



Temporal Reliability of ChatGPT-Generated Bilingual Texts: Implications for AI-Assisted Language Learning and Multilingual Information Access

FU Zhihan XU Yaowei

Zhejiang Yuexiu University, China

Received: July 24, 2026

Accepted: August 26, 2026

Published: September 30, 2026

To cite this article: FU Zhihan & XU Yaowei. (2026). Temporal Reliability of ChatGPT-Generated Bilingual Texts: Implications for AI-Assisted Language Learning and Multilingual Information Access. *Asia-Pacific Journal of Humanities and Social Sciences*, 6(3), 266–276, DOI: 10.53789/j.1653–0465.2026.0603.027

To link to this article: <https://doi.org/10.53789/j.1653–0465.2026.0603.027>

Abstract: Large language models (LLMs) increasingly support multilingual information processing, translation, and language learning. However, continuously updated commercial services may produce unstable outputs over time. To investigate this temporal reliability, four successive ChatGPT snapshots (GPT-5.0, GPT-5.4, GPT-5.5, and GPT-5.6 Sol) were evaluated under identical zero-shot prompting conditions. The assessment utilized fixed Chinese source texts and professional English references. Three authentic texts from finance, technology, and public policy were analyzed via COMET, readability indices, sentence length, lexical diversity, adjacent overlap, and connective incidence. Results reveal no consistent improvement across updates. Instead, semantic reliability and textual characteristics related to accessibility followed diverging trajectories. Public-policy COMET scores improved, but later outputs became less readable. Conversely, technological translations featured shorter sentences and stronger lexical cohesion, yet their COMET scores declined after GPT-5.4. These findings underscore the necessity for continuous, domain-specific evaluations of LLM services in bilingual information access and language learning applications.

Keywords: ChatGPT; COMET; bilingual reading support; temporal reliability; readability; machine translation

Notes on the contributors: FU Zhihan is an undergraduate student at the School of English Studies, Zhejiang Yuexiu University, Shaoxing 312000, China. His academic interests include educational technology, AI-assisted language learning, large language model evaluation, and technology-enhanced bilingual learning. His email address is fzh_0813@qq.com. XU Yaowei is Director of the Foreign Language Experimental Teaching Center and Lecturer at the School of English Studies, Zhejiang Yuexiu University, Shaoxing 312000, China. He holds a PhD in Translation Studies from Shanghai International Studies University. His research interests include digital humanities, corpus-based translation studies, translation technology, and human-AI collaborative education. His email address is 20231027@zyufl.edu.cn.

1. Introduction

Large language models (LLMs) frequently provide translation assistance and bilingual reading support. Beyond mere translation, these generated texts are explored as resources for multilingual information access and AI-supported language learning. Consequently, the reliability of these outputs has become a critical concern in educational technology (Denny et al. 1; Huang et al. 1).

Unlike static educational software, commercial LLM services undergo continuous updates. Users submitting identical source texts and prompts may receive completely different outputs over time. Terminology, sentence structure, meaning preservation, and discourse organization can all shift unexpectedly. Therefore, a one-time evaluation of an LLM often loses validity following subsequent updates (Chen et al. 1).

Current research primarily compares different models or prompting strategies cross-sectionally. While longitudinal studies have tracked shifts in reasoning, safety, and instruction following, translation-based learning support remains under-researched. Furthermore, semantic metrics alone fail to capture changes in readability and broader linguistic characteristics.

To address this gap, our research investigates the stability of LLM-generated bilingual texts over time. We track four successive ChatGPT snapshots using fixed Chinese source texts, identical zero-shot prompts, and professional English references. By assessing financial, technological, and public-policy texts, we map temporal changes in both semantic reliability and textual accessibility. This approach challenges the assumption that newer updates are universally superior by revealing a complex and domain-specific evolutionary pattern.

