
LLMs as Translator Training Partners: A Multi-Agent Approach

Ming Qian

Human-Centered AI, Charles River Analytics, Cambridge, MA, USA

qianmi@gmail.com

Luyi Yang

Certified Translator and Interpreter, Toronto, ON, Canada

luyiyang335@gmail.com

Abstract

Peer-review-based translator training promotes reflection and collaborative critique. This study investigates whether GPT-5.4, guided by Multidimensional Quality Metrics (MQM) style prompts, can function as a peer-review training partner rather than a grading tool. Using translated passages from a translator practice group, we compared GPT-5.4's feedback with human peer-review comments on the same translation segments, including both negative and positive judgments, and used a second large language model (DeepSeek) to arbitrate disagreements between human and AI judgments. GPT-5.4 agreed with human trainees on the existence of translation issues in 77.8% of negative flags and 88.9% of positive flags and achieved an F1 score of 0.875 against in-depth verification by two human evaluators. Although GPT-5.4 occasionally produced incorrect judgments, it also generated more comprehensive and analytical feedback than human trainees and frequently offered alternative but equally valid interpretations that could stimulate deeper reflection. DeepSeek successfully resolved a substantial proportion of human-LLM disagreements, demonstrating the value of an independent LLM arbitrator. These findings suggest that a multi-agent framework combining an MQM-guided reviewer, human deliberation, and LLM-based arbitration is a feasible and promising approach for developing reliable AI training partners for translator education.

1 Introduction

Peer-review-based learning asks learners to evaluate and comment on one another's work. By comparing peers' feedback, learners engage in metacognitive reflection: examining alternative strategies, identifying strengths and weaknesses, and reconsidering their own choices. Exposure to diverse perspectives can further improve learning outcomes and strengthen group problem solving.

LLM-based quality checks, or "LLM-as-judge" systems, are fast and scalable but still struggle to capture nuance, address critical issues, and detect subtle domain-specific errors (Li et al., 2024; Wang, R., et al., 2025). Bavaresco et al. (2025) questioned their validity as substitutes for human judges in translation and NLP tasks, citing biases and inconsistencies with human assessments. Wang, G., et al. (2025) proposed a multi-agent automated translation system using the Multidimensional Quality Metrics (MQM) framework for fine-grained error detection and refinement but did not report empirical results. Overall, available evidence (Feng et al., 2025; Gu et al., 2024; Szymanski et al., 2025; Ye et al., 2024) suggests that LLM-as-judge systems are best used as auxiliary quality-control tools, not replacements for expert human evaluators, especially in high-stakes contexts such as ATA exam grading.

In this study, we examine LLMs as training partner

need training partners that are readily available, can identify problematic text, and provide feedback that encourages deeper reflection, even if they do not match grader-level accuracy. The core argument is that a training partner is valuable not because they are always correct, but because they create productive friction. If they sometimes challenge you, disagree with you, or even make occasional mistakes, they can still help you learn by forcing you to think, test your assumptions, and externalize your reasoning more clearly.

LLMs are well suited to this role because they are consistently accessible and offer a low-stress environment for practice. We used ChatGPT5.4 to generate peer-review feedback based on the MQM framework. MQM assesses translation quality through customizable error categories and severity levels, including accuracy, fluency, terminology, and style (Lommel et al., 2014; MQM Council, 2025). Prior work has used multi-agent LLM systems to automatically evaluate distinct MQM categories (Fernandes et al., 2023; Lu et al., 2025; Wang G. et al., 2025).

To evaluate our approach, we compared GPT-5.4-generated MQM evaluations with human trainees' peer-review comments on the same human-translated passage. The human review data came from a peer-led translation practice group organized by the ATA Chinese

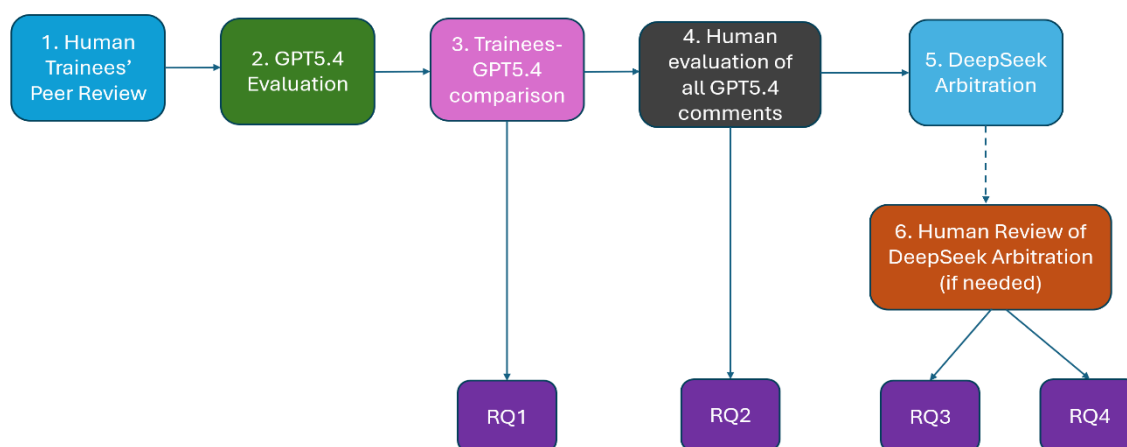


Figure 1: Experimental workflow and analysis process.

(American Translators Association – Community Led Division, n.d.). Four trainees translated a short exam-style passage and reviewed anonymized peer translations in a shared Google Doc. Trainees provided both positive and negative review comments. Notably, one trainee had passed the ATA certification exam before the practice in December 2023 and two trainees passed the ATA certification exam sometime after the practice.

We then prompted GPT-5.4 using an MQM-style prompt to generate positive and negative review comments for the same passage. First, our analysis focused on translation segments that were annotated by both the human trainees and GPT-5.4. Specifically, we addressed two research questions:

1. What percentage of the human trainees' positive and negative comments were also identified by GPT-5.4? How do comments with the same overall polarity differ in error types and severity?
2. Since GPT-5.4 generated substantially more comments than the human trainees, two human evaluators assessed all GPT5.4 comments to determine what percentage of its positive and negative comments could be confirmed as valid (we evaluate both true and false positives and negatives and calculate the F1 score as a quality indicator).

A second large language model (DeepSeek-V4) was used to arbitrate cases in which GPT-5.4 disagreed with the human trainees and/or evaluators. For cases where DeepSeek's judgments conflicted with the evaluators' assessments, the two evaluators conducted detailed reviews to reach a final consensus on the correctness of DeepSeek's arbitration. We investigated the following research questions:

3. What percentage of DeepSeek's arbitration decisions were judged to be correct?
4. How much does the human-LLM deliberation workflow, followed by LLM arbitration,

process?

Figure 1 illustrates the experimental workflow and analysis process.

2 Experimental Data and MQM-like Prompts

Four trainees independently translated the same English-to-Chinese passage and evaluated each other's work. MQM-style prompts were then used to generate GPT5.4 evaluation comments. Because GPT5.4 produced far more comments than the human trainees, two evaluators reviewed all GPT5.4 comments to validate them. Implementation details follow.

2.1 Experimental data

Four trainees were each assigned to translate the same English passage—comparable in length and difficulty to those on the ATA certification exam—into Chinese. All translations were submitted anonymously in a shared Google Doc, after which each trainee reviewed the others' work. *Below is the English passage (an excerpt from: <https://www.mckinsey.com/featured-insights/future-of-asia/videos/is-asia-on-the-cusp-of-a-new-era>):*

“We believe that the world is on the cusp of a new era with very different underlying forces than those that shaped our world over the past 30 years. Asia will be at the nexus of these forces and will experience heightened versions of the trends affecting the rest of the world.

Asia is the world's trade crossroads. It's home to 18 of the 20 fastest-growing trade corridors and 13 of the 20 largest. But now, it could find itself in the crosshairs of trade tensions. Asia's integration has come about through commercial pragmatism rather than political alignment. Can that continue if trade tensions rise?

Asia excels in manufacturing new technology in four areas in particular: consumer electronics, automobiles, and

semiconductors. In these places, Asia has more than 40 percent of the revenues, R&D, and patents. But the value created by tech is moving away from hardware and into new transversal technologies. So, the question is whether Asia can reinvent itself as a tech creator as well as a tech manufacturer.

Asia has largely been in a demographic sweet spot with large pools of young workers and surging productivity in some economies. But now, aging is fiercest in the high-productivity economies of the Pacific Rim. Asia has the people to fuel growth, but 90 percent of the expected increase in Asia's non-farm workforces by 2050 will be where productivity is relatively low. Can Asia shift value chains to where the workers are and raise productivity everywhere?"

2.2 MQM-like Prompts

For GPT5.4-generated negative comments, we used the following prompt based on the prompt proposed in Wang et al. (2025):

"The whole source English text segment: ...

The source text sentence:

"We believe that the world is on the cusp of a new era with very different underlying forces than those that shaped our world over the past 30 years."

The translated target text sentence:

"我们相信新的时代马上就要来临，这个新时代深层暗涌的力量与在过去30年促成世界格局的力量大不相同。"

You are a professional translator responsible for evaluating the quality of the translated target text sentence based on the Multidimensional Quality Metrics (MQM) framework.

Your task is to identify all errors in the translation and classify them into the following dimensions:

Accuracy (Mistranslation, omission, addition, over-translation),

Terminology (Incorrect term choice, missing glossary application),

Linguistic Conventions (Grammar mistakes, punctuation errors, typographical issues),

Style (Unnatural flow, awkward expressions, overly literal phrasing),

Locale Conventions (Wrong format, incorrect symbols),

Audience Appropriateness (Overly technical terms for general audiences, offensive wording),

Design and Markup (Missing markup, broken text formatting, layout inconsistencies),

Each error must

Critical: Critical errors inhibit comprehension of the text.

Major: Major errors disrupt the flow, but what the text is trying to say is still understandable.

Minor: Minor errors are technically errors, but do not disrupt the flow or hinder comprehension.

If no error is detected, return "no-error" in its severity."

For GPT5.4-generated positive comments, we used the following prompt:

"The whole source English text segment: ...

The source text sentence:

"We believe that the world is on the cusp of a new era with very different underlying forces than those that shaped our world over the past 30 years."

The translated target text sentence:

"我们相信新的时代马上就要来临，这个新时代深层暗涌的力量与在过去30年促成世界格局的力量大不相同。"

You are a professional translator responsible for evaluating the quality of the translated target text sentence based on the Multidimensional Quality Metrics (MQM) framework.

Please give some feedback on where the translation is done exceptionally well and commendable:

Accuracy (meaning is accurately conveyed across language and culture, no mistranslation, no omission, no addition, no over-translation),

Terminology (correct term choice),

Linguistic Conventions (no grammar mistakes, no punctuation errors, no typographical issues),

Style (flows eloquently and naturally as if a native speaker who has a good command of the target language wrote it; no awkward expressions and not read like a rigid translation, no overly literal phrasing)

Locale Conventions (correct format, correct symbols),

Audience Appropriateness (proper technical terms for educated laymen, no offensive wording).

If none is detected as exceptionally well, return "Nothing is found."

..

It's important to note that, before the prompts, we include the full source paragraph, the specific source sentence of interest, and the translated target sentence that is under evaluation.

2.3 DeepSeek Arbitration Prompt

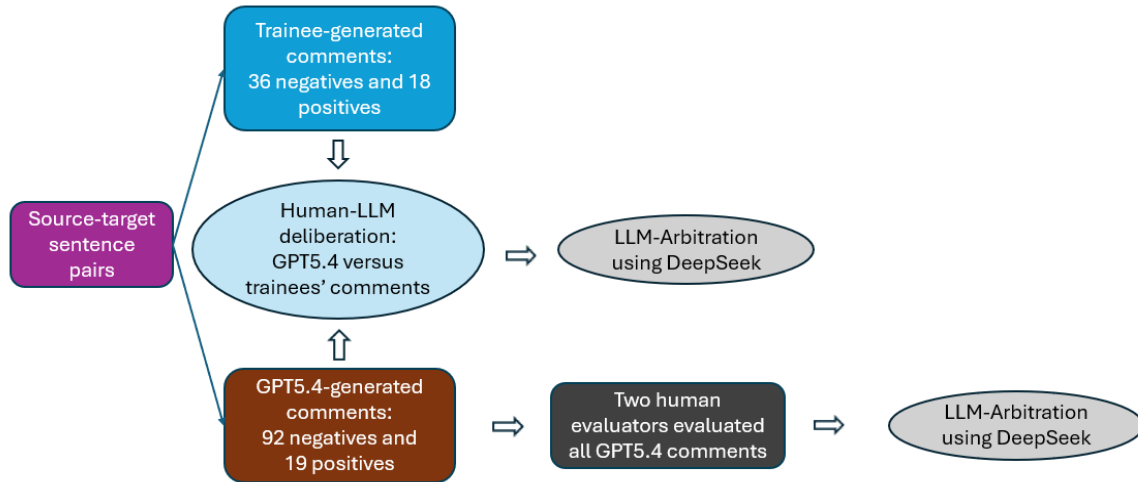


Figure 2: Collecting performance statistics based on trainee-generated and GPT5.4-generated comments through human-LLM deliberation and LLM-arbitration.

We used the following prompt to elicit DeepSeek arbitration:

“There are two comments on the translation pair below:

“Asia has more than 40 percent of the revenues”

“亚洲占了超过百分之四十的收益”

Comment 1: “40% “ should be written in Chinese text ” 百分之四十 “. Are there any specific formatting rules?

Comment 2: No problem detected.

Which comment is better? Why?

”

2.4 Evaluations

We established a human-generated baseline by collecting four trainees’ translation outcomes and their review comments towards each other. Altogether, we collected 32 negative, 18 positive trainee-generated comments, and the corresponding source/target sentence pairs.

Using the collected source–target sentence pairs and the MQM-style prompts described in Section 2.2, we generated both positive and negative comments with GPT5.4. Based on these results, we computed the following statistics:

- (1) What percentage of trainee-generated negative comments were flagged by GPT5.4 using the MQM-style prompt?
- (2) What percentage of trainee-generated positive comments were flagged by GPT5.4 using the MQM-style prompt?
- (3) What percentage of GPT-generated negative comments were confirmed by the two human evaluators?
- (4) What percentage of GPT-generated positive comments were confirmed by the two human evaluators?

DeepSeek (a second LLM) was used to arbitrate cases in which GPT-5.4 disagreed with the human trainees and/or evaluators. For cases where DeepSeek’s judgments conflicted with the evaluators’ assessments, the two evaluators conducted (additional) detailed reviews to reach a final consensus on the correctness of DeepSeek’s arbitration. Based on these results, we computed the following statistics

- (5) What percentage of DeepSeek’s arbitration decisions were judged to be correct?
- (6) How much does the human–AI deliberation workflow, followed by LLM arbitration, improve the overall evaluation process?

