Toward Measurement Equivalence: LLM-Powered Translation of Critical Thinking Assessments

Euigyum Kim[1], Hyo Jeong Shin[1,*], Alina A. von Davier[2] and Salah Khalil[3]

[1]Sogang University, Seoul, South Korea
[2]Duolingo, Pittsburgh, PA, USA
[3]MACAT International Ltd., London, United Kingdom

Abstract

This study investigates the feasibility of large language models (LLMs) for item translation in cross-lingual assessment contexts with the aim of supporting test equivalence and measurement invariance. To this end, we propose an LLM-based multilingual translation framework and evaluate its utility. As a case study, we translated 48 English critical thinking assessment items into Bulgarian through the proposed framework, and evaluated the quality of the translations. The LLM translations demonstrated high linguistic and semantic preservation in the evaluation based on back-translation, and a human expert further confirmed their appropriateness for the target age group and cultural context. However, the expert revised the LLM translations to adopt a more academically precise style, favoring direct addressee forms and concise instructional expressions over the LLM's more descriptive tone. Despite these linguistic differences, the LLM translations demonstrated strong semantic preservation and linguistic accuracy. These findings suggest that the proposed framework effectively produces translation drafts suitable for testing contexts with minimal expert adaptation, highlighting its potential as an efficient and scalable approach to multilingual assessment development using LLMs.

Keywords

Multilingual Assessment, Measurement Equivalence, Agentic AI, Large Language Models, Machine Translation

1. Introduction

The emergence of foundation models, particularly large language models (LLMs), is transforming the landscape of natural language processing (NLP) research. By consolidating traditionally segmented NLP tasks, researchers no longer need to focus on complex model architecture design, extensive experimentation, and elaborate text preprocessing jobs to address individual NLP tasks [1]. Instead, end-to-end task execution with LLMs is now easily achievable through conversational interfaces such as ChatGPT or application programming interfaces (API). This shift has substantially broadened the accessibility of NLP technology and accelerated the broad adoption of LLMs in educational research, where a growing body of literature reports both their efficacy and critical concerns [2].

The field of educational measurement and evaluation is actively adopting artificial intelligence (AI) and LLM technologies [2, 3], leading to rapid growth in research on automated scoring (AS) and automatic item generation (AIG), while the use of computational psychometrics to analyze data collected through these AI-driven assessment processes has also been emerging [4]. However, studies on the use of LLMs for test adaptation are relatively scarce and applications to multilingual assessments remain particularly limited. One notable area of interest is the test adaptation that preserves measurement equivalence [5]. Although some studies have explored the use of machine translation to score multilingual responses in AS contexts [6, 7], research that attempts to translate assessment items using LLMs with the focus on test adaptation or measurement equivalence is yet limited.

In fact, even the most well-known international educational assessments, such as the Programme for International Student Assessment (PISA), human-translated items often struggle to maintain construct

LLM Psychometrics: Applying Assessment Theories to Evaluate LLMs and LLM-Supported Learners, LAK 2026 Workshop
*Corresponding author.
egkim@sogang.ac.kr (E. Kim); hshinedu@sogang.ac.kr (H. J. Shin); avondavier@duolingo.com (A. A. von Davier); salah@macat.com (S. Khalil)
0009-0007-3716-3945 (E. Kim); 0000-0003-2867-494X (H. J. Shin); 0000-0002-5505-5311 (A. A. von Davier)
© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2026-11-15

CEUR-WS.org/Vol-4330/paper2.pdf

and measurement equivalence across countries and languages when items are translated by humans, despite the implementation of a rigorous and systematic protocol for translation and adaptation [8, 9]. Consequently, multilingual item translation aimed at securing measurement invariance in large-scale international assessments has been considered a technically and methodologically challenging agenda. This instance underscores the complexity of cross-lingual item translation, while also presenting opportunities to leverage the potential of LLMs that demonstrate stable performance across languages that have been pre-trained with extensive multilingual corpus [10, 11].

In this regard, this study explores the feasibility of LLMs for multilingual translation of assessment items. To this end, we propose an agentic framework for LLM-based cross-lingual item translation and report findings from a case study with the translation of English critical thinking (CT) items into Bulgarian. Through this exploratory study, we aim to offer practical insights and key considerations for the adoption of LLM-based translation methodologies in cross-lingual comparative assessments, with particular focus on achieving effective test adaptation and measurement equivalence.

2. Methodology

2.1. Assessment Items as Source Data

The dataset for this study consists of 48 English CT assessment items developed for Year 6 students. These items were developed under the PACIER framework, which defines six sub-domains of CT [12]. All items take a multiple-choice format consisting of a short passage, a question, and a set of statements with options. An example structure of the CT item is presented in Figure 1.