2. Literature Review

2.1 *Largelanguage models and AI-generated bilingual reading support*

LLMs can generate fluent and context-sensitive translations, but professional texts continue to present difficulties involving terminology, numerical information, omission, and logical relations (Fu et al. 1). At the same time, LLM translation is increasingly used not only as a target-language product but also as a resource for bilingual reading and information access. Research on technology-enhanced EFL learning also shows how intelligent platforms can reshape learning environments and learner interaction (Gong 165). Educational-text research emphasizes two requirements: generated materials should preserve source meaning while presenting language at an appropriate level of difficulty (Denny et al. 1; Huang et al. 1). In the present study, the outputs are treated as potential bilingual reading-support texts rather than fully designed instructional materials.

2.2 *Fromtranslation quality assessment to educational text evaluation*

Translation technology has increasingly developed into a distinct area requiring its own theoretical framework and research pathways (Lan 47). Translation quality is a multidimensional construct involving more than intuitive judgments of fluency (House 1). COMET evaluates semantic correspondence by jointly considering the source text, candidate translation, and human reference, and it has shown stronger alignment with human judgment than metrics based mainly on surface overlap (Rei et al. 2685). Nevertheless, automatic metrics may overlook critical



errors involving omissions, restrictions, or altered relations, so one score cannot provide a complete account of translation quality (Saadany and Orăsan 1).

This limitation is particularly relevant when translations are used for reading support. Educational-text studies show that improving readability may reduce semantic preservation, while linguistically accurate texts may remain difficult for learners (Day et al. 1; Huang et al. 1). Readability, sentence structure, lexical diversity, and discourse cohesion can complement semantic evaluation (Graesser et al. 193). Xu and Fu (61) similarly found that semantic and linguistic indicators revealed different patterns across professional translation domains.

2.3 Temporal reliability of commercial large language models

Large language models have also reshaped language knowledge production and translation practice (Feng and Zhang 1). Commercial LLMs are frequently evaluated as if they were stable systems, although their interfaces and underlying services are regularly updated. Chen et al. (1) compared different temporal versions of GPT-3.5 and GPT-4 and found that some capabilities improved while others declined. Their findings show that one-time evaluation may not provide a lasting description of a commercial service.

Liu et al. (1) similarly reported non-monotonic changes across longitudinal versions of GPT, Llama, and Qwen. Improvements in one form of robustness did not guarantee gains in other dimensions. These findings suggest that release order, model size, and general capability cannot independently establish task-specific reliability.

Most temporal studies focus on reasoning, safety, or robustness rather than translation-based support. Xu and Fu (61) identified strong domain effects in a cross-sectional comparison of six LLMs, but did not track successive snapshots of one service. It remains unclear whether semantic reliability, readability, and discourse characteristics remain stable as a commercial LLM changes over time.

2.4 Summary and research gap

Translation evaluation, educational-text accessibility, and temporal LLM change have largely been studied separately. The present study connects these areas by evaluating four ChatGPT snapshots with fixed source texts, prompts, and professional references. COMET and complementary linguistic measures are used to identify domain-specific temporal changes in bilingual reading-support texts.

3. Methodology

3.1 Research questions and research design

This study adopts a controlled longitudinal design to examine the temporal reliability of AI-generated bilingual reading-support texts. Rather than comparing unrelated models, it tracks four successive snapshots of the same commercial service: GPT-5.0, GPT-5.4, GPT-5.5, and GPT-5.6 Sol. Each snapshot translated the same three Chinese texts under an identical zero-shot condition. The texts represent financial communication, technological documentation, and public policy. Professional human translations were used as references for semantic evaluation. The outputs were examined as potential bilingual reading-support texts rather than as fully designed instructional materials.

The study addresses the following research questions:

RQ1: How does the semantic reliability of Chinese-English translations vary across successive ChatGPT service snapshots and the three selected domains?

RQ2: How do the textual accessibility and linguistic characteristics of the translations change across successive ChatGPT snapshots?

RQ3: Do newer ChatGPT snapshots show consistent improvement, or do semantic reliability and textual accessibility follow different temporal patterns?

The controlled design enabled descriptive comparison of temporal and domain-specific variation.