In terms of performance measures:

- (1) We calculated the accuracy and misclassification rates of GPT5.4 in confirming human-generated negative and positive comments.
- (2) For GPT5.4-generated negative and positive comments, we calculated true positives, true negatives, false positives, false negatives, and the F1 score based on the two evaluators’ investigation.

Figure 2 illustrates how performance statistics are collected based on trainee-generated and GPT5.4-generated comments through human-LLM deliberation and LLM-arbitration.

3 Results and Discussion

We measured both GPT5.4 accuracy relative to trainees’ judgments (using trainee-identified negatives/positives as the baseline) and classification performance of GPT5.4-generated comments (using GPT5.4-generated negatives/positives as the baseline).

3.1 GPT5.4 Accuracy Relative to Trainees’ Judgments

Appendix Table 1 summarizes the 36 errors

error existence was treated as a binary decision. Human trainees and GPT-5.4 could agree that an error existed while disagreeing on its type or severity. GPT-5.4 agreed with the trainees on error existence in 28 of the 36 cases. Among these shared errors, they agreed on the error type in 17 of 28 cases and on error severity in 18 of 28 cases. In the remaining 8 cases, GPT-5.4 did not identify an error that the human trainees had flagged. The two human evaluators determined that these disagreements primarily reflected trainee nitpicking of very minor issues, suggesting that GPT-5.4 reduced over-flagging of trivial issues. However, the evaluators also found that GPT-5.4 overstated error severity in 9 cases by classifying minor errors as major, indicating a tendency to overestimate severity after the error existence was established.

In four instances, GPT5.4 and human trainees marked the same text span as erroneous but assigned different error types. In text segment 8, a human trainee labeled “生产力最高的经济体” as a mistranslation of “high-productivity economies,” because “最高” means “highest,” whereas the source indicates “high,” is not a superlative. GPT5.4 instead labeled it as a style error, noting that the rendering was overly literal and less natural than standard Chinese economic terminology. In this case, the trainee’s annotation is more accurate. Also in text segment 8, a human trainee flagged “老龄化……最为激烈” as an unnatural expression, while GPT5.4 classified it as a word-choice nuance error. In this case, neither annotation is clearly preferable. In text segment 9, a human trainee found that “fuel growth” was mistranslated, whereas GPT5.4 described the translation as overly literal. The human trainee’s annotation is more accurate. In text segment 13, a human annotator flagged an incorrect translation of “surging,” while GPT5.4 classified it as an awkward translation. Here again, the trainee’s annotation provides the more accurate explanation. Overall, when human trainees and GPT5.4 disagree on error type, the human annotations are generally more accurate and address root causes instead of surface effects.

Appendix Table 2 lists 18 cases where human trainees identified notable strengths rather than errors. When evaluating the same sentence pairs, GPT5.4 agreed in 16 of 18 cases. In the remaining two cases, GPT5.4 judged the translations as merely acceptable, whereas trainees found them to be exceptional—the two evaluators determined that the two cases were not strong at all. Humans often elaborate more on criticism than praise because of negativity bias (“Negativity Bias,” 2026). In actual study groups, human trainees’ praise is often shorter and more judicious.

pre-designed feedback typically requested in controlled experiments. In contrast, GPT5.4, when properly prompted, tends to give more analytical, evaluative, and detailed explanations for positive feedback. For example,

Trainee’s positive comment: “新时代的风口浪尖” is a vivid translation.

GPT5.4: The translation, “新时代的风口浪尖”, adds rhetorical strength and make the sentence vivid, while still staying close to the original meaning. It does not feel mechanical or overly literal.

3.2 Classification Performance of GPT5.4-generated comments

For the same 28 sentence pairs that human translators marked with 36 negative or 18 positive comments, GPT5.4 generated 92 negative comments and 19 positive comments. Appendix Table 3 lists the 92 cases in which GPT5.4 flagged errors; the two authors (evaluators) evaluated these cases and found that 77 were reasonable and 15 were false. Appendix Table 4 lists the 19 cases in which GPT5.4 noted strengths; the two evaluators found that 12 were reasonable and 7 were false.

To evaluate GPT5.4 as an AI system for automatically commenting on translation quality, the F1 score is calculated, based on true positives, false positives, true negatives, and false negatives, as $(2 \times 77) / (2 \times 77 + 15 + 7) = 0.875$. There is no universal “good” F1-score; it depends on the task and the cost of false positives versus false negatives. In practice, 0.8 to 0.9 is often considered good, and above 0.9 is usually excellent. Therefore, GPT5.4, prompted by the MQM-like prompt can be considered a good AI system for this small task in this study.

Below is a list of reasons we identified where GPT5.4 generated the false negatives (examples in which GPT5.4 identified as false while human evaluators deemed acceptable):

- Misclassifying a suboptimal but acceptable translation as unacceptable.
- Incorrectly flagging meaning omissions or additions: translation is not always a one-to-one mapping. A source sentence may be split into multiple target sentences, or several source clauses may be combined into one coherent translated sentence.
- False fault isolation: mistranslation may stem from a misunderstanding of the source sentence as a whole but GPT5.4 may incorrectly identify the translation of a specific source word as the cause of the problem instead of a sentence-level issue.

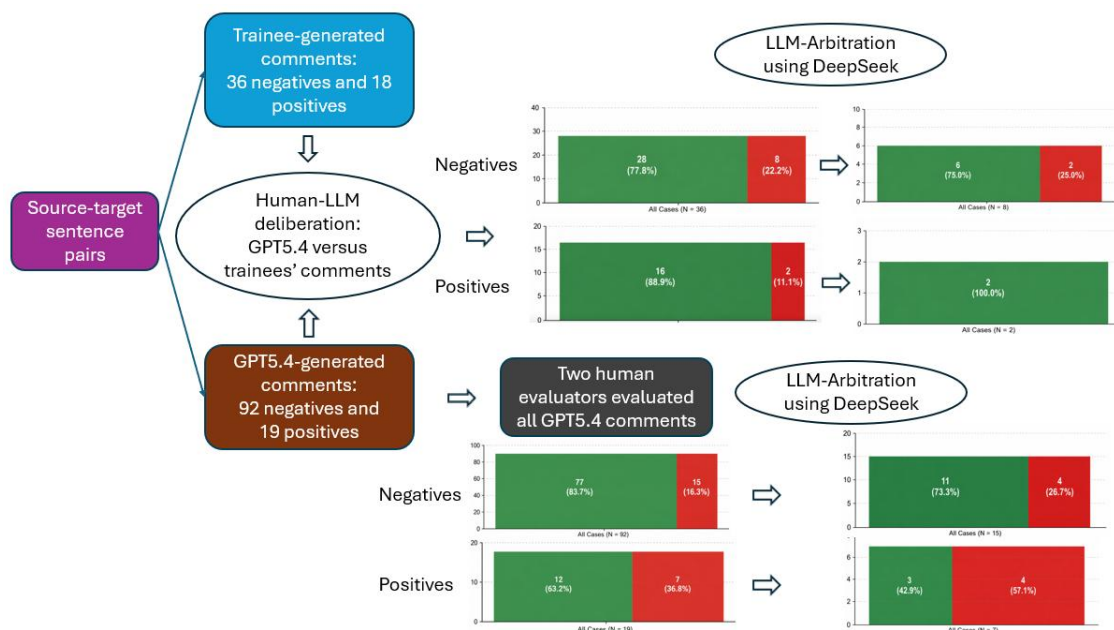


Figure 3: Comparison of LLM deliberation and arbitration results with those of human trainees and evaluators.

ignore the fact that syntax works differently across languages and expect parts of speech in the target text to align perfectly with the source, even when the same meaning is expressed naturally through different structures.

- Overly strict structural matching: GPT5.4 may complain that the target expression introduces an structure not present in the source sentence and treat it as a mistranslation, even when that structure makes the translation more natural in the target language.

- Overreliance on common translation choices: GPT5.4 may rigidly insist on the most commonly seen translation and ignore contextual factors that could justify alternative choices.

- Overly narrow word interpretation: GPT5.4 may narrowly limit a word's meaning or function in the target language and ignore other valid meanings or alternative functions the word can serve.

- Inconsistency: Inconsistent error detection: GPT5.4 may flag a repeated translation phrase as a mistranslation in one instance while ignoring the same issue in other occurrences.

- Ignoring contextual meaning: GPT5.4 may criticize the translation for adding meaning, even when that meaning is already implied or established by the surrounding context.

Below is a list of reasons we identified where GPT5.4 generated the false positives (examples in which GPT5 noted positive strength while human evaluators disagree):

- Over-praising weak word choices: GPT5.4 may praise a suboptimal word choice even when it dist

- Over-praising over-translation: GPT5.4 may praise an over-translated phrase even when it adds meaning, emphasis, or specificity that were not present in the source.
- Over-praising fancy translation choices: GPT5.4 may praise a polished or sophisticated translation choice even when it introduces added inaccuracy or distorts the source meaning.
- Over-praising poorly adopted word choices: GPT5.4 may praise a word choice that is not well adopted or natural in the target language, even when it sounds awkward or non-idiomatic.
- Over-praising idiomatic but inaccurate translations: GPT5.4 may praise an idiomatic translation even when it distorts the source meaning.

GPT5.4 appears to evaluate each translation segment across all error categories (that could be a strength), sometimes forcing comments even when a category is not relevant. Human trainees, by contrast, typically identify the most salient error category and leave unnecessary categories unaddressed (saliency bias). When an error could fall under multiple categories, humans focus on the root cause. Criticizing a single sentence across many categories can be overly harsh and unnecessary; it is unlikely that one sentence would contain every possible type of translation error.

In contrast, GPT-5.4 can analyze subtle and nuanced cases consistently and in detail without experiencing the fatigue that may affect

3.3 Arbitration Performance of DeepSeek

DeepSeek-V4 was used to arbitrate disagreements between GPT-5.4 and the human trainees. Appendix Table 5 summarizes DeepSeek's arbitration of disagreements on negative exemplars. After further review, the human evaluators determined that DeepSeek's decisions were correct in 6 of the 8 cases. Appendix Table 6 summarizes DeepSeek's arbitration of disagreements on positive exemplars. The human evaluators confirmed that DeepSeek's decisions were correct in both cases (2 of 2).

DeepSeek was also used to arbitrate disagreements between GPT5.4 and the two human evaluators. Appendix Table 7 lists DeepSeek arbitration towards disagreements on negative exemplars between human evaluators and GPT5.4. Based on the human evaluators' further investigation and final judgments, DeepSeek's arbitration was correct in 11 of the 15 cases. Appendix Table 8 lists DeepSeek arbitration towards disagreements on positive examples between human trainees and GPT5.4. Based on the human evaluators' further investigation and final judgments, DeepSeek's arbitration was correct in 3 of the 7 cases.

Figure 3 shows comparison of LLM deliberation and arbitration results with those of human trainees and evaluators.

Discussion and Conclusion

Melby (2012) defined TTT (Translation Turing Tests or Translation Technology Tests) using an operational approach: human, machine, or hybrid translation providers are expected to generate translation results that meet detailed specifications. ATA exam passages are typically written at “educated layperson level” and comes with a brief instruction of the intended audience (e.g., “for university students learning about the subject”). Other than a general readability level and a brief indication of the intended audience, the specification of ATA exam passages is relatively general.

General specifications allow greater room for human judgment. Unlike human graders in translation certification exams such as the ATA exam, who focus on formal assessment, human trainees may be less accurate but can still provide valuable opportunities for learning and practice. Because we compared GPT-5.4 with human trainees rather than professional graders, the human benchmark represents a less stringent standard. Therefore, the Turing test posed to GPT-5.4 is relatively easier to pass.

What we need most from training partners is availability, attention, the ability to be a “more knowledgeable other”, and productive friction that encourages critical, creative, and deeper reflection.

available and can provide a low-stress practice environment, since learners may feel less fear of embarrassment or negative judgment even when making serious mistakes. In addition, unlike human partners, LLMs would never experience fatigue.

We identify the following strengths of using LLMs as a training partner (guided by an MQM-style prompt):

- It can identify most major/critical errors identified by human trainees.
- It can reduce over-flagging of trivial (nitpicking) issues.
- It can detect and document a greater number of errors (without fatigue) and achieve good performance (F1 score = 0.875 based on our small experiment).
- With appropriate prompting, it can provide reasons and explanations for positive cases.
- It could generate alternative error types that encourage reflection and the human-LLM team to reach a more balanced judgment.

We identify the following weaknesses of using LLMs as a training partner (guided by an MQM-style prompt):

- It can overstate/inflate error severity.
- It can generate less accurate error types.
- It can generate a greater number of errors without proper saliency bias (identify the most salient error category while leave out secondary issues)
- LLMs, like human, can make poor judgment. The issues included misclassifying suboptimal but acceptable translations as unacceptable; incorrectly flagging meaning omissions or additions; attributing errors to the wrong linguistic element; requiring overly rigid syntax or structural alignment between source and target; relying too heavily on common translation choices or narrow word interpretations; applying inconsistent standards; ignoring contextual meaning; and over-praising weak, overly literal, overly embellished, or poorly adapted word choices, even when they introduced inaccuracies.

DeepSeek was used to arbitrate disagreements regarding the existence of translation errors between GPT-5.4 and both the human trainees and the human evaluators. Our findings suggest that incorporating a second LLM as an independent arbitrator can help resolve a substantial proportion of human-LLM disagreements, providing an additional layer of

assisted translator training.

In a peer-learning setting, stronger trainees can serve as a more knowledgeable other, helping their peers while also reinforcing their own learning by explaining their reasoning, answering questions, and comparing approaches. In our experiment, GPT-5.4 demonstrated the characteristics of a more knowledgeable other by identifying most of the errors detected by human trainees, as well as a greater number of subtle and nuanced errors. It also maintained a high F1 score, indicating both high precision and a low false negative rate. However, we would not conclude that GPT-5.4 definitively passes the translation Turing test. A Turing test is not simply about achieving a high F1 score; it asks whether humans can reliably distinguish GPT-5.4 from human trainees/evaluators based on relevant behaviors. Although we observed substantial behavioral similarity, we cannot establish indistinguishability.

Although our study was conducted on a relatively small dataset, the results provide encouraging evidence that a multi-agent approach to developing LLM-based training partners for translators is both feasible and promising. The combination of an MQM-guided reviewer, human deliberation, and a second LLM acting as an independent arbitrator demonstrates the potential to provide reliable feedback, encourage critical reflection, and improve judgment quality. Future work should validate this approach on larger datasets, additional language pairs, and more diverse translation domains to further assess its effectiveness and generalizability.