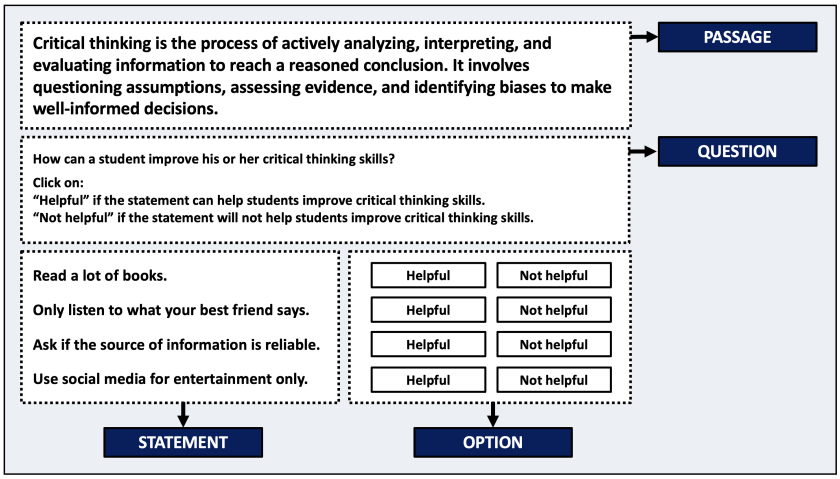


Figure 1: A fictitious English written CT assessment item based on MACAT's PACIER framework. Each item includes a passage and a question, followed by statements to which examinees select one of two options.

The items had previously undergone pilot testing and psychometric evaluation, demonstrating stable reliability and valid estimation of examinees' latent abilities [12]. Building on the established item parameters and ability estimates from prior research, the present study applies a cross-lingual translation framework to translate the items from English into Bulgarian, in preparation for the forthcoming administration of the assessment in Bulgaria. As this study is a work-in-progress, we focus mainly on linguistic equivalence, while the evaluation of measurement equivalence is reserved for future research.

2.2. Framework for Multilingual Item Translation

We propose a framework that leverages LLMs to achieve stable and linguistically and semantically equivalent translations for cross-lingual assessment items. The framework is designed to operate in situations where no gold-standard annotations or large parallel corpora exist for the target language, particularly for low-resource languages. As part of a human–AI collaborative model that integrates the

general capabilities of LLMs with the final judgment of human experts, this framework aims to maintain the linguistic equivalence of translated items through an iterative process of translation, evaluation, and revision, ultimately targeting a test adaptation that achieves measurement equivalence. An overview of the proposed framework is presented in Figure 2.

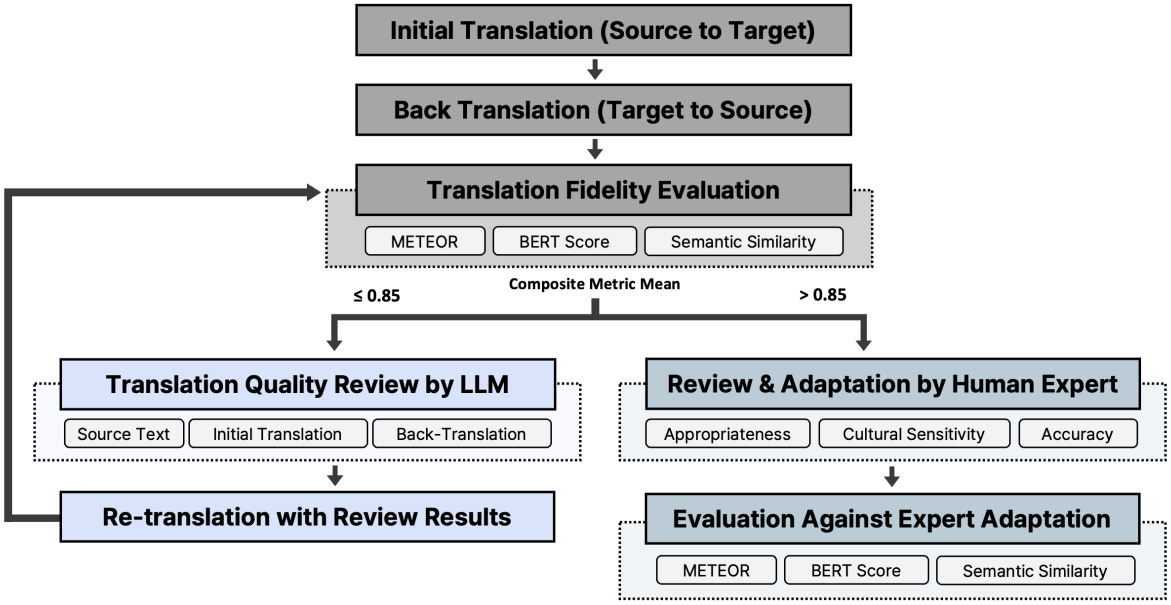


Figure 2: Workflow of LLM-powered translation framework for assessment items. From an agentic AI perspective, the framework follows a goal–action–evaluation–revision loop [13], in which the LLM performs translation, invokes predefined functions for quality evaluation, infers potential issues in the initial translation, and incorporates feedback into subsequent re-translation for self-correction and refinement. Rather than granting full agent autonomy, these agentic mechanisms in our proposed framework are constrained within a rule-based sequence under structured guidance, prioritizing human-governed control and stable performance.

.