3.2 *Materials and data collection*

Three authentic Chinese informational texts were selected: the newspaper articles Financial “Living Water” Should Flow Continuously and A Single Network, Multiple Services for Enterprises, and the technological manual Solution for Connecting Photovoltaic Power Plants to the FusionSolar Intelligent PV Management System. The texts contained approximately 800–1,100 Chinese characters and averaged about eight paragraphs.

The materials were not pedagogically adapted. Their generated translations were examined as potential resources for bilingual reading support. Professional English translations served as references for COMET scoring and linguistic comparison.

Data were collected from October 2025 to July 2026, covering GPT-5.0, GPT-5.4, GPT-5.5, and GPT-5.6 Sol. Each snapshot translated the three texts using the same prompt:

Please translate the following text into English: [source text]

Outputs were retained without linguistic revision. Only paragraph numbers, interface text, spacing, punctuation, and paragraph formatting were standardized.

3.3 *Evaluation measures*

Semantic reliability was assessed using reference-based COMET. Each evaluation included the Chinese source, the ChatGPT translation, and the professional English reference. The translations were processed in three or four aligned segments, after which a file-level COMET score was reported. Higher values indicated stronger semantic correspondence.

Textual accessibility was examined using Flesch Reading Ease, Flesch-Kincaid Grade Level, word count, and average sentence length. Lexical and discourse characteristics were measured through type-token ratio, adjacent content-word overlap, adjacent noun overlap, and connective incidence per 1,000 words. The same measures were calculated for the professional references.

TTR was interpreted cautiously because it is sensitive to text length. The human translations were used as domain-specific comparison points rather than absolute standards. These measures describe textual properties and do not directly measure learner performance.

3.4 *Data analysis*

Scores were compared descriptively within each domain and across the four snapshots. Because each domain contained only one source text, no inferential tests or cross-domain averages were used. COMET and linguistic measures were then compared to identify aligned or divergent temporal patterns, with the human translations serving as domain-specific reference points.

4. Results

4.1 Semantic reliability across ChatGPT snapshots

Table 1 presents the file-level COMET scores for the four ChatGPT service snapshots across the financial, technological, and public-policy domains. Figures 1 and 2 further display the temporal trajectories, the distribution of scores across domains, and the changes relative to the GPT-5.0 baseline.

The three domains showed different temporal trajectories. Finance remained nearly unchanged from GPT-5.0 to GPT-5.4, reached its highest score under GPT-5.5, and declined slightly under GPT-5.6 Sol. Technology initially increased and peaked under GPT-5.4, but then declined under GPT-5.5 and GPT-5.6 Sol. Its final score was also below the GPT-5.0 baseline. Public policy was the only domain to improve across every successive snapshot. Its COMET score increased from 0.8220 under GPT-5.0 to 0.8393 under GPT-5.6 Sol. Consequently, each domain reached its highest semantic score under a different snapshot: GPT-5.5 for finance, GPT-5.4 for technology, and GPT-5.6 Sol for public policy.

Public policy was the only domain to improve across every successive snapshot. Its COMET score increased from 0.8220 under GPT-5.0 to 0.8393 under GPT-5.6 Sol. Consequently, each domain reached its highest semantic score under a different snapshot: GPT-5.5 for finance, GPT-5.4 for technology, and GPT-5.6 Sol for public policy.

The direction of COMET change depended on the domain, and improvement was not sustained across all successive updates. The comparisons are descriptive because each domain contained only one source text.

ChatGPT snapshot	Finance	Technology	Public policy
GPT-5.0	0.8062	0.8572	0.8220
GPT-5.4	0.8064	0.8719	0.8267
GPT-5.5	0.8231	0.8313	0.8369
GPT-5.6 Sol	0.8184	0.8267	0.8393