Acknowledgements

We sincerely thank Roanne Cheung, Jessie Doherty, and Zhitong Zou from the ATA-CLD translation practice group for their dedicated efforts in generating high-quality annotations and generously sharing their results, which were essential to this study's dataset.

References

American Translators Association – Community-Led Division. (n.d.). *Translation Practice Group*. ATA Divisions. <https://www.ata-divisions.org/CLD/resources/translation-practice-group/>

Bavaresco, A., Bernardi, R., Bertolazzi, L., Elliott, D., Fernández, R., Gatt, A., ... & Testoni, A. (2025, July).

Proceedings of the 17th Conference of the Association for Machine Translation in the Americas, LLMs instead of hu Québec City, Canada, August 31 - September 2, 2026. Volume 1: Research Track

study across 20 nlp evaluation tasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 238-255).

Feng, Z., Su, J., Zheng, J., Ren, J., Zhang, Y., Wu, J., ... & Liu, Z. (2025, July). M-MAD: Multidimensional multi-agent debate for advanced machine translation evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 7084-7107).

Fernandes, P., Deutsch, D., Finkelstein, M., Riley, P., Martins, A. F., Neubig, G., ... & Firat, O. (2023, December). The devil is in the errors: Leveraging large language models for fine-grained machine translation evaluation. In *Proceedings of the Eighth Conference on Machine Translation* (pp. 1066-1083).

Kocmi, T., & Federmann, C. (2023, December). GEMBA-MQM: Detecting translation quality error spans with GPT-4. In *Proceedings of the Eighth Conference on Machine Translation* (pp. 768-775).

Li, H., Dong, Q., Chen, J., Su, H., Zhou, Y., Ai, Q., ... & Liu, Y. (2024). LLMs-as-judges: a comprehensive survey on llm-based evaluation methods. *arXiv preprint arXiv:2412.05579*.

Lommel, A., Uszkoreit, H., & Burchardt, A. (2014). Multidimensional quality metrics (MQM): A framework for declaring and describing translation quality metrics. *Tradumática*, (12), 0455-463.

Lu, Q., Ding, L., Zhang, K., Zhang, J., & Tao, D. (2025, January). MQM-APE: toward high-quality error annotation predictors with automatic post-editing in LLM translation evaluators. In *Proceedings of the 31st International Conference on Computational*

Melby, A. (2012). Human and machine translation quality: Definable? Achievable? Desirable. In LACUS Forum (Vol. 39, pp. 1-29). Available online: www.ttt.org/melbyak [last accessed 18 August. 2026]

MQM Council. (2025). *MQM (Multidimensional Quality Metrics)*. <https://themqm.org/>

Negativity bias (2026, April 30th). In Wikipedia https://en.wikipedia.org/wiki/Negativity_bias

Szymanski, A., Ziems, N., Eicher-Miller, H. A., Li, T. J. J., Jiang, M., & Metoyer, R. A. (2025, March). Limitations of the llm-as-a-judge approach for evaluating llm outputs in expert knowledge tasks. In *Proceedings of the 30th international conference on intelligent user interfaces* (pp. 952-966).

Wang, G., Hu, J., & Ali, S. (2025). MAATS: A Multi-Agent Automated Translation System Based on MQM Evaluation. *arXiv preprint arXiv:2505.14848*.

Wang, R., Guo, J., Gao, C., Fan, G., Chong, C. Y., & Xia, X. (2025). Can llms replace human evaluators? an empirical study of llm-as-a-judge in software engineering. *Proceedings of the ACM on Software Engineering*, 2(ISSTA), 1955-1977.

Ye, J., Wang, Y., Huang, Y., Chen, D., Zhang, Q., Moniz, N., ... & Zhang, X. (2024). Justice or prejudice? quantifying biases in llm-as-a-judge. *arXiv preprint arXiv:2410.02736*.

Appendix

#	Source and Target Sentences	Human & GPT5 generated feedback
1	“We believe that <i>the world</i> is on the cusp of a new era...” “我们相信新的时代马上就要来临...”	Trainee: “the world” was not translated. (Severity: Minor) GPT5: omission (Severity: Minor) Same error type
2	“But now, it could find itself in the crosshairs of trade tensions.” “当今贸易 紧张局势 也在这里得到了格外强烈的 体现 ”	Trainee: “crosshairs of trade tensions: ➔ “紧张局势.....的体现” is a mis-translation. (Severity: Minor) GPT5: mis-translation (Severity: Major) Same error type
3	“Asia’s integration has come about through commercial pragmatism rather than political alignment.” “亚洲的融入是通过商业务实性，而不是政治 取向 一致性。”	Trainee: The phrase “取向” is a mis-translation. (Severity: Minor) GPT5: Incorrect term choice (Severity: Major) Both error types are acceptable. Mistranslation is “meaning error” while the term choice “取向” is incorrect.
4	“Can that continue if trade tensions rise?” “亚洲还可以维持现状吗？”	Trainee: "That" in the original text refers more to the strategy of commercial pragmatism, so it doesn't mean “maintaining the status quo.” (“维持现状”) (Severity: Major) GPT5: Mis-translation (Severity: Major)

meaning error.

- 5 “Asia excels in manufacturing new technology in four areas in particular”
“亚洲在如下四项新型科技制造方面超群拔类”
Trainee: The original text only discussed technology; translating it as “科技” adds meaning not present in the original text. (Severity: Minor)
GPT5: incorrect term choice; vague and non-standard (Severity: Major)
Both error types are acceptable: minor addition versus incorrect term choice.
- 6 “consumer electronics, industrial electronics, electric vehicles, and semiconductors.”
“消费电子产品、工业电子产品、电动汽车、半导体。”
Trainee: The last item in a list must be connected with a “和”. (Severity: No error)
GPT5: No problem detected
- 7 “Asia has more than 40 percent of the revenues”
“亚洲占了超过百分之四十的收益”
Trainee: Why “40% “ should be written in Chinese text “百分之四十”. Are there any specific formatting rules? (Severity: No error)
GPT5: No problem detected
- 8 “But now, aging is fiercest in the high-productivity economies of the Pacific Rim.”
“但是现在人口老龄化在环太平洋地区生产力最高的经济体最为激烈。”
Trainee: “high-productivity” does not mean the highest productivity. (Severity: Minor)
GPT5: style problem: unnatural expression (Severity: Minor)
Trainee-generated error type is more reasonable. It should be a mistranslation instead of an unnatural expression.
Trainee: “老龄化 ... 最为激烈 “ is an unnatural expression (Severity: Minor)
GPT5: word choice nuance (Severity: Minor)
Both error types are acceptable: unnatural expression vs. word choice nuance.
- 9 “Asia has the people to fuel growth, but 90 percent of the expected increase in Asia’s non-farm workforces by 2050 will be where productivity is relatively low.”
“亚洲拥有发展所需的人口，但2050年前亚洲非务农工作力的预期增长的90%将处于生产力相对低的地区。”
Trainee: “non-farm workforces” → “非务农工作力” is a mistranslation (Severity: Minor)
GPT5: bad term choice (Severity: Major)
Both error types are acceptable: unnatural expression vs. word choice nuance.
- 10 “Can Asia shift value chains to where the workers are and raise productivity everywhere?”
“Can Asia shift value chains to where the workers are and raise productivity everywhere?”
Trainee: “everywhere” was mistranslated: should be “地区”, not “方面”. (Severity: Minor)
GPT5: Mis-translation (Severity: Major)
Both error types are acceptable: unnatural expression vs. overly literal phrasing.

	在的地区，并在所有方面都提高生产力呢？”	meaning error.
11	“But now, it could find itself in the crosshairs of trade tensions.” “然而目前亚洲也发现自己在紧张的贸易关系中成了靶子。”	Trainee: Translation for “find itself in” (“发现自己……”) is too literal (Severity: Minor) GPT5: Mis-translation (Severity: Major) Disagreement on error types: “too literal” versus mistranslation (meaning error). “too literal” is more reasonable.
12	“But the value created by tech is moving away from hardware and into new transversal technologies. So, the question is whether Asia can reinvent itself as a tech creator as well as a tech manufacturer.” “然而创造价值的科技产业正逐渐从硬体转变为新式横向科技，因此问题在于亚洲是否能在科技创新和科技制造业这些领域进行自我重塑。”	Trainee: The translation of “hardware” (“硬体”) is an expression used in Taiwan. But it is not a conventional expression in mainland China. (Severity: Minor) GPT5: No problem detected.
13	“Asia has largely been in a demographic sweet spot with large pools of young workers and surging productivity in some economies.” “亚洲有大批的年轻工人，当中有些经济体的生产力又有所提高，这些在很大程度上都是最有利的人口特征。”	Trainee: “surging” means increasing significant. The degree of improvement was not translated properly. (Severity: Major) GPT5: Unnatural/awkward expression (Severity: Minor) Disagreement on error type: The trainee is correct. The issue is a mistranslation rather than an unnatural expression.
14	“We believe that the world is on the cusp of a new era with very different underlying forces than those that shaped our world over the past 30 years.” “我们认为全球正处在一个新时代的风口浪尖，各种深层力量相互角逐。而这几股力量与过去30年间塑造世界的力量截然不同。	Trainee: “相互角逐” is added meaning that does not exist in the source sentence. (Severity: Minor) GPT5: meaning addition (Severity: Minor) Same error type
15	“Asia will be at the nexus of these forces and will experience heightened versions of the trends affecting the rest of the world.” “亚洲则将处于这几股力量的角力旋涡中，更深刻地体会这种发展趋势，进而影响着世界其他地方。	Trainee: Mis-translation on “nexus” (center of things) because “旋涡” carries the meaning of struggling. (Severity: Major) GPT5: Mis-translation (Severity: Major) Same error type

		the world" is a mis-translation. (Severity: Major) GPT5: Mis-translation (Severity: Major) Same error type
16	“Asia is the world’s trade crossroads. It’s home to 18 of the 20 fastest-growing trade corridors and 13 of the 20 largest.” “亚洲是全球贸易的岔道口。这里盘踞着全球20条增长最快的贸易走廊中的18条，20条规模最大的贸易走廊中的13条。”	Trainee: Crossroad-> “岔道口” is a mis-translation (Severity: Major) GPT5: Incorrect term choice (mis-translation) (Severity: Major) Both error types are acceptable: Incorrect term choice leads to mistranslation. Trainee: Need a connection word after the comma. (severity: Minor) GPT5: Grammar/Punctuation (severity: Minor) Equivalent error type
17	“But now, it could find itself in the crosshairs of trade tensions.” “但如今，该地区却被卷入盘根错节中。”	Trainee: Mis-translate “in the crosshairs” (Severity: Major) GPT5: Over-translation (Severity: Major) Equivalent error type: both point to meaning error.
18	“So, the question is whether Asia can reinvent itself as a tech creator as well as a tech manufacturer.” “问题就变成亚洲是否能够不仅充当技术生产者，还能将自己重塑成为一个技	Trainee: “reinvent itself” means reinvent for both technology creator and manufacturer. So the verb should cover both. (Severity: Minor) GPT5: No problem detected.
19	“Asia has largely been in a demographic sweet spot with large pools of young workers and surging productivity in some economies.” “亚洲从很大程度上一直拥有人口红利，在某些经济体中拥有大量的年轻工人和高涨的生产力。”	Trainee: “Demographic sweet spot” translated as “人口红利”, is a mis-translation. GPT5: No problem detected. Trainee: “在某些经济体中拥有” is unnatural Chinese expression. (Severity: Minor) GPT5: Unnatural flow (Severity: Minor) Equivalent error type
20	“But now, aging is fiercest in the high-productivity economies of the Pacific Rim.” “但如今，环太平洋的高效经济体中的老龄化问题也最为严重。”	Trainee: Added meaning with the word “也” (“as well”). (Severity: Minor) GPT5: added meaning (Severity: Minor) Same error type
21	“but 90 percent of the expected increase in Asia’s non-farm workforces by 2050 will be where productivity is relatively low.” “但在截至2050年的亚洲非农劳动力预	Trainee: “90%的增长” (90% increase) is redundant. “90%” is sufficient. (Severity: Minor) GPT5: unnecessary redundancy reduces fluency (Severity: Minor) Same error type

力相对较低的地区。”

- 22 “Asia will be at the nexus of these forces and will experience heightened versions of the trends affecting the rest of the world..”
“亚洲将成为这些力量的纽带，在影响世界其他地区的趋势上，也将经历更高版本的转机。”
- 23 “It’s home to 18 of the 20 fastest-growing trade corridors and 13 of the 20 largest.”
“20个增速最快贸易通道中的18个位于亚洲，20个最大贸易通道中的13个同样起源于亚洲。”
- 24 “But now, it could find itself in the crosshairs of trade tensions.”
“但如今，亚洲发现自己正处于贸易压力的十字准线上。”
- 25 “Asia’s integration has come about through commercial pragmatism rather than political alignment.”
“亚洲的整合来自于商业的实用主义而不是政策导向。”
- 26 “Can that continue if trade tensions rise?”
“如果贸易压力增高，这种整合还会继续吗？。”
- 27 “Asia excels in manufacturing new technology in four areas in particular: consumer electronics, industrial electronics, electric vehicles, and semiconductors”
“亚洲在新科技生产方面遥遥领先，尤其在消费电子产品、工业电子、电动汽车和半导体四个方面。”

Trainee: “Trend”, translated as “转机”, is a mis-translation.
(Severity: Critical)
GPT5: Mis-translation (Severity: Critical)
Same error type

Trainee: “条” is a more appropriate quantity word.
(Severity: Minor)
GPT5: No problem detected.

Trainee : “亚洲” was repeated, which is unnatural. It is better to use a different word “亚洲” (“this place”).
(Severity: Minor)
GPT5: No problem detected.

