cultural sensitivity	Clarity	Practicality
Mean	-5.68	4.89	5.09	5.11	5.18	5.05	4.57
Standard deviation	-7.92	1.61	1.46	1.33	1.54	1.44	1.50
Max	5.00	1.00	1.00	3.00	1.00	2.00	1.00
Min	-25.00	7.00	7.00	7.00	7.00	7.00	7.00
Kurtosis	-0.21	0.215	-0.98	-1.12	-1.21	-0.55	-0.81
Skewness	-1.07	-0.911	-0.23	-0.18	-0.28	-0.73	-0.25

f. Descriptive statistics of DeepL

	Error	Cohesion	Norms	Style, tone, and register	Cultural sensitivity	Clarity	Practicality
Mean	-4.80	5.43	5.47	5.59	5.36	5.37	5.03
Standard deviation	-6.87	1.57	1.35	1.19	1.66	1.67	1.64
Max	0.00	2.00	2.00	2.00	1.00	1.00	1.00
Min	-25.00	7.00	7.00	7.00	7.00	7.00	7.00
Kurtosis	-1.41	-0.93	-0.54	-0.11	-0.28	-0.22	-0.73
Skewness	-1.48	-0.19	-0.70	-0.17	-0.08	-0.41	-0.54

Appendix C. MFRM-calibrated fair scores

	Cohesion	Norms	Style, tone, and register	Cultural sensitivity	Clarity	Practicality
ChatGPT 0 shot	6.02	5.29	5.24	5.63	5.43	4.68
ChatGPT 1 shot	6.63	6.02	6.11	6.57	6.14	5.68
ChatGPT context	6.5	5.76	5.73	6.27	5.79	5.42
Google	6.1	5.64	5.71	5.82	5.81	5.41
MS	6.08	5.28	5.45	5.76	5.43	5.06
DeepL	6.26	5.47	5.69	6.01	5.72	5.42

Appendix D. Correlation coefficients between human and automated assessment

	BLEU	chrF	COMET	BERTScore
Error penalty	0.19*	0.2*	0.36*	0.41*
Cohesion	0.08	0.13	0.33*	0.32*
Adherence to norms	0.07	0.09	0.24	0.23
Style, tone, and register	0.13	0.13	0.21	0.31*
Cultural sensitivity	0.11	0.08	0.17	0.19
Clarity	0.16	0.11	0.30*	0.34*
Practicality	0.04	0.14	0.19	0.23*

Note: *$p < 0.05$