Table 1. COMET scores across ChatGPT snapshots and domains

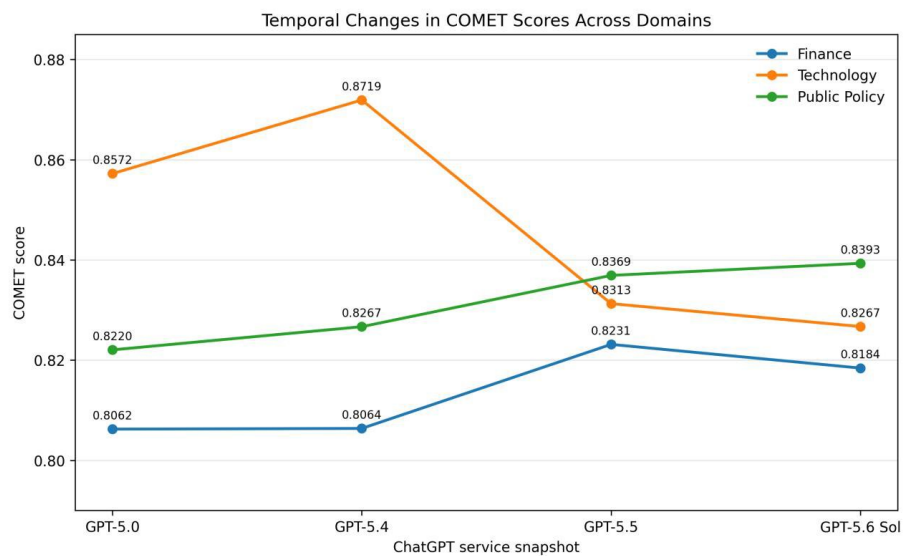


Figure 1. Temporal changes in COMET scores across domains

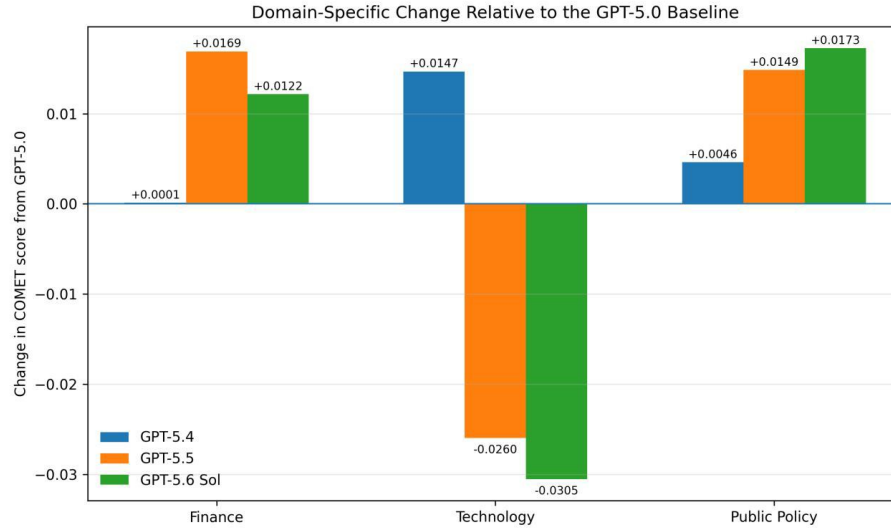


Figure 2. Domain-specific changes relative to the GPT-5.0 baseline

4.2 Textual accessibility and linguistic characteristics

Because the texts differed in domain and structure, linguistic measures were compared within each domain rather than averaged across domains. The professional translations served as domain-specific reference points.

4.2.1 Readability and structural features

Table 2 reports the Flesch Reading Ease score, Flesch-Kincaid Grade Level, word count, and average sentence length for the four ChatGPT snapshots and the professional reference translations.

Domain	Translation	FRE	FKGL	Word count	Average sentence length
Finance	GPT-5.0	19.69	15.58	619	23.81
Finance	GPT-5.4	23.57	15.56	680	25.19
Finance	GPT-5.5	22.31	15.41	704	23.47
Finance	GPT-5.6 Sol	18.07	16.20	685	24.46
Finance	Human reference	29.91	13.33	526	18.79
Technology	GPT-5.0	37.51	12.20	656	17.73
Technology	GPT-5.4	43.13	12.02	812	19.33
Technology	GPT-5.5	39.62	12.31	850	18.48
Technology	GPT-5.6 Sol	40.18	12.15	836	18.17
Technology	Human reference	37.87	15.20	786	29.11
Public policy	GPT-5.0	19.90	16.10	629	23.30
Public policy	GPT-5.4	26.18	15.29	757	24.42
Public policy	GPT-5.5	18.03	16.06	732	22.88
Public policy	GPT-5.6 Sol	18.41	16.24	750	24.19
Public policy	Human reference	30.54	14.37	710	22.19