Trainee: “Tension”, translated as “压力”, is a mis-translation.
(Severity: Major)
GPT5: Mis-translation (Severity: Major)
Same error type
Trainee: “crosshairs”, directly translated as “十字准线”, is awkward in the target language.
(Severity: Minor)
GPT5: Incorrect term choice (Severity: Major)
Both error types are acceptable: incorrect term choice leads to awkwardness in the target language

Trainee: “Political alignment”, translated as “政策导向” is a mis-translation.
(Severity: Major)
GPT5: Mis-translation (Major)
Same error types

Trainee: “trade tension”, translated as “贸易压力”, is a mis-translation.
(Severity: Major)
GPT5: Mis-translation (Major)
Same error types

Human: The source text only mentioned “technology” (“技术”). In contrast, the target text phrase “科技” covers both “science” and “technology”. Therefore, there is added meaning.
(Severity: Minor)
GPT5: No problem detected.

reinvent itself as a tech creator as well as a tech manufacturer.”	clause used the verb “作为”, it would be better to use the same verb in the second clause as well.
“如此一来, 问题就变成亚洲是否能不	(Severity: Minor)
仅作为技术生产者, 也是技术创新者进行再次创新。”	GPT5: “Re-invent itself”→ “进行再次创新” is a major mistranslation. The structure “不仅作为技术生产者, 也是技术创新者进行再次创新” is syntactically awkward.
	(Severity: Major)
	Disagreement on error type: there is a major mistranslation that is identified by GPT5.

Table 1: Negative exemplars (Human trainees versus GPT5). Agreement rate= 28/36 (77.8%). Agreement refers only to whether an error exists, not to its type or severity.

#	Source and Target Sentences	Human & GPT5 generated feedback
1	“We believe that the world is on the cusp of a new era with very different underlying forces than those that shaped our world over the past 30 years.” “我们相信新的时代马上就要来临, 这个新时代深层暗涌的力量与在过去30年促成世界格局的力量大不相同。”	Trainee: “the world is on the cusp of a new era” → “促成世界格局”is excellent. GPT5: It is an appropriate terminology choice.
2	“Asia is the world’s trade crossroads. It’s home to 18 of the 20 fastest-growing trade corridors and 13 of the 20 largest.” “亚洲是世界的贸易交叉衔接地, 坐拥20条发展最快的贸易通道中的18条, 和20条最大的贸易通道中的13条。”	Trainee: “crossroads” → “交叉衔接地”is excellent translation. GPT5: It is a highly accurate translation to convey faithful meaning. Trainee: “It is home to...” → “坐拥” is a good word to use in the context. GPT5: strength in style; slightly elevated and fluent style that avoids rigid literalness while remaining accurate.
3	“So, the question is whether Asia can reinvent itself as a tech creator as well as a tech manufacturer.” “因此, 问题在于亚洲是否能向兼顾科技创造和科技制造转型。”	Trainee: The phrase “兼顾” is well selected. GPT5: strength in style; avoids awkward literalism, making the sentence feel native rather than translated.
4	“We believe that the world is on the cusp of a new era with very different underlying forces than those that shaped our world over the past 30 years.” “我们相信世界正处于新时代的开端, 与过去30年相比, 影响世界的根本势力已变得非常不同。”	Trainee: “underlying force”→” 根本势力” is a good target language expression. GPT5: a highly accurate terminology. It conveys structural or fundamental drivers rather than superficial changes, aligning well with the analytical tone of the original text.
5	“Asia’s integration has come about through commercial pragmatism rather than political alignment. Can that continue if trade tensions rise?” “亚洲一体化通过商业实用主义而非政治结盟所产生的。一旦贸易紧张局势升级, 一体化的进程还能继续下去吗?”	Trainee: “一体化的进程还能继续下去吗?” is a smooth sentence structure. GPT5: Good at following conventional linguistic convention, and the idiomatic and engaging translation reads smoothly and naturally.
6	“So, the question is whether Asia can reinvent itself as a tech creator as well as a tech manufacturer.” “因此问题在于亚洲是否能在科技创新和科技制造业这些领域进行自我重	Trainee: “reinvent itself”→ “自我重塑” is a good target phrase choice. GPT5: accurate translation captures the idea of transformation rather than mere

- 7 “We believe that the world is on the cusp of a new era with very different underlying forces than those that shaped our world over the past 30 years.”
The translated target text sentence:
“我们认为全球正处在一个新时代的风口浪尖，各种深层力量相互角逐。而这几股力量与过去30年间塑造世界的力量截然不同。”
- 8 “Asia is the world’s trade crossroads. It’s home to 18 of the 20 fastest-growing trade corridors and 13 of the 20 largest.”
“亚洲是全球贸易的岔道口。这里盘踞着全球20条增长最快的贸易走廊中的18条，20条规模最大的贸易走廊中的13条。”
- 9 “Asia’s integration has come about through commercial pragmatism rather than political alignment.”
“亚洲的整合仰赖的是商业实用主义，而非政治联盟。”
- 10 “Can that continue if trade tensions rise?”
“这条道路还行得通吗?”
- 11 “In these places, Asia has more than 40 percent of the revenues, R&D, and patents.”
“在这些领域，亚洲坐拥超过40%的利润、研发和专利。”
- 12 “Asia has the people to fuel growth, but 90 percent of the expected increase in Asia’s non-farm workforces by 2050 will be where productivity is relatively low.”
“亚洲虽然拥有人力来推动增长，但在截至2050年的亚洲非农劳动力预期增长中，其中90%的增长都来自生产力相对较低的地区。”
- 13 “We believe that the world is on the cusp of a new era with very different underlying forces than those that shaped our world over the past 30 years.”
“人们认为相较于过去30年塑形世界的力量，新时代截然不同的潜在力量，将世界推向风口浪尖。”
- Trainee: “the world is on the cusp of a new era” → “新时代的风口浪尖” is a vivid translation.
GPT5: add rhetorical strength and make the sentence vivid, while still staying close to the original meaning. It does not feel mechanical or overly literal.
- Trainee: “home to” → “盘踞” is a good word choice.
GPT5: No comment on the word choice.
- Trainee: “come about through” → “仰赖” is a good word choice.
GPT5: The causal nuance of “has come about through... rather than...” is accurately captured by “仰赖的是...而非...”. It is correct and idiomatic in Chinese.
- Trainee: The target sentence follows the source sentence and is very natural.
GPT5: The core meaning of the source sentence is well preserved. The style is the strongest aspect. The phrasing is natural, idiomatic, and fluent in Chinese. It avoids overly technical or literal phrasing and instead uses a familiar metaphor (“道路”) that resonates well in Chinese discourse.
- Trainee: has → “坐拥” is a good word choice.
GPT5: it is particularly effective in terms of style: “坐拥” adds a sense of possession and prominence that reads naturally and fluently in Chinese, avoiding a stiff or overly literal rendering. It enhances readability while preserving the emphasis of the original sentence.
- Trainee: “其中90%的增长” is a good translation.
GPT 5.4: the style is strong: creates a smooth, idiomatic flow rather than a rigid, word-for-word rendering. It reads naturally and fluently in Chinese
- Trainee: “on the cusp of” → “推向风口浪尖” is a vivid translation.
GPT5: strong style (fluency and naturalness): effectively convey a sense of urgency and transformation, making the sentence engaging and vivid rather than stiff or overly literal.

	and will experience heightened versions of the trends affecting the rest of the world.” “亚洲将成为这些力量的纽带，在影响世界其他地区的趋势上，也将经历更高版本的转机。”	GPT5: the term “纽带” is an appropriate and understandable choice for a general audience when translating “nexus,” avoiding overly technical or obscure wording.
15	“Asia is the world’s trade crossroads. It’s home to 18 of the 20 fastest-growing trade corridors and 13 of the 20 largest.” “亚洲是世界贸易枢纽。20个增速最快贸易通道中的18个位于亚洲，20个最大贸易通道中的13个同样起源于亚洲。”	Trainee: “trade crossroads”→“贸易枢纽” is an excellent target phrase choice. GPT5: Key economic and trade-related terms are well chosen. “贸易枢纽” appropriately reflects “trade crossroads,” aligning with standard usage in economic and policy contexts.
16	“Asia excels in manufacturing new technology in four areas in particular: consumer electronics, industrial electronics, electric vehicles, and semiconductors.” “亚洲在新科技生产方面遥遥领先，尤其在消费电子产品、工业电子、电动汽车和半导体四个方面。”	Trainee: “excel”→“遥遥领先” is an excellent choice. GPT5: No comment on the word choice.
17	“But the value created by tech is moving away from hardware and into new transversal technologies.” “然而技术产生的价值正从硬件生产转移到新兴交叉科技。”	Trainee: “transversal technologies”→“交叉科技” is an excellent target phrase choice. GPT5: The rendering of “transversal technologies” as “交叉科技” is particularly strong. It reflects an understanding of the concept as cross-domain or interdisciplinary technologies, which is consistent with how such ideas are expressed in Chinese technical and business contexts.
18	“But now, aging is fiercest in the high-productivity economies of the Pacific Rim.” “可是如今，老龄化在太平洋地区高生产力的经济体中成为最棘手的问题。”	Trainee: “fiercest”→“最棘手的问题” is an very interesting translation. GPT5: The core meaning of the source sentence is well preserved. “aging is fiercest” is effectively rendered as “老龄化...成为最棘手的问题,” which captures the intensity and seriousness of the demographic challenge, even though it is slightly interpretive. The relationship between aging and “high-productivity economies of the Pacific Rim” is clearly and correctly conveyed.

Table 2: Positive exemplars (Human trainees versus GPT5). Agreement rate 16/18 (88.8%). Agreement refers only to whether a positive translation exists. Human trainees’ comments tend to be shorter and judgment-based while GPT5 comments are more analytical and evaluative.

#	Source and Target Sentences	GPT5 generated feedback (highlighted are false negatives identified by humans)
1	“We believe that <i>the world</i> is on the cusp of a new era...” “我们相信新的时代马上就要来临...”	Accuracy/mistranslation issue: “on the cusp of” →“马上就要来临” [Human evaluators’ disagreement: The translation is acceptable as long as it reflects the sentiment of “change is about to happen” or “a new era is about to start”.] Omission: “the world” is omitted. Style: slightly informal; less precise tone.

		Mistranslation: “in the crosshairs of trade tensions”→ “得到了格外强烈的体现”. Omission: “could find itself” Addition: “也” [Human evaluators’ disagreement: We don’t think it’s an addition here. 也 isn’t “also” all the time. Here it looks like 也 is used to continue on the context of the previous sentence (Asia is centre of trade paths, it may now be the centre of trade tensions).]
2	“But now, it could find itself in the crosshairs of trade tensions.” “当今贸易 紧张局势 也在这里得到了格外强烈的 体现”	Style: Unnatural / awkward expression Awkward referential expression (minor): “这里” is vague and weak.
3	“Asia’s integration has come about through commercial pragmatism rather than political alignment.” “亚洲的融入是通过商业务实性，而不是政治 取向 一致性。”	Mistranslation: integration”→ “融入” Terminology error: “political alignment”→ “政治取向一致性” Style – Unnatural / awkward expression: : “是通过商业务实性，而不是政治取向一致性” is very literal and stiff. Linguistic convention issue: “通过商业务实性” is grammatically acceptable but unnatural. [Human evaluators’ disagreement: 基于/源于 are better but 通过 isn’t wrong and still flies, We wouldn’t flag this as an error.]
4	“Can that continue if trade tensions rise?” “亚洲还可以维持现状吗？”	Omission: “if trade tensions rise”. The LLM did not go back and check the context. [Human evaluators’ disagreement: There is no omission here. The translation of “Can that continue if trade tensions rise?” is broken down into 2 clauses in Chinese separated by a comma, the translation is “如果贸易紧张局势加剧，亚洲还可以维持现状吗？” . The alleged omission is placed before the comma, the LLM simply missed it. Translation is not simple mapping, sometimes one sentence in the source is broken down into several in the target, sometimes several clauses can be joined together to form one wholesome sentence in the translation.] Mistranslation: “that continue” is rendered as “维持现状” Style (Slightly awkward phrasing): “that continue” → “维持现状” [Human evaluators’ disagreement: the mistranslation seems to be caused by a misunderstanding of the source sentence and it is not the word “continue”.]
5	“Asia excels in manufacturing new technology in four areas in particular” “亚洲在如下四项新型科技 制造方面超群拔类”	Accuracy issue: “如下四项” (the following four items) Terminology issue: “manufacturing new technology”→ “新型科技制造” Style issue: “如下四项” is slightly formal/awkward, not natural. [Human evaluators’ disagreement: 以下 sounds

		that it should not be counted as an error.] Style issue: “超群拔类” is overly literary and formal, not consistent with the analytical tone of the source text.
6	“consumer electronics, industrial electronics, electric vehicles, and semiconductors.” “消费电子产品、工业电子产品、电动汽车、半导体。”	GPT5 did not find any problem.
7	“Asia has more than 40 percent of the revenues, R&D, and patents.” “亚洲占了超过百分之四十的收益、研发以及专利。”	Terminology issue: “revenues” → “收益” Style issue: “占了超过百分之四十的...” is slightly less natural.
8	“But now, aging is fiercest in the high-productivity economies of the Pacific Rim.” “但是现在人口老龄化在环太平洋地区生产力最高的经济体最为激烈。”	Linguistic convention issue: Missing structural particles (“的”“中”) in “在环太平洋地区生产力最高的经济体最为激烈”. Style issue: “生产力最高的经济体” is an unnatural expression that is overly literal. Style (word choice nuance): “最为激烈” used for “aging”
9	“Asia has the people to fuel growth, but 90 percent of the expected increase in Asia’s non-farm workforces by 2050 will be where productivity is relatively low.” “亚洲拥有发展所需的人口，但2050年前亚洲非务农工作力的预期增长的90%将处于生产力相对低的地区。”	Accuracy issue: “will be where productivity is relatively low” → “将处于生产力相对低的地区” Terminology issue: “non-farm workforces” → “非务农工作力” [Human evaluators’ disagreement: 非务农 and 非农业 both should work. It is not an error]. Linguistic convention issue: “预期增长的90%将处于...” structure is grammatically acceptable but slightly clunky due to repeated “的” and long modifier stacking. That is awkward in Chinese. Style issue: “发展所需的人口” is slightly stiff.
10	“Can Asia shift value chains to where the workers are and raise productivity everywhere?” “亚洲是否可以将价值链转移到工人所在的地区，并在所有方面都提高生产力呢？”	Accuracy issue: “raise productivity everywhere” → “在所有方面都提高生产力” Style issue: “是否可以将价值链转移到...” is slightly literal phrasing
11	“But now, it could find itself in the crosshairs of trade tensions.” “然而目前亚洲也发现自己紧张的贸易关系中成了靶子。”	Accuracy issue: “trade tensions” → “紧张的贸易关系”. The source refers specifically to trade tensions/conflicts, which implies geopolitical friction or disputes. The translation weakens the meaning and changes the nuance. [Human evaluators’ disagreement: The turn of phrase is a bit awkward but it for Machine Translation in the Americas,

		over translation for the exam.] Accuracy issue: “find itself in the crosshairs” is rendered as “发现自己...成了靶子”. Omission/meaning shift: The modal “could” (indicating possibility) is not clearly conveyed. Grammar: The phrase “紧张的贸易关系中成了靶子” is grammatically awkward. And there are two related style problems.
12	“But the value created by tech is moving away from hardware and into new transversal technologies. So, the question is whether Asia can reinvent itself as a tech creator as well as a tech manufacturer.” “然而创造价值的科技产业正逐渐从硬体转变为新式横向科技，因此问题在于亚洲是否能在科技创新和科技制造业这些领域进行自我重塑。”	Accuracy issue: changes the subject from “value created” to “the tech industry”. Incorrect term: “横向科技” is not a standard Chinese rendering. Accuracy issue: “tech creator”→ “科技创新” Entire sentence structure is an over-expansion not present in source. [Human evaluators’ disagreement: The source words “creator” and “manufacturer” are stylistic choice. The “base” meaning of the segment is “create/invent tech” and “manufacture tech”, which is preserved. We would not expect the parts of speech in the translation to align perfectly with the source - syntax works differently in different languages.]
13	“Asia has largely been in a demographic sweet spot with large pools of young workers and surging productivity in some economies.” “亚洲有大批的年轻工人，当中有些经济体的生产力又有所提高，这些在很大程度上都是最有利的人口特征。”	Unnatural collocation: “新式横向科技”. Rigid translation: “因此问题在于” Accuracy issue: “demographic sweet spot” → “最有利的人口特征”. Accuracy issue: 这些在很大程度上都是” (these largely are) introduces an explanatory structure not present in the source sentence. [Human evaluators’ disagreement: We think 很大程度上 matches “largely” in the source text, what’s changed is the sentence structure - but the meaning of “largely” is preserved. This is not an error.] Unnatural flow: “这些在很大程度上都是最有利的人口特征” is awkward and not idiomatic in Chinese.
14	“We believe that the world is on the cusp of a new era with very different underlying forces than those that shaped our world over the past 30 years.” “我们认为全球正处在一个新时代的风口浪尖，各种深层力量相互角逐。而这几股力量与过去30年间塑造世界的力量截然不同。”	Over-translation/Addition: “On the cusp”→“风口浪尖” Style: “新时代的风口浪尖” is rhetorically strong and journalistic; unnatural or inappropriate metaphorical tone Linguistic conventions: The translation splits the original single sentence into two sentences
15	“Asia will be at the nexus of these forces and will in turn influence other places.”	Accuracy issue: “进而影响着世界其他地方” means “and in turn influence other places”

	the trends affecting the rest of the world.” “亚洲则将处于这几股力量的角力旋涡中，更深刻地体会这种发展趋势，进而影响着世界其他地方。”	Accuracy issue: “nexus”→ “角力旋涡” Style: Unnatural / Overly Dramatic Expression. “角力旋涡” and “更深刻地体会” Awkward expression: 更深刻地体会这种发展趋势”
16	“Asia is the world’s trade crossroads. It’s home to 18 of the 20 fastest-growing trade corridors and 13 of the 20 largest.” “亚洲是全球贸易的岔道口。这里盘踞着全球20条增长最快的贸易走廊中的18条，20条规模最大的贸易走廊中的13条。”	Accuracy issue: “home to”→“盘踞着” Incorrect term choice: “crossroads”->“岔道口” Style: the second sentence is an unnatural expression. Linguistic conventions: lack parallel structure clarity.
17	“But now, it could find itself in the crosshairs of trade tensions.” “但如今，该地区却被卷入盘根错节中。”	Accuracy issue: The phrase “盘根错节的” (complicated) introduces an additional semantic element that does not exist in the source text. Style: The phrase “盘根错节的贸易紧张局势” sounds stylistically heavy and somewhat literary compared with the neutral analytical tone of the source text.
18	“So, the question is whether Asia can reinvent itself as a tech creator as well as a tech manufacturer.” “问题就变成亚洲是否能够不仅充当技术生产者，还能将自己重塑成为一个技术创新者。”	Accuracy issue: “So, the question is”→ “问题就变成” Incorrect term choice: “tech manufacture”→ “技术生产者” Style: The sentence structure is somewhat heavy and verbose compared with standard Chinese economic writing.
19	“Asia has largely been in a demographic sweet spot with large pools of young workers and surging productivity in some economies.” “亚洲从很大程度上一直拥有人口红利，在某些经济体中拥有大量的年轻工人和高涨的生产力。”	Accuracy issue: “surging productivity”→“高涨的生产力” Incorrect term choice: “productivity”→“生产力”: “productivity” is translated as “生产力” instead of the more precise economic term “生产率”. [Human evaluators’ disagreement: We don’t think 生产力 is wrong. Here the paragraph is talking about demographics and labour, and with young workers it naturally factors that the output capability is high. “Productivity” in this context refers more likely to the ability to produce, rather than the ratio of material output over input.] Style: “从很大程度上一直拥有” is somewhat awkward and verbose. Style: “大量的年轻工人” is understandable but slightly literal.
20	“But now, aging is fiercest in the high-productivity economies of the Pacific Rim.”	Accuracy issue: The word “也” (also) is added in the translation, but the source sentence does not contain the meaning “also.” [Human evaluators’ disagreement: We

	老龄化问题也最为严重。”	one is to back-translate this into English. However, there is an implied contrast between this sentence and the previous sentence in the source. “Asia population situation good. (But) some countries have population problems”. 也 is used as an emphasis for the aging issue, which is the point of contrast. Nothing is wrong]. Incorrect terminology: “high productivity economics”→ “高效经济体” Style: The structure “环太平洋的高效经济体中的老龄化问题” is somewhat heavy and indirect compared with natural Chinese. Accuracy issue: “will be where productivity is relatively low” → “来自生产力相对较低的地区”. The source sentence means the workforce increase will occur in low-productivity areas, not that it comes from those areas. “来自” introduces an unintended source meaning.
21	“but 90 percent of the expected increase in Asia’s non-farm workforces by 2050 will be where productivity is relatively low.” “但在截至2050年的亚洲非农劳动力预期增长中，其中90%的增长都来自生产力相对较低的地区。”	[Human evaluators’ disagreement: We think this is fine and would not cause misunderstanding on the part of the reader. The core message refers to where labor growth is, the reader would think about location rather than where the people are born (“from”).] Accuracy issue: “by 2050”→ “截至2050年” (slightly alters the meaning) Style: Repetition in “预期增长中，其中90%的增长” creates redundancy and slightly reduces fluency. Style: “亚洲非农劳动力预期增长中” is awkward phrasing
22	“Asia will be at the nexus of these forces and will experience heightened versions of the trends affecting the rest of the world..” “亚洲将成为这些力量的纽带，在影响世界其他地区的趋势上，也将经历更高版本的转机。”	Incorrect term: “Nexus” → “纽带” Structure: “在影响世界其他地区的趋势上” is grammatically awkward. Accuracy issue: “heightened versions”→ “更高版本”. Accuracy issue: “trends”→ “转机” Style: Overall phrasing is unnatural and overly literal.
23	“It’s home to 18 of the 20 fastest-growing trade corridors and 13 of the 20 largest.” “20个增速最快贸易通道中的18个位于亚洲，20个最大贸易通道中的13个同样起源于亚洲。”	Inaccuracy: “home to”→ “起源于亚洲” Incorrect term choice: “trade corridors”→ “贸易通道” Style: Use of “同样” is awkward/unnecessary emphasis
24	“But now, it could find itself in the crosshairs of trade tensions.” “但如今，亚洲发现自己正处于贸易压力的十字准线上。”	Inaccuracy: “trade tensions” → “贸易压力” Incorrect term choice: “crosshairs” → “十字准线” Style: “find itself”→ “发现自己正处于” Style: “贸易压力的十字准线上” is stylistically unnatural and reads like a word-for-word rendering rather than

25	“Asia’s integration has come about through commercial pragmatism rather than political alignment.” “亚洲的整合来自于商业的实用主义而不是政策导向。”	Accuracy issue: “political alignment” is translated as “政策导向” Style: “来自于商业的实用主义” is slightly stiff and literal Suboptimal term choice: “商业的实用主义”
26	“Can that continue if trade tensions rise?” “如果贸易压力增高，这种整合还会继续吗？。”	Accuracy issue: “trade tensions” → “贸易压力” Style: “这种整合还会继续吗” is understandable but slightly stiff and literal Accuracy issue: “excels in” → “遥遥领先” Style: “新科技生产” is not a natural Chinese collocation.
27	“Asia excels in manufacturing new technology in four areas in particular: consumer electronics, industrial electronics, electric vehicles, and semiconductors” “亚洲在新科技生产方面遥遥领先，尤其在消费电子产品、工业电子、电动汽车和半导体四个方面。”	Accuracy issue: “consumer electronics” → “消费电子产品” Adding “产品” slightly expands the meaning. While common in Chinese, it is not strictly present in the source text. [Human evaluators’ disagreement: Such term exists and is used in Chinese. It is also the term of choice in 3 of the 4 translations (the other one being 消费者电子产品), but the MQM did not pick out the term until version 4.] Accuracy issue: “reinvent itself” → “进行再次创新” Addition: The phrase “如此一来” (“thus/therefore”) is added but does not appear in the source sentence. [Human evaluators’ disagreement: We don’t think the translation is wrong. It is likely serves the function “so” serves in the source. “Given the above context” is the idea “so” and 如此一来 express. The translation is appropriate.]
28	“So, the question is whether Asia can reinvent itself as a tech creator as well as a tech manufacturer.” “如此一来，问题就变成亚洲是否能不仅作为技术生产者，也是技术创新者进行再次创新。”	Style: The structure “不仅作为技术生产者，也是技术创新者进行再次创新” is syntactically awkward. Style: The phrase “问题就变成” is slightly heavy and literal. Terminology: “tech manufacturer” → “技术生产者”

Table 3: True and false negative exemplars generated by GPT5. Percentage of false negatives found by human evaluators is 15/92 (16.3%).

#	Source and Target Sentence	GPT5 generated feedback (highlighted are false positives identified by humans)
1	“We believe that the world is on the cusp of a new era with very different underlying forces than those that shaped our world over the past 30 years.” “我们相信新的时代马上就要来临，这个新时代深层暗涌的力量与在过去30年促成世界格局的力量大不相同。 “Asia is the world’s trade crossroads. It’s	“that shaped our world” → “促成世界格局”: The choice of “世界格局” for “our world” is appropriate in this context, as it reflects a macro-level, analytical perspective consistent with the tone of the source.

	corridors and 13 of the 20 largest.” “亚洲是世界的贸易交叉衔接地，坐拥20条发展最快的贸易通道中的18条，和20条最大的贸易通道中的13条。”	meaning is conveyed very faithfully. “trade crossroads” is effectively rendered as “贸易交叉衔接地,” capturing the idea of Asia as a central hub.
3	“So, the question is whether Asia can reinvent itself as a tech creator as well as a tech manufacturer.” “因此，问题在于亚洲是否能向兼顾科技创造和科技制造转型。”	“as well as”→“兼顾：The phrase “向兼顾...转型” is concise and avoids awkward literalism, making the sentence feel native rather than translated. “underlying force” →“根本势力”: The phrase “very different underlying forces” is effectively translated as “根本势力已变得非常不同,” preserving the conceptual emphasis on foundational drivers of change. There is no omission or distortion of meaning. [Human evaluators’ disagreement: Translating “underlying” as 根本 is not wrong enough to warrant a point deduction, but 根本 can mean “source/origin” whereas the meaning of “underlying” is more like “deep down/deep within”. Therefore, this is a minor false positive.]
4	“We believe that the world is on the cusp of a new era with very different underlying forces than those that shaped our world over the past 30 years.” “我们相信世界正处于新时代的开端，与过去30年相比，影响世界的根本势力已变得非常不同。”	
5	“Asia’s integration has come about through commercial pragmatism rather than political alignment. Can that continue if trade tensions rise?” “亚洲一体化通过商业实用主义而非政治结盟所产生的。一旦贸易紧张局势升级，一体化的进程还能继续下去吗？”	“Can that (integration) continue if trade tensions rise”→“一体化的进程还能继续下去吗？”: Linguistic Conventions The sentence is grammatically sound, with correct punctuation and no typographical issues. The use of the question mark and the conditional structure (“一旦...还能...吗”) is properly handled. Style The translation reads smoothly and naturally in Chinese, particularly the second sentence, which is idiomatic and engaging. The rhetorical question structure is well rendered and aligns with native Chinese discourse patterns.
6	“So, the question is whether Asia can reinvent itself as a tech creator as well as a tech manufacturer.” “因此问题在于亚洲是否能在科技创新和科技制造业这些领域进行自我重塑。”	“reinvent itself”→“自我重塑”: the notion of “reinvent itself” is effectively rendered as “进行自我重塑,” which captures the idea of transformation rather than mere adjustment. [Human evaluators’ Disagreement: 重塑 is used well, but 自我 personifies the subject Asia too much (not intended in source text).]
7	“We believe that the world is on the cusp of a new era with very different underlying forces than those that shaped our world over the past 30 years.” The translated target text sentence: “我们认为全球正处在一个新时代的风口浪尖，各种深层力量相互角逐。而这几股力量与过去30年间塑造世界的力量截然不同。”	“the cusp of a new era “→“新时代的风口浪尖”: The core meaning of the source sentence is conveyed clearly and faithfully. The notion of “on the cusp of a new era” is effectively captured by “正处在一个新时代的风口浪尖,” which reflects a moment of transition and significance. [Human evaluators’ disagreement: Net negative. The idiom is good (so the comment about it being lively isn’t wrong). But it does not accurately

threshold” or “change is about to happen”, the segment about the underlying forces describes the subject of “on the cusp”. 风口浪尖 is more on the side or “at the height of tensions”; it signifies underlying struggle, which is not suggested in the source text and therefore distorts the meaning.

Translation can be subjective. There are readers who marvel at the use of fancy words without thinking deeper at the meaning. 用词漂亮差不多得了. But we think here since the meaning is not fully reflective, the negative outweighs the positive.]

8 “Asia’s integration has come about through commercial pragmatism rather than political alignment.”
“亚洲的整合仰赖的是商业实用主义，而非政治联盟。”

“come about”→“仰赖”: the causal nuance of “has come about through... rather than...” is well captured by “仰赖的是...而非...”.

9 “Can that continue if trade tensions rise?”
“这条道路还行得通吗?”

“Can that continue...”→ “这条道路还行得通吗?”:

Accuracy: The core meaning of the source sentence is well preserved. The rhetorical question “Can that continue...” is effectively conveyed through “还行得通吗,” which captures the sense of feasibility and continuity under changing conditions. While “that” is slightly interpreted as “这条道路,” this is a justified and contextually appropriate explication rather than an error.