Table 2. Readability and structural features of model and human translations

In finance, all four model translations were less readable than the professional reference. They had lower

Flesch Reading Ease scores, higher grade-level estimates, and sentences approximately five to six words longer. GPT-5.4 was the most readable model snapshot, whereas GPT-5.6 Sol had the lowest Reading Ease score.

Public policy showed a similar contrast. All model outputs were less readable than the reference, although GPT-5.4 was the closest. Readability declined sharply in GPT-5.5 and remained low in GPT-5.6 Sol. Because average sentence length remained relatively close to the reference, this difference cannot be explained by sentence length alone.

Technology produced a different pattern. Model outputs received similar or higher Reading Ease scores and substantially lower grade-level estimates than the professional translation. Their average sentences contained approximately 18–19 words, compared with 29.11 in the reference. Thus, the model outputs divided technical information into shorter sentence units. Across all three domains, GPT-5.4 had the highest Reading Ease score, but later snapshots did not maintain this pattern. Values above zero indicate that the model output received a higher Reading Ease score than the corresponding human reference. Positive values indicate longer model sentences, while negative values indicate shorter model sentences.

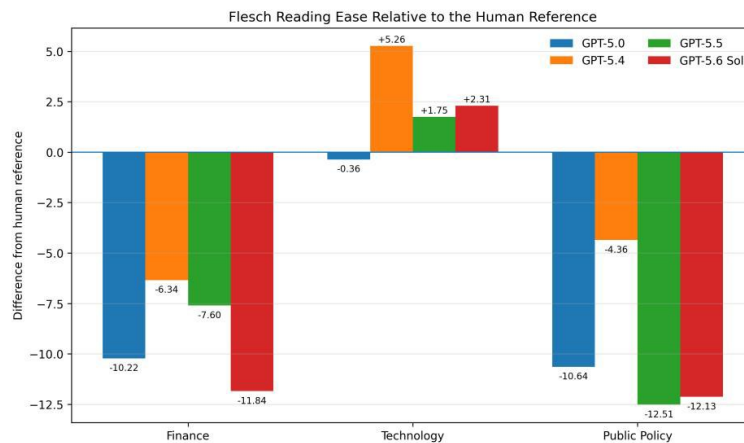


Figure 3. Flesch reading ease relative to the human reference

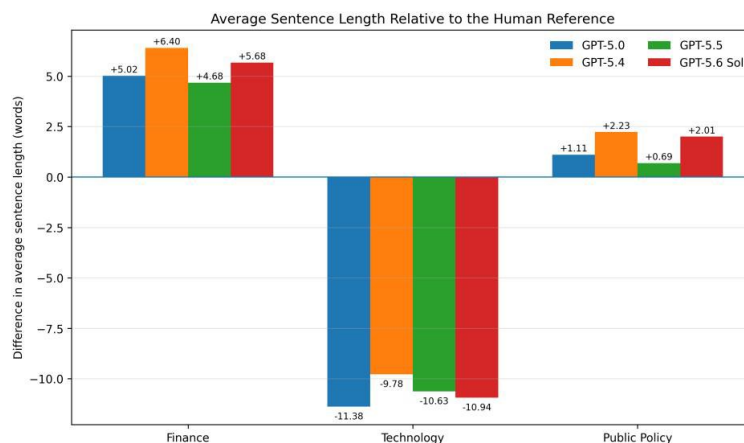


Figure 4. Average sentence length relative to the human reference

4.2.2 Lexical and discourse features

Table 3 reports type-token ratio, adjacent content-word overlap, adjacent noun overlap, and total connective

incidence. These measures were used to describe lexical variation and explicit discourse organization.

Domain	Translation	TTR	Content-word overlap	Noun overlap	Total connective incidence
Finance	GPT-5. 0	0. 520	0. 050	0. 039	67. 85
Finance	GPT-5. 4	0. 484	0. 076	0. 065	69. 12
Finance	GPT-5. 5	0. 442	0. 071	0. 046	62. 50
Finance	GPT-5. 6 Sol	0. 461	0. 092	0. 063	70. 07
Finance	Human reference	0. 563	0. 056	0. 063	51. 33
Technology	GPT-5. 0	0. 500	0. 148	0. 201	50. 30
Technology	GPT-5. 4	0. 458	0. 182	0. 255	55. 42
Technology	GPT-5. 5	0. 435	0. 192	0. 259	60. 00
Technology	GPT-5. 6 Sol	0. 445	0. 215	0. 292	59. 81
Technology	Human reference	0. 410	0. 190	0. 221	64. 89
Public policy	GPT-5. 0	0. 533	0. 093	0. 134	65. 18
Public policy	GPT-5. 4	0. 473	0. 147	0. 233	73. 98
Public policy	GPT-5. 5	0. 510	0. 116	0. 198	64. 21
Public policy	GPT-5. 6 Sol	0. 500	0. 106	0. 142	66. 67
Public policy	Human reference	0. 492	0. 146	0. 231	53. 52

Table 3. Lexical and discourse features of model and human translations

TTR did not show a consistent temporal direction and was affected by differences in output length. It was therefore treated only as a descriptive lexical indicator.

Adjacent lexical overlap varied by domain. In finance, content-word overlap generally increased but fluctuated between snapshots. Public policy reached its highest content-word and noun overlap under GPT-5. 4, with values close to the professional reference, before declining in the later snapshots.

Technology showed the clearest temporal pattern. Both content-word and noun overlap increased across the four snapshots. By GPT-5. 6 Sol, both values exceeded those of the professional reference, indicating stronger lexical recurrence between adjacent sentences.

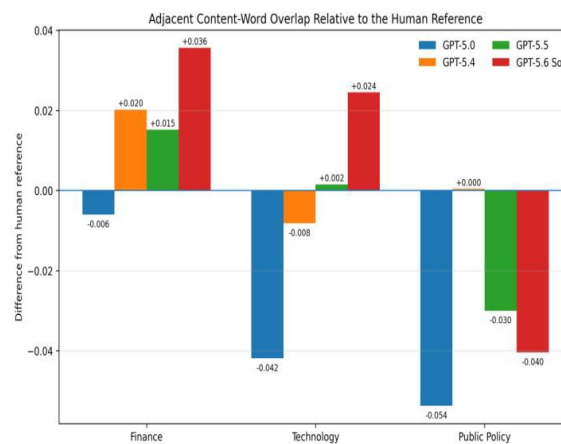


Figure 5. Adjacent content-word overlap relative to the human reference

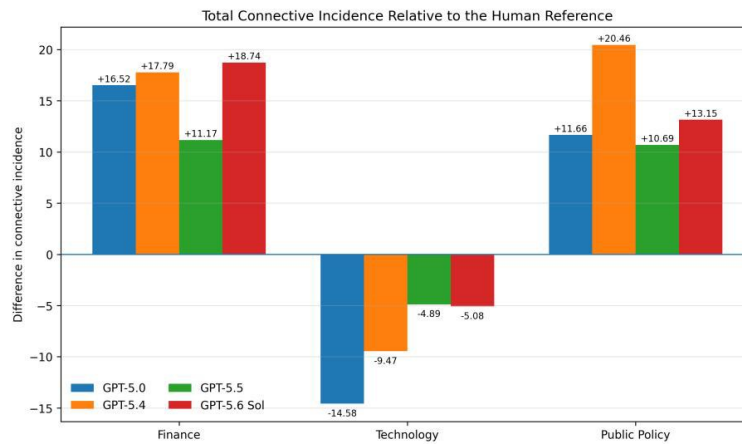


Figure 6. Totalconnective incidence relative to the human reference

Connective incidence was consistently higher than the human reference in finance and public policy, but lower in technology. This pattern describes differences in explicit discourse marking and should not be interpreted as a direct measure of discourse quality.