Style: This is the strongest aspect. The phrasing is natural, idiomatic, and fluent in Chinese. “这条道路还行得通吗” reads like original Chinese writing rather than a translation, and it maintains the rhetorical tone of the source text effectively.

Audience Appropriateness: The sentence is clear and accessible to educated general readers. It avoids overly technical or literal phrasing and instead uses a familiar metaphor (“道路”) that resonates well in Chinese discourse.

10 “In these places, Asia has more than 40 percent of the revenues, R&D, and patents.”
“在这些领域，亚洲坐拥超过40%的利润、研发和专利。”

“has”→“坐拥”:

Style: The phrase “亚洲坐拥超过40%” is particularly effective. “坐拥” adds a sense of possession and prominence that reads naturally and fluently in Chinese, avoiding a stiff or overly literal rendering. It enhances readability while preserving the emphasis of the original sentence.

11 “Asia has the people to fuel growth, but 90 percent of the expected increase in Asia’s non-farm workforces by 2050 will be where productivity is relatively low.”
Proceedings of the 17th Conference of the Association for Machine Translation in the Americas, Québec City, Canada, August 31 - September 2, 2026. Volume 1: Research Track

“90 percent of the expected increase”→“其中90%的增长”:
The translation reads naturally and fluently in Chinese. The use of “截至”

	截至2050年的亚洲非农劳动力预期增长中，其中90%的增长都来自生产力相对较低的地区。”	自…” creates a smooth, idiomatic flow rather than a rigid, word-for-word rendering. The sentence maintains a formal, analytical tone consistent with the original text. [Human evaluators’ disagreement: Neutral, the translation is correct, but nothing outstanding, and using 增长 twice in close succession may seem redundant to some.]
12	“We believe that the world is on the cusp of a new era with very different underlying forces than those that shaped our world over the past 30 years.” “人们认为相较于过去30年塑形世界的力量，新时代截然不同的潜在力量，将世界推向风口浪尖。”	“the cusp of a new era “→“新时代的 风口浪尖”: The sentence reads smoothly and idiomatically in Chinese. Expressions like “风口浪尖” effectively convey a sense of urgency and transformation, making the sentence engaging and vivid rather than stiff or overly literal. [Human evaluators’ disagreement: Translation can be subjective. There are readers who marvel at the use of fancy words without thinking deeper at the meaning. But We think here since the meaning is not fully reflective, the negative outweighs the positive.]
13	“Asia will be at the nexus of these forces and will experience heightened versions of the trends affecting the rest of the world.” “亚洲将成为这些力量的纽带，在影响世界其他地区的趋势上，也将经历更高版本的转机。”	“nexus”→”纽带”: The term “纽带” is an appropriate and understandable choice for a general audience when translating “nexus,” avoiding overly technical or obscure wording.
14	“Asia is the world’s trade crossroads. It’s home to 18 of the 20 fastest-growing trade corridors and 13 of the 20 largest.” “亚洲是世界贸易枢纽。20个增速最快贸易通道中的18个位于亚洲，20个最大贸易通道中的13个同样起源于亚洲。”	“crossroad”→“枢纽”: Key economic and trade-related terms are well chosen. “贸易枢纽” appropriately reflects “trade crossroads,” and “贸易通道” is a suitable and consistent rendering of “trade corridors,” aligning with standard usage in economic and policy contexts.
15	“But the value created by tech is moving away from hardware and into new transversal technologies.” “然而技术产生的价值正从硬件生产转移到新兴交叉科技。”	“transversal technologies”→“交叉科技”: The rendering of “transversal technologies” as “交叉科技” is particularly strong. It reflects an understanding of the concept as cross-domain or interdisciplinary technologies, which is consistent with how such ideas are expressed in Chinese technical and business contexts. [Human evaluators’ disagreement: this term does not seem adopted in Chinese and does not make much sense at face value.]
16	“But now, aging is fiercest in the high-productivity economies of the Pacific Rim.” “可是如今，老龄化在太平洋地区高生产力的经济体中成为最棘手的问题。”	“fiercest”→“最棘手的问题”: The core meaning of the source sentence is well preserved. “aging is fiercest” is effectively rendered as “老龄化…成为最棘手的问题,” which captures the intensity and seriousness of the demographic challenge, even though it is slightly interpretive. The relationship

clearly and correctly conveyed.
[Human evaluators' disagreement: The translation is idiomatic, but it also distorts the meaning.]

Table 4: True and false positive exemplars generated by GPT5. Percentage of false positive found by human evaluators is 7/19 (36.84%)

#	Different opinions from human trainees (comment 1) and GPT5.4 (comment 2)	DeepSeek arbitrations and human evaluator's final judgments
1	There are two comments on the translation pair below: English: "consumer electronics, industrial elec-tronics, electric vehicles, and semiconductors." Chinese: "消费电子产品、工业电子产品、电动汽车、半导体。" Comment 1: The last item in a list must be connected with a "和". Comment 2: No problem detected. Which comment is better? Why?	DeepSeek: Comment 1 is better. The English list explicitly includes "and"; before the last item, so the Chinese translation should reflect that with a conjunction like "和"; for naturalness and fidelity. While omitting it is not strictly incorrect in all contexts, the "must" in Comment 1 is slightly overstated, but it correctly identifies the improvement needed. Comment 2 overlooks this common convention. DeepSeek agrees with the human trainee. Human evaluators: disagree with DeepSeek: Such a convention does not exist. DeepSeek is wrong!
2	There are two comments on the translation pair below: "Asia has more than 40 percent of the revenues" "亚洲占了超过百分之四十的收益" Comment 1: Why "40%" " should be written in Chinese text "百分之四十". Are there any specific formatting rules? Comment 2: No problem detected. Which comment is better? Why?	DeepSeek: Comment 2 is better because it correctly identifies that the translation is accurate and follows standard Chinese localization practices. In Chinese, percentages are commonly written as "百分之四十"; rather than "40%"; in formal or narrative text, especially when the number is written out in full. Comment 1's concern is unnecessary, as there is no specific formatting rule violated—this usage is perfectly natural and idiomatic in Chinese. DeepSeek agrees with GPT5.4: this usage is perfectly natural and idiomatic in Chinese. Human evaluators: agree with DeepSeek. DeepSeek is right.
3	There are two comments on the translation pair below: "But the value created by tech is moving away from hardware and into new transversal technologies. So, the question is whether Asia can reinvent itself as a tech creator as well as a tech manufacturer." "然而创造价值的科技产业正逐渐从硬体转变为新式横向科技, 因此问题在于亚洲是否能在科技创新和科技制造业这些领域进	DeepSeek: Comment 1 is better because it accurately identifies a regional terminology issue. The word "硬件" is standard in mainland China, while "硬体" is primarily used in Taiwan. Since the translation likely targets a mainland Chinese audience, using "硬体" could be considered unnatural or a regional mismatch. Comment 2 overlooks this nuance.

	Comment 1: The translation of “hardware” (硬件) is an expression used in Taiwan. But it is not a conventional expression in mainland China. Comment 2: No problem detected. Which comment is better? Why?	feedback for improving the translation’s appropriateness. DeepSeek agrees with human trainee’s point: the regional terminology issue exists. Human evaluators: agree with DeepSeek. DeepSeek is right.
4	There are two comments on the translation pair below: “So, the question is whether Asia can reinvent itself as a tech creator as well as a tech manufacturer.” “问题就变成亚洲是否能够不仅充当技术生产者，还能将自己重塑成为一个技术创新者。” Comment 1: “reinvent itself” means reinvent for both technology creator and manufacturer. So the verb should cover both. Comment 2: No problem detected. Which comment is better? Why?	DeepSeek: Comment 2 is better because the translation accurately captures the meaning of the original. In the English sentence, “reinvent itself”; refers specifically to the transformation into a tech creator, given that Asia is already a tech manufacturer. The Chinese translation reflects this by using “充当” (act as) for the existing role and “将自己重塑成为”; (reinvent itself as) for the new role, which correctly separates the two concepts. Comment 1 incorrectly suggests that the verb must cover both roles, but that would be redundant or misleading since the manufacturing role does not require reinvention. DeepSeek agrees with GPT5.4: , “reinvent itself”; refers specifically to the transformation into a tech creator, given that Asia is already a tech manufacturer. Human evaluators: agree with DeepSeek. DeepSeek is right.
5	There are two comments on the translation pair below: “Asia has largely been in a demographic sweet spot with large pools of young workers and surging productivity in some economies.” “亚洲从很大程度上一直拥有人口红利，在某些经济体中拥有大量的年轻工人和高涨的生产力。” Comment 1: “Demographic sweet spot”. translated as “人口红利”, is a mis-translation. Comment 2: No problem detected. Which comment is better? Why?	DeepSeek: Comment 2 is better because the translation “人口红利”; accurately captures the meaning of “demographic sweet spot”; In economic and demographic contexts, “demographic sweet spot”; refers to a period when a large working-age population relative to dependents creates favorable conditions for economic growth—exactly what “人口红利”; (demographic dividend) denotes. The Chinese version also correctly conveys the additional details about young workers and surging productivity. Therefore, the translation is correct, and Comment 2 is valid. DeepSeek agrees with GPT5.4: “人口红利” is a good translation. Human evaluator: disagrees with DeepSeek: It changes the meaning too much. DeepSeek is wrong!
6	There are two comments on the translation pair below: “It’s home to 18 of the 20 fastest-growing trade corridors and 13 of the 20 largest.” “20个增速最快贸易通道中的18个位于亚洲，20个最大贸易通道中的13个同样起源于亚洲。” Comment 1: “条” is a more appropriate quantity word.	DeepSeek: Comment 1 is better because the Chinese translation uses the generic measure word “个” (gè) for “贸易通道” (trade corridors), but the more precise and natural measure word for “通道” (corridor) is “条” (tiáo). While “个” is acceptable in informal contexts, “条” is more accurate and idiomatic in

Which comment is better? Why?

describing physical or conceptual routes. Comment 2 overlooks this subtle but important nuance. DeepSeek agrees with human trainee. **Human evaluators:** agree with DeepSeek: “条” is better. DeepSeek is correct.

7 There are two comments on the translation pair below: “It’s home to 18 of the 20 fastest-growing trade corridors and 13 of the 20 largest.” “20个增速最快贸易通道中的18个位于亚洲，20个最大贸易通道中的13个同样起源于亚洲。” Comment 1: “亚洲” was repeated, which is unnatural. It is better to use a different word “此地” (“this place”). Comment 2: No problem detected. Which comment is better? Why?

DeepSeek: Comment 2 is better. The repetition of “亚洲” in the Chinese translation is not unnatural—it maintains clarity and emphasis in a parallel structure. Using “此地” (“this place”) could introduce ambiguity if the referent (Asia) is not explicitly stated nearby, and it sounds slightly archaic or formal in this context. The original translation is clear and idiomatic, so no change is needed. DeepSeek agrees with GPT5.4. That is not a problem to repeat “亚洲” . **Human evaluators:** agree with DeepSeek. DeepSeek is correct.

8 There are two comments on the translation pair below: ““Asia excels in manufacturing new technology in four areas in particular: consumer electronics, industrial electronics, electric vehicles, and semiconductors” “亚洲在新科技生产方面遥遥领先，尤其在消费电子产品、工业电子、电动汽车和半导体四个方面。” Comment 1: The source text only mentioned “technology” (“技术”). In contrast, the target text phrase “科技” covers both “science” and “technology”. Therefore, there is added meaning. Comment 2: No problem detected. Which comment is better? Why?

DeepSeek: Comment 2 is better because the translation is accurate and natural in Chinese. The term “科技” in Chinese commonly serves as a shorthand for “技术” (technology) in contexts like manufacturing and industry, especially when paired with “新” (new). While “科技” literally covers both science and technology, its usage here is idiomatic and does not introduce unwarranted meaning—the source’s “new technology” is faithfully rendered. Comment 1 is overly pedantic and fails to account for standard Chinese lexical conventions. DeepSeek agrees with GPT5.4: It is a standard Chinese lexical convention. **Human evaluators:** agree with DeepSeek. DeepSeek is correct.

Table 5: DeepSeek arbitrated disagreements on negative exemplars between human trainees and GPT-5.4. Based on the human evaluators' final judgments, DeepSeek's arbitration was correct in 6 of the 8 cases.

#	Different opinions from human trainees (comment 1) and GPT5.4 (comment 2)	DeepSeek’s arbitration and Human evaluator’s Re-evaluation.
1	There are two comments on the translation pair below: “Asia is the world’s trade crossroads. It’s home to 18 of the 20 fastest-growing trade corridors and 13 of the 20 largest.” “亚洲是全球贸易的岔道口。这里盘踞着全球20条增长最快的贸易走廊中的18条，20条规模最大的贸易走廊中的13条。”	DeepSeek: Comment 1 is better because it provides specific, constructive feedback by praising a particular word choice “盘踞”) that effectively captures the meaning and tone of “home”; in the context. It acknowledges a successful translation decision, which is helpful for the translator and encourages good

	word choice. Comment 2: No positive comments. Which comment is better? Why?	Comment 2 offers no insights or justification, making it unhelpful for improving translation quality or understanding. DeepSeek agrees with human trainees that particular word choice is a successful translation decision. Human evaluators: agree with DeepSeek and human trainee. positive, good language quality. 盘踞 is a verb suitable to describe something long and thin, which is in accordance with the concept of “corridor”. DeepSeek is correct.
2	There are two comments on the translation pair below: “Asia excels in manufacturing new technology in four areas in particular: consumer electronics, industrial electronics, electric vehicles, and semiconductors.” “亚洲在新科技生产方面遥遥领先，尤其在消费电子产品、工业电子、电动汽车和半导体四个方面。” Comment 1: “excel”→“遥遥领先” is an excellent choice. Comment 2: No positive comments. Which comment is better? Why?	DeepSeek: Comment 1 is better because it provides specific, constructive feedback that highlights a successful translation choice. The phrase :遥遥领先” effectively captures the meaning of “excels”; in this context, adding emphasis that fits naturally in Chinese. This type of positive comment encourages the translator and offers a clear example of good practice. Comment 2, on the other hand, offers no useful information—it neither acknowledges the quality of the translation nor helps the translator improve. In a review or learning setting, specific praise (or critique) is far more valuable than a generic dismissal. DeepSeek agrees with human trainee’s comment. Human evaluators: agree with DeepSeek: Positive, good language quality. DeepSeek is correct.

Table 6: DeepSeek arbitrated disagreements on positive exemplars between human trainees and GPT-5.4. Based on the human evaluators' final judgments, DeepSeek's arbitration was correct in two of the two cases.

#	GPT5.4 generated explanation (comment 1) and human evaluator’s disagreements (comment 2)	DeepSeek’s arbitration and Human evaluator’s Re-evaluation
1	There are two comments on the translation pair below: English: “We believe that the world is on the cusp of a new era...” Chinese: ““我们相信新的时代马上就要来临...”” Comment 1: Accuracy/mistranslation issue: “on the cusp of” □ “马上就要来临” Comment 2: The translation is acceptable as long as it reflects the sentiment of “change is about to happen” or “a new era is about to start”. Which comment is better? Why?	Deep Seek: Comment 2 is better because it recognizes that the translation effectively conveys the core meaning and sentiment of the original phrase, even if it doesn’t use a literal equivalent of “on the cusp of”; In translation, capturing the intended meaning and naturalness in the target language is often more important than word-for-word accuracy. The Chinese expression “马上就要来临” naturally communicates the immediacy and impending change, which aligns with the original’s sense of being on the verge of a new era. Comment 1 is overly literal and fails to account for

		agrees with human evaluator's judgment. Evaluators' further investigation confirms the original judgment. DeepSeek is correct.
2	There are two comments on the translation pair below: English: "But now, it could find itself in the crosshairs of trade tensions." Chinese: "当今贸易紧张局势也在这里得到了格外强烈的体现" Comment 1: Add meaning: "也" added meaning in Chinese Comment 2: We don't think it's an addition here. 也 isn't "also" all the time. Here it looks like 也 is used to continue on the context of the previous sentence (Asia is centre of trade paths, it may now be the centre of trade tensions). Which comment is better? Why?	DeepSeek: Comment 2 is better because it recognizes the pragmatic and cohesive function of "也" in Chinese discourse. In this context, "也" does not primarily mean "also" in a literal additive sense; rather, it serves as a connective device that ties the current statement to the preceding context (e.g., Asia being the center of trade paths). This is a natural and idiomatic usage in Chinese, not an unnecessary addition of meaning. Comment 1 oversimplifies the issue by treating "也" as a simple semantic addition, ignoring its textual role. Therefore, Comment 2 provides a more accurate and nuanced analysis. DeepSeek agrees with evaluators' disagreement. Evaluators' further investigation confirms the original judgment. DeepSeek is correct.
3	There are two comments on the translation pair below: English: "Asia's integration has come about through commercial pragmatism rather than political alignment." Chinese: "亚洲的融入是通过商业务实性，而不是政治取向一致性。" Comment 1: Linguistic convention issue: "通过商业务实性" is grammatically acceptable but unnatural. "基于/源于" is better. Comment 2: 基于/源于 are better but 通过 isn't wrong and still flies, We wouldn't flag this as an error. Which comment is better? Why?	DeepSeek: Comment 1 is better because it identifies a genuine nuance in idiomatic expression. While "通过" is grammatically acceptable, the English phrase "come about through" implies a causal foundation or driving force—not merely a method or channel. In Chinese, "基于" (based on) or "源于" (deriving from) more naturally convey this underlying reason, making the translation sound more native. Comment 2 is too permissive: it fails to recognize that a small collocation shift can improve clarity and fluency, which is the goal of good translation. Therefore, Comment 1 offers a useful, targeted improvement rather than just excusing a passable but suboptimal choice. DeepSeek does not agree with evaluators' disagreement. Evaluators' further investigation confirm the original judgment. DeepSeek is incorrect.
4	There are two comments on the translation pair below: English: "Can that continue if trade tensions rise?" Chinese: "如果贸易紧张局势加剧，亚洲还可以维持现状吗?" Comment 1: Omission: "if trade tensions rise". Comment 2: There is no omission here. The translation of "Can that continue if trade tensions rise?" is broken down into 2 clauses in Chinese separated by a comma, the translation is "如果贸易紧张局势加	DeepSeek: Comment 2 is better because the Chinese translation provided — "如果贸易紧张局势加剧，亚洲还可以维持现状吗?" — actually includes the conditional clause "if trade tensions rise" (如果贸易紧张局势加剧). Comment 1 incorrectly claims an omission. Comment 2 correctly identifies that the clause is present, placed before the comma, and explains why the earlier alleged omission was mistaken. DeepSeek agrees with evaluators' disagreement. Evaluators' further investigation confirms the original judgment. DeepSeek is correct.

	alleged omission is placed before the comma, the LLM simply missed it. Which comment is better? Why?	Evaluators' further investigation confirms the original judgment. DeepSeek is correct.
5	There are two comments on the translation pair below: English: ""Can that continue if trade tensions rise?" Chinese: "亚洲还可以维持现状吗?" Comment 1: Style (Slightly awkward phrasing) problem to translate "that continue" as "维持现状" Comment 2: The mistranslation seems to be caused by a misunderstanding of the source sentence and it is not the word "continue". Which comment is better? Why?	DeepSeek: Comment 2 is better because it correctly identifies the root cause of the mistranslation: the word "that" in the English source was misunderstood or arbitrarily replaced with "亚洲" (Asia) in the Chinese, which is a semantic error. Comment 1 focuses on a style issue with the translation of "continue" as "维持现状," which is a minor phrasing choice; the real problem is the insertion of "亚洲" that has no basis in the original sentence. Thus, Comment 2 addresses the more serious mistake. DeepSeek agrees with evaluator's disagreement. Evaluators' further investigation confirms the original judgment. DeepSeek is correct.
6	There are two comments on the translation pair below: English: "Asia excels in manufacturing new technology in four areas in particular." Chinese: "亚洲在如下四项新型科技 制造方面超群拔类" Comment 1: Accuracy issue: translate "the following four items" as "如下四项" Comment 2: "以下" sounds better than "如下", but this minor enough that it should not be counted as an error.] Which comment is better? Why?	DeepSeek: Comment 1 is better because it identifies a genuine accuracy issue: the English source does not include the phrase "the following" — it simply says "in four areas in particular." The Chinese translation adds "如下" (as follows), which implies that those areas are about to be listed, a nuance not present in the original. This is an unnecessary and potentially misleading addition. Comment 2, on the other hand, focuses on a minor stylistic preference ("以下" vs. "如下") that does not affect meaning and explicitly states it should not count as an error. Therefore, Comment 1 addresses a more substantive problem with the translation. DeepSeek agrees more with GPT5.4's point. Evaluators' further investigation confirms the original judgment. DeepSeek is incorrect.
7	There are two comments on the translation pair below: English: "Asia has the people to fuel growth, but 90 percent of the expected increase in Asia's non-farm workforces by 2050 will be where productivity is relatively low." Chinese: "亚洲拥有发展所需的人口, 但2050年前亚洲非务农工作力的预期增长的90%将处于生产力相对低的地区。" Comment 1: Terminology issue: translate "non-farm workforces" as "非务农工作力" Comment 2: 非务农 and 非农业 both should work. It is not an error]. Which comment is better? Why?	DeepSeek: Comment 2 is better because it correctly recognizes that both "非务农" and "非农业" are acceptable translations for "non-farm" in this context, and it wisely avoids insisting on a single, less idiomatic phrasing. The term "非农业劳动力" is more standard and widely used in Chinese economic texts, while "非务农劳动力" is also understandable but less common. Additionally, "工作力" in Comment 1 is less natural than the standard "劳动力" for "workforce." Comment 2's open-minded assessment is more accurate and helpful for translation quality. DeepSeek agrees with human disagreements. Evaluators' further

		judgment. DeepSeek is correct.
8	There are two comments on the translation pair below: English: “But now, it could find itself in the crosshairs of trade tensions.”” Chinese: “然而目前亚洲也发现自己紧张的贸易关系中成了靶子。” Comment 1: Accuracy issue: translate “trade tensions” as “紧张的贸易关系”. The source refers specifically to trade tensions/conflicts, which implies geopolitical friction or disputes. The translation weakens the meaning and changes the nuance. Comment 2: The turn of phrase is a bit awkward but it isn’t an error in accuracy. Interpreting it as “conflicts” may be considered an over translation for the exam.] Which comment is better? Why?	DeepSeek: Comment 2 is better. The translation “紧张的贸易关系” (tense trade relations) is a standard and accurate rendering of “trade tensions” — it does not weaken the meaning, as both terms refer to the same geopolitical friction. Comment 1 overstates the issue; there is no significant loss of nuance. Comment 2 correctly identifies the awkwardness in phrasing (the Chinese sentence is grammatically clunky) but recognizes that this is a stylistic flaw, not an accuracy error. Furthermore, Comment 2’s point about over-translation into “conflicts” is valid, as that would be a shift in register. Therefore, Comment 2 offers a more balanced and precise critique. DeepSeek agrees with human disagreements. Evaluators’ further investigation confirms the original judgment. DeepSeek is correct.
9	There are two comments on the translation pair below: English: “But the value created by tech is moving away from hardware and into new transversal technologies. So, the question is whether Asia can reinvent itself as a tech creator as well as a tech manufacturer.” Chinese: “然而创造价值的科技产业正逐渐从硬件转变为新式横向科技，因此问题在于亚洲是否能在科技创新和科技制造业这些领域进行自我重塑。” Comment 1: Accuracy issue to translate “tech creator” as “科技创新” Entire sentence structure is an over-expansion not present in source. Comment 2: The source words “creator” and “manufacturer” are stylistic choice. The “base” meaning of the segment is “create/invent tech” and “manufacture tech”, which is preserved. We would not expect the parts of speech in the translation to align perfectly with the source - syntax works differently in different languages. Which comment is better? Why?	DeepSeek: Comment 1 is better. Reasoning: Comment 1 correctly identifies two specific problems: 1. Accuracy issue: The source uses “tech creator” (an entity that creates technology), but the translation renders it as “科技创新” (technological innovation), which shifts the meaning from a role/agent to an abstract activity. This is a semantic change, not a mere stylistic or syntactic adjustment. 2. Over-expansion: The Chinese adds “这些领域” (these fields) and restructures the sentence into a “reinvent in the fields of X and Y” format, while the source says “reinvent itself as X as well as Y.” This changes the core relationship from identity (being a creator and manufacturer) to domain (operating within innovation and manufacturing fields). Comment 2 attempts to defend the translation by claiming the “base meaning” is preserved and that part-of-speech alignment is unnecessary. However, this overlooks the crucial distinction between being a creator (+ noun, subject) and engaging in innovation (+ abstract noun, field). The translation’s structural expansion also introduces a meaning not present in the source, which goes beyond acceptable syntactic variation. DeepSeek agrees with GPT5.4. Evaluators’ further investigation confirms the original judgment: comment 2 is better. The base meaning analysis in comment 2 is accurate. Comment 1 rigidly separates create(or) and 创新 as two different things, whereas in this context they refer to the

- There are two comments on the translation pair below:
 English: "Asia has largely been in a demographic sweet spot with large pools of young workers and surging productivity in some economies."
 Chinese: "亚洲有大批的年轻工人，当中有些经济体的生产力又有所提高，这些在很大程度上都是最有利的人口特征。"
- 10 Comment 1: Accuracy issue: "these largely are" translated as "这些在很大程度上都是" introduces an explanatory structure not present in the source sentence.
 Comment 2: We think 很大程度上 matches "largely" in the source text, what's changed is the sentence structure - but the meaning of "largely" is preserved. This is not an error.
 Which comment is better? Why?
- DeepSeek:** Comment 2 is better because it correctly recognizes that translation often requires restructuring sentences while preserving meaning. The Chinese version accurately conveys "largely" as "在很大程度上" and the overall idea—that Asia's demographic conditions (young workers and rising productivity) constitute a "demographic sweet spot." The added explanatory structure ("这些...都是") in the Chinese does not distort the source meaning; it simply makes the reference explicit, which is a natural adaptation in Chinese. Comment 1 is overly literal and fails to account for acceptable rephrasing in translation. DeepSeek agrees with human disagreement. **DeepSeek is correct.**
- There are two comments on the translation pair below:
 English: "Asia has largely been in a demographic sweet spot with large pools of young workers and surging productivity in some economies."
 Chinese: "亚洲从很大程度上一直拥有人口红利，在某些经济体中拥有大量的年轻工人和高涨的生产力。"
- 11 Comment 1: "productivity" is translated as "生产力" instead of the more precise economic term "生产率".
 Comment 2: 生产力 is not wrong. Here the paragraph is talking about demographics and labour, and with young workers it naturally factors that the output capability is high. "Productivity" in this context refers more likely to the ability to produce, rather than the ratio of material output over input.
 Which comment is better? Why?
- DeepSeek:** Comment 1 is better. While "生产力" (productive forces) is not entirely incorrect, "生产率" (productivity rate/efficiency) is the standard and more precise economic term for "productivity" in this context. The English sentence explicitly separates "large pools of young workers" (a demographic factor) from "surging productivity" (an efficiency factor), so the latter refers to output per worker or per unit of input, not just the general ability to produce. Using "生产率" avoids ambiguity and aligns with common usage in economic translations. DeepSeek agrees with GPT5.4. Evaluators' further investigation confirms the original judgment. **DeepSeek is incorrect.**
- There are two comments on the translation pair below:
 English: "But now, aging is fiercest in the high-productivity economies of the Pacific Rim."
 Chinese: "但如今，环太平洋的高效经济体中的老龄化问题也最为严重。"
- 12 Comment 1: The word "也" (also) is added in the translation, but the source sentence does not contain the meaning "also."
 Comment 2: We can see how 也 may be a problem if one is to back-translate this into English. However, there is an implied contrast between this sentence and the previous sentence in the source. "Asia population situation good. (But) some countries have population problems". 也 is used as an emphasis for the aging issue,
- DeepSeek:** Comment 2 is better because it takes into account the broader context and the pragmatic function of "也" in Chinese. While Comment 1 correctly notes that the word "也" is not directly present in the English source, translation fidelity is not simply about word-for-word equivalence. The source sentence begins with "But now," which signals a contrast with prior discourse. If the preceding text discussed aging in other regions (e.g., "Aging was moderate in other areas"), then using "也" in Chinese can serve to highlight that the same issue is now also present in the Pacific Rim, effectively conveying the contrast and emphasis intended by the source's "But" and "fiercest." In Chinese, "也" often carries such emphatic or contrastive nuance beyond

Which comment is better? Why?

There are two comments on the translation pair below:

English: “but 90 percent of the expected increase in Asia’s non-farm workforces by 2050 will be where productivity is relatively low.”

Chinese: “但在截至2050年的亚洲非农劳动力预期增长中，其中90%的增长都来自生产力相对较低的地区。”

Comment 1: Accuracy issue: “will be where productivity is relatively low” was translated as “来自生产力相对较低的地区”. The source sentence means the workforce increase will occur in low-productivity areas, not that it comes from those areas. “来自” introduces an unintended source meaning.