4.3 Cross-metric temporal patterns

COMET and linguistic measures did not follow a common temporal trajectory. In finance, semantic reliability peaked under GPT-5.5, whereas Reading Ease peaked under GPT-5.4. Technology initially showed aligned improvement under GPT-5.4, but later COMET scores declined while adjacent lexical overlap continued to increase. Public policy showed the clearest divergence: COMET improved across all four snapshots, but readability declined after GPT-5.4.

Comparison with the professional references also remained domain-dependent. Finance and public policy model outputs were generally more difficult, whereas the technological outputs used substantially shorter sentence units. These results answer RQ3 by showing that semantic reliability, readability, and discourse characteristics did not improve together, and that no snapshot was optimal across all domains and dimensions.

5. Discussion

5.1 Temporalreliability and non-monotonic model change

The non-monotonic results are consistent with previous longitudinal studies of commercial LLMs. Chen et al. (1) and Liu et al. (1) showed that service updates may improve some capabilities while weakening others. The present study extends this concern to Chinese-English translation, where each domain reached its highest semantic score under a different snapshot.

Commercial updates may simultaneously target instruction following, safety, efficiency, style, reasoning, and user preferences. Changes that improve general interaction may not produce equivalent gains in domain-specific translation. Terminology choice, sentence restructuring, or information compression may benefit one type of text while reducing correspondence in another. Because internal service updates are not fully disclosed,

the technical causes of the observed variation cannot be determined.

Model recency should not be treated as evidence of domain-specific reliability. Users interact with a changing service rather than a permanently fixed research model, so a translation workflow validated at one time may require renewed assessment after an update. This does not show that newer snapshots are generally inferior; it shows that improvement cannot be assumed across tasks and domains.

5.2 Semantic reliability and textual accessibility in bilingual reading-support texts

Semantic reliability and textual accessibility are related but distinct dimensions of bilingual reading-support texts. A translation must preserve source information, terminology, and rhetorical relations, but semantic correspondence alone does not determine how easily the English output can be processed. Sentence structure, lexical choice, repetition, and discourse marking shape the form in which translated information is presented.

The public-policy and technological texts illustrate opposite forms of divergence. Public-policy COMET scores increased while later outputs became less readable, showing that stronger meaning preservation may be accompanied by greater linguistic difficulty. In technology, shorter sentences and lower predicted grade levels did not prevent semantic reliability from declining after GPT-5.4. Surface simplification therefore did not guarantee reliable transfer of technical content.

The technological manual requires particular caution. Technical sentences may encode conditions, prohibitions, consequences, locations, and safety relations within one structure. Sentence splitting may reduce apparent complexity while redistributing modifiers or weakening the relationship between a condition and its consequence. Similarly, increasing lexical repetition can improve local continuity without confirming that restrictions or technical relations have been preserved accurately.

The metrics captured different dimensions of text quality. COMET describes semantic correspondence; readability and sentence measures describe predicted processing demands; lexical overlap and connective incidence describe aspects of discourse form. No single measure can determine whether a translated text is suitable for bilingual reading support. The appropriate balance depends on the domain, intended reader, and purpose of use.

5.3 Educational implications, limitations, and future research

These findings suggest that teachers, learners, and material developers should not assume that the newest service snapshot is automatically more reliable or accessible. Semantic reliability and textual accessibility should be checked separately, and professional review remains important for specialized texts containing terminology, restrictions, warnings, or complex logical relations.

Several limitations restrict the conclusions. Each domain contained only one source text, and only one output was collected for each snapshot-text combination. The observed variation may reflect both service changes and stochastic generation. The three texts were authentic informational materials rather than pedagogically adapted lessons, and the technological text was a professional safety and operation manual. The analysis relied on automatic indicators and did not include human error annotation, learner comprehension, reading-time data, or classroom observation. Readability and discourse measures consequently describe textual properties rather than learning outcomes.