Comment 2: We think this is fine and would not cause misunderstanding on the part of the reader. The core message refers to where labor growth is, the reader would think about location rather than where the people are born (“from”).

Which comment is better? Why?

There are two comments on the translation pair below:

English: “Asia excels in manufacturing new technology in four areas in particular: consumer electronics, industrial electronics, electric vehicles, and semiconductors”

Chinese: “亚洲在新科技生产方面遥遥领先，尤其在消费电子产品、工业电子、电动汽车和半导体四个方面。”

Comment 1: Accuracy issue: “consumer electronics” is translated as “消费电子产品”. Adding “产品” slightly expands the meaning. While common in Chinese, it is not strictly present in the source text.

Comment 2: Such term exists and is used in Chinese. It does not lead to any misunderstanding.

Which comment is better? Why?

There are two comments on the translation pair below:

English: “So, the question is whether Asia can reinvent itself as a tech creator as well as a tech manufacturer.”

Chinese: “如此一来，问题就变成亚洲是否能不仅作为技术生产者，也是技术创新者进行再次创新。”

Comment 1: Addition: The phrase “如此一来” (“thus/therefore”) is added but does

overlooks these pragmatic considerations and imposes a overly literal standard, whereas Comment 2 demonstrates a more nuanced understanding of translation as conveying meaning and discourse cohesion.

DeepSeek agrees with human evaluator. Evaluators’ further investigation confirms the original judgment. **DeepSeek is correct.**

DeepSeek: Comment 2 is better because it recognizes that the translation is idiomatic and does not cause misunderstanding. The Chinese phrase “来自...地区” is commonly used to indicate where something occurs or is located (e.g., “90% of the increase comes from low-productivity areas” naturally means the increase happens in those areas), not necessarily implying origin or birthplace. The core message—that the workforce growth will happen in low-productivity regions—is preserved. Comment 1 over-interprets the preposition “来自” and applies a stricter, less context-sensitive standard that would make many natural translations seem inaccurate.

DeepSeek agrees with human evaluator. Evaluators’ further investigation confirms the original judgment. **DeepSeek is correct.**

DeepSeek: Comment 2 is better. The addition of “产品” (products) in the Chinese translation is natural and idiomatic in Chinese. “Consumer electronics” (消费电子产品) is a standard, widely accepted term in Chinese that accurately conveys the intended meaning. The slight expansion does not create any misunderstanding or distort the original message, and it aligns with common usage rather than being a strict accuracy error. Comment 1 is overly pedantic, as it fails to account for natural target-language conventions.

DeepSeek agrees with human evaluator. Evaluators’ further investigation confirms the original judgment. **DeepSeek is correct.**

DeepSeek: Comment 2 is better because it goes beyond mere observation to evaluate the translation's appropriateness. It correctly recognizes that “如此一来” is a natural Chinese discourse marker that fulfills the same cohesive function as “So” in the source—linking the statement to previous context. Comment 1 simply notes an addition without considering

is wrong. It is likely serves the function “so” serves in the source. “Given the above context” is the idea “so” and 如此一来 express. The translation is appropriate.]

translation, missing the key principle that good translation often requires adding or omitting elements for naturalness and clarity in the target language.

Which comment is better? Why?

DeepSeek agrees with human evaluator. Evaluators’ further investigation confirms the original judgment. **DeepSeek is correct.**

Table 7: DeepSeek arbitrated disagreements on negative exemplars between GPT5.4 and human evaluators. Based on the human evaluators’ further investigation and final judgments, DeepSeek’s arbitration was correct in 11 of the 15 cases.

#	GPT5.4 generated explanation (comment 1) and human evaluator’s disagreements (comment 2)	DeepSeek’s arbitration and Human evaluator’s Re-evaluation
1	There are two comments on the translation pair below: English: “We believe that the world is on the cusp of a new era with very different underlying forces than those that shaped our world over the past 30 years.” Chinese: “我们相信世界正处于新时代的开端，与过去30年相比，影响世界的根本势力已变得非常不同。” Comment 1: “underlying force” is translated as “根本势力”: The phrase “very different underlying forces” is effectively translated as “根本势力已变得非常不同,” preserving the conceptual emphasis on foundational drivers of change. There is no omission or distortion of meaning. Comment 2: Translating “underlying” as 根本 is not wrong enough to warrant a point deduction, but 根本 can mean “source/origin” whereas the meaning of “underlying” is more like “deep down/deep within”. Therefore, this is a minor false positive.] Which comment is better? Why?	DeepSeek: Comment 2 is better because it offers a more nuanced and accurate linguistic critique. While Comment 1 states that the translation preserves meaning without distortion, Comment 2 correctly identifies a subtle semantic shift: “根本” leans toward “source/origin,” whereas “underlying” conveys “deep down / deep within.” This distinction matters in professional translation evaluation, as it reflects a slight misalignment in connotation. Comment 2 is also more helpful because it acknowledges the translation is acceptable (no point deduction) while still highlighting a minor false positive, thus providing balanced and precise feedback. Comment 1 oversimplifies the issue by claiming no distortion at all. DeepSeek agrees with human evaluator. Evaluators’ further investigation confirms the original judgment. DeepSeek is correct.
2	There are two comments on the translation pair below: English: “So, the question is whether Asia can reinvent itself as a tech creator as well as a tech manufacturer.” Chinese: “因此问题在于亚洲是否能在科技创新和科技制造业这些领域进行自我重塑。” Comment 1: “reinvent itself”→“自我重塑”: the notion of “reinvent itself” is effectively rendered as “进行自我重塑,” which captures the idea of transformation rather than mere adjustment. Comment 2: 重塑 is used well, but 自我 personifies the subject Asia too much (not intended in source text). Which comment is better? Why?	DeepSeek: Comment 1 is better because it accurately recognizes that the source text intentionally personifies Asia through the reflexive pronoun “itself” in “reinvent itself.” The Chinese translation “进行自我重塑” faithfully preserves this personification, as the metaphor of reinvention inherently implies agency. Comment 2 incorrectly claims the personification is unintended, when in fact the English explicitly uses “itself” to personify Asia, making “自我” an appropriate and natural rendering. DeepSeek agrees with GPT5.4. Evaluators’ further investigation (自我重塑 isn’t wrong but it personifies Asia (as one entity). Asia does not act as one entity, what the writer means is whether the countries in Asia can make a shift) confirms the original judgment. DeepSeek is incorrect.

pair below:

English: "We believe that the world is on the cusp of a new era with very different underlying forces than those that shaped our world over the past 30 years."

Chinese: "我们认为全球正处在一个新时代的风口浪尖, 各种深层力量相互角逐。而这几股力量与过去30年间塑造世界的力量截然不同。"

Comment 1: "the cusp of a new era" → "新时代的风口浪尖": The core meaning of the source sentence is conveyed clearly and faithfully. The notion of "on the cusp of a new era" is effectively captured by "正处在一个新时代的风口浪尖," which reflects a moment of transition and significance.

Comment 2: The idiom is good (so the comment about it being lively isn't wrong). But it does not accurately reflect "cusp". "Cusp" means "at the threshold" or "change is about to happen", the segment about the underlying forces describes the subject of "on the cusp". 风口浪尖 is more on the side or "at the height of tensions"; it signifies underlying struggle, which is not suggested in the source text and therefore distorts the meaning.

Which comment is better? Why?

There are two comments on the translation pair below:

English: "Asia has the people to fuel growth, but 90 percent of the expected increase in Asia's non-farm workforces by 2050 will be where productivity is relatively low."

Chinese: "亚洲虽然拥有人力来推动增长, 但在截至2050年的亚洲非农劳动力预期增长中, 其中90%的增长都来自生产力相对较低的地区。"

4 Comment 1: "90 percent of the expected increase" is translated as "其中90%的增长":

The translation reads naturally and fluently in Chinese. The use of "截至2050年" and "其中90%的增长都来自..." creates a smooth, idiomatic flow rather than a rigid, word-for-word rendering. The sentence maintains a formal, analytical tone consistent with the original text.

Comment 2: Neutral, the translation is correct, but nothing outstanding, and using 增长 twice in close succession may seem redundant to some.

Which comment is better? Why?

Deep Seek: Comment 2 is better because it identifies a significant semantic inaccuracy in the translation. The English phrase "on the cusp of a new era" means being at the threshold or on the verge of a new era, emphasizing imminent change—not being in the midst of a struggle. The Chinese idiom "风口浪尖" (feng kou lang jian) literally means "at the mouth of the wind and the tip of the waves," and figuratively refers to being at the center of a storm or in a highly tense, contentious situation. This connotation of active conflict or turmoil is absent from the original, which simply states that the underlying forces are different, not that they are locked in competition. Additionally, the translation gratuitously adds "各种深层力量相互角逐" ("various deep forces are competing with each other"), which further distorts the original meaning. Comment 1 mistakenly equates "cusp" with "风口浪尖" by focusing only on the transitional sense, ignoring that the idiom carries a strong implication of struggle that misrepresents the source text. Thus, Comment 2 provides a more accurate, linguistically informed critique.

DeepSeek agrees with human evaluators. Evaluators' further investigation confirms the original judgment. **DeepSeek is correct.**

DeepSeek: Comment 1 is better. It provides a detailed, positive assessment of the translation's natural fluency and idiomatic structure, highlighting specific elements like the handling of the temporal phrase "by 2050" and the use of "其中...都来自" to create a smooth flow. This analysis demonstrates a deeper understanding of translation quality, focusing on overall effectiveness rather than a minor stylistic quibble.

Comment 2, while valid in noting the repeated use of "增长," is overly neutral and focuses on a potential redundancy that does not significantly detract from the translation's correctness or readability. In Chinese, such repetition is often acceptable for clarity and emphasis, making this critique less impactful. A good translation critique should weigh strengths against minor weaknesses, and Comment 1 does that more insightfully.

DeepSeek agrees with GPT5.4. Evaluators' further investigation confirms the original judgment.

DeepSeek is correct.

DeepSeek: Comment 2 is better because it prioritizes semantic fidelity

English: “We believe that the world is on the cusp of a new era with very different underlying forces than those that shaped our world over the past 30 years.”

Chinese: “人们认为相较于过去30年塑形世界的力量，新时代截然不同的潜在力量，将世界推向风口浪尖。”

Comment 1: “the cusp of a new era “□“新时代的风口浪尖” : The sentence reads smoothly and idiomatically in Chinese. Expressions like “风口浪尖” effectively convey a sense of urgency and transformation, making the sentence engaging and vivid rather than stiff or overly literal.

Comment 2: Translation can be subjective. There are readers who marvel at the use of fancy words without thinking deeper at the meaning. But We think here since the meaning is not fully reflective, the negative outweighs the positive.] Which comment is better? Why?

There are two comments on the translation pair below:

English: “But the value created by tech is moving away from hardware and into new transversal technologies.”

Chinese: “然而技术产生的价值正从硬件生产转移到新兴交叉科技。”

Comment 1: “transversal technologies” is translated as “交叉科技”: The rendering of “transversal technologies” as “交叉科技” is particularly strong. It reflects an understanding of the concept as cross-domain or interdisciplinary technologies, which is consistent with how such ideas are expressed in Chinese technical and business contexts.

Comment 2: this term does not seem adopted in Chinese and does not make much sense at face value.

Which comment is better? Why?

There are two comments on the translation pair below:

English: “But now, aging is fiercest in the high-productivity economies of the Pacific Rim.”

Chinese: “可是如今，老龄化在太平洋地区高生产力的经济体中成为最棘手的问题。”

Comment 1: The core meaning of the source sentence is well preserved. “aging is fiercest” is effectively rendered as “老龄化...成为最棘手的问题,” which captures the intensity and seriousness of the demographic challenge, even though it is slightly interpretive. The relationship between aging and “high-productivity economies of the Pacific Rim” is clearly and correctly conveyed.

Comment 2: The translation is idiomatic, but it also distorts the meaning.

forward-looking perspective (“on the cusp of a new era,” “We believe”), while the Chinese translation introduces the dramatic, crisis-laden idiom “风口浪尖” (at the crest of a wave / in the teeth of a storm) and changes the subject from “We” to “人们” (people). This alters both the tone and the agency. Comment 1 merely praises the vividness without addressing these inaccuracies, which is a superficial evaluation for a translation task. Comment 2 correctly identifies that the meaning is not fully reflected, and that the negative impact of distorting the original outweighs any stylistic gains. A good translation must first be faithful; flair is secondary.

DeepSeek with human evaluators. Evaluators’ further investigation confirms the original judgment.

DeepSeek is correct.

DeepSeek: Comment 1 is better because it accurately recognizes that “交叉科技” is a valid and idiomatic translation of “transversal technologies” in Chinese technical and business discourse. The term conveys the idea of cross-domain or interdisciplinary technologies, which aligns with common usage in Chinese innovation contexts. Comment 2’s assertion that the term is not adopted and makes little sense is incorrect—it is actually established in relevant literature.

DeepSeek agrees with GPT5.4.

Evaluators’ further investigation confirms the original judgment (交叉科技 does not make sense to native Chinese speaker). **DeepSeek is incorrect.**

DeepSeek: Comment 1 is better because it accurately assesses the translation’s strengths and acknowledges its slight interpretive nature without overstating the flaws. Why Comment 1 is more accurate: The translation “老龄化在太平洋地区高生产力的经济体中成为最棘手的问题” indeed captures the core meaning of the source: it conveys that aging is a severe and pressing issue in the specified economies. The shift from “is fiercest” (suggesting intensity) to “成为最棘手的问题” (becoming the most troublesome problem) is a natural idiomatic adjustment in Chinese. It preserves the essence—that aging is at its most acute in these regions—while sounding natural in the target language. Comment 1 correctly notes this adaptation as “slightly interpretive” but

Comment 2 claims the translation “distorts the meaning,” which is too strong. While there is a minor nuance change (from a description of intensity to a problem-oriented phrasing), the overall meaning—that aging is most severe in high-productivity Pacific Rim economies—is not distorted. The translation does not introduce false information or contradict the source; it simply uses a different but equivalent expression for “fiercest” in Chinese. Therefore, calling it a distortion is an overstatement and fails to recognize the acceptable degree of idiomatic adaptation.

In summary: Comment 1 provides a balanced, context-aware evaluation; Comment 2 is overly critical and not fully justified by the actual translation. DeepSeek agrees with GPT5.4. Evaluators’ further investigation confirms the original judgment
DeepSeek is incorrect.

Table 8: DeepSeek arbitrated disagreements on positive exemplars between GPT5.4 and human evaluators. Based on the human evaluators' further investigation and final judgments, DeepSeek's arbitration was correct in 3 cases out of 7.