Future studies should include multiple texts and repeated generations within each domain, combine automatic scores with human error analysis, and test the resulting materials with learners at different proficiency levels.



6. Conclusion

This study found no uniform improvement across four successive ChatGPT service snapshots. Semantic reliability changed in domain-specific and non-monotonic ways, while readability and discourse characteristics often followed different trajectories. Public-policy reliability increased as later outputs became less readable, whereas technological outputs showed shorter sentences and stronger lexical overlap without maintaining earlier COMET performance. These findings suggest that updated LLM services should not be evaluated only according to release order or general capability improvements. When LLM-generated bilingual texts are used for reading support and language learning purposes, their semantic reliability and linguistic characteristics require separate consideration. Continuous and domain-specific evaluation remains necessary because different types of texts may respond differently to service updates.

References

- Graesser, Arthur C., Danielle S. McNamara, Max M. Louwerse, and CAI Zhiqiang. "Coh-Metrix: Analysis of Text on Cohesion and Language." *Behavior Research Methods, Instruments, & Computers*, vol. 36, no. 2, 2004, pp. 193–202.
- CHEN Lingjiao, Matei Zaharia, & ZOU James. "How Is ChatGPT's Behavior Changing over Time?" *arXiv*, 2023. Preprint.
- FENG Zhiwei & ZHANG Dengke. "Artificial Intelligence and Large Language Models." *Foreign Language and Literature*, vol. 40, no. 3, 2024, pp. 1–29.
- FU Yuxin, SI Shijing, MAI Leyi, & LI Xi-ang. "FFN: A Fine-Grained Chinese-English Financial Domain Parallel Corpus." *arXiv*, 2024. Preprint.
- GONG Wen. "An Empirical Study of College English Smart Teaching Driven by Intelligent Technology." *Asia-Pacific Journal of Humanities and Social Sciences*, vol. 2, no. 3, 2022, pp. 165–172.
- Saadany, Hadeel, & Constantin Orăsan. "BLEU, METEOR, BERTScore: Evaluation of Metrics Performance in Assessing Critical Translation Errors in Sentiment-Oriented Text." *arXiv*, 2021. Preprint.
- HUANG Chieh-Yang, WEI Jing, & HUANG Ting-Hao Kenneth. "Generating Educational Materials with Different Levels of Readability Using LLMs." *arXiv*, 2024. Preprint.
- House, Juliane. *Translation Quality Assessment: Past and Present*. London: Routledge, 2015.
- LAN Hongjun. "Theoretical Research of Translation Technology: Problems and Pathways." *English Studies*, no. 1, 2023, pp. 47–58.
- LIU Yugeng, CONG Tianshuo, ZHAO Zhengyu, Michael Backes, SHEN Yun, & ZHANG Yang. "Robustness over Time: Understanding Adversarial Examples' Effectiveness on Longitudinal Versions of Large Language Models." *arXiv*, 2023. Preprint.
- Denny, Paul, Hassan Khosravi, Arto Hellas, Juho Leinonen, & Sami Sarsa. "Can We Trust AI-Generated Educational Content? Comparative Analysis of Human and AI-Generated Learning Resources." *arXiv*, 2023. Preprint.
- Rei, Ricardo, Craig Stewart, Ana C. Farinha, & Alon Lavie. "COMET: A Neural Framework for MT Evaluation." *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, 2020, pp. 2685–2702.
- Day, Stephanie L., Jacapo Cirica, Steven R. Clapp, Veronika Penkova, Amy E. Giroux, Abbey Banta, Catherine Bordeau, Poojitha Muttineni, & Ben D. Sawyer. "Evaluating GenAI for Simplifying Texts for Education: Improving Accuracy and Consistency for Enhanced Readability." *arXiv*, 2025. Preprint.
- XU Yaowei & FU Zhihan. "Comparing Chinese-English Translation Performance of Large Language Models: A Multidimensional Evaluation of Financial, Technological, and Political Texts." *Journal of Computer Science and Frontier Technologies*, vol. 3, no. 3, 2026, pp. 61–74.

(Editors: LIU Rong & LI Ruobing)