The first stage of the framework involves generating an initial translation of the source item into the target language. In this research, we used GPT-5 main model that has demonstrated strong performance across various languages in the language evaluation benchmark [14]. In the initial forward-translation stage, the LLM produces a draft translation using zero-shot prompts that only incorporate the translation instruction and the CT assessment specifications. This output is not intended as a final translation, but rather as a preliminary draft to be refined in the following stages.

The second stage is back-translation to the original source language. When sufficient annotated data are unavailable, back-translation is commonly employed as an effective strategy [15]. By translating the draft translation back into the source language, it becomes possible to assess its linguistic alignment with the original one and detect major omissions or distortions of the translation at an early stage.

The third stage involves comparing and evaluating the original item and its back-translated version. Using metrics commonly employed in NLP translation tasks, we compute quantitative indices that capture meaning preservation and linguistic accuracy. Low-resource language translations often lack one-to-one correspondence with the source language and may differ substantially in cultural or stylistic conventions. In such situations, these metrics function as a criterion for determining whether the degree of variation remains within acceptable bounds for maintaining linguistic equivalence. Specifically, we select three complementary metrics and compute their average score for each item. Items whose averaged metric score exceeds a predefined threshold are accepted, whereas those falling below the threshold are flagged for further review by LLMs and carried forward to the re-translation stage.

The fourth and fifth stages involve LLM-based error diagnosis and guided re-translation. For translations flagged in the previous step, LLMs are serving as a judge [16], comparing the original source text, the initial translation, and the back-translation results to infer sources of error and propose optimal

revision directions. This feedback is then incorporated into the initial translation prompt as additional translation guidelines, forming an augmented prompt that enables a more informed re-translation rather than a simple repetition of the earlier process. The revised translation is re-evaluated using the same metrics, and this procedure is repeated until the output surpasses the predefined threshold.

The final stage is the review of human experts. Even when a translation has passed the threshold, human judgment remains essential, particularly for low-resource languages, where automated systems may not fully capture cultural norms, local usage, cultural sensitivity, or the level of adaptation required for young learners given national curricula and developmental considerations [11, 17]. Therefore, in the final stage, professional human translators evaluate the quality of the translation in terms of age appropriateness, cultural relevance, and linguistic precision. They then make final adaptations to ensure that the item is fully aligned with the educational and cultural context of the target country.

2.3. Metrics for the Comparison of Linguistic Properties

In order to evaluate similarity and translation fidelity between the back-translated and original source texts to identify translation errors, and further validate model performance by comparing the LLM-generated translation with expert-adapted translations using final human adaptations as the reference, this study employs three evaluation metrics commonly used in NLP translation tasks [18].

Table 1
Evaluation Metrics for LLM Translation Quality and Performance

Metric	Description
METEOR	– Extends n-gram–based evaluation by combining exact, stem, and synonym matching, while explicitly accounting for the word order [19]. – Suitable for assessing the accuracy of the translation and the lexical appropriateness.
BERTScore	– Computes similarity between the translated and reference texts using contextualized embeddings by pretrained BERT model at the token or word level. – Captures semantic equivalence and the preservation of the meaning, even when the surface forms of the words differ [20].
Semantic Similarity	– Projects translated and reference texts into a shared vector space using Transformer-based embeddings. – Quantifies sentence-level cosine similarity to estimate semantic fidelity [18].

By combining these three complementary metrics, this study evaluates the translation performance of the LLMs from surface-level lexical matching to contextual and semantic fidelity, and explores the systematic difference in translations between LLM and human expert.

3. Results

Building on the proposed framework, we analyzed the translation fidelity between the English source items and their initial Bulgarian translations. This fidelity is assessed through back-translation analysis, professional human translator's qualitative evaluations and resulting revisions, and comparisons between the final adapted items and the initial LLM-generated translations. Then, we examined whether the LLM outputs are sufficiently robust to support linguistic equivalence and elicit the implications for LLM-powered test adaptations.

3.1. Evaluation of Back-Translation Results

To begin with, we generated English back-translations of the initial LLM-produced Bulgarian translations. We then evaluated the extent to which the original items' semantic and linguistic features were preserved. As presented in Table 2, in all 48 items and their components, the metrics exceeded the score of 0.95. For some item components including the case where METEOR score fell below 0.8, lower values appeared as

they reflect alternative phrasings rather than a significant decrease in translation quality. This claim is further supported by the extremely high scores observed for BERTScore and semantic similarity. Using the arithmetic mean of the three metrics as the decision criterion, all items exceeded the threshold 0.85, allowing us to proceed directly to expert review without additional LLM evaluation or re-translation stage. These results suggest that the LLM could produce translations with sufficient semantic fidelity, even without domain- or language-specific fine-tuning measures.

Table 2
Summary Statistics for Translation Fidelity Between Source Items and Back-Translation Results (N=48)

	METEOR				BERTScore				Semantic Similarity			
Item Type	Mean (SD)	Min	Median	Max	Mean (SD)	Min	Median	Max	Mean (SD)	Min	Median	Max
Passage	0.96 (0.07)	0.74	1.00	1.00	0.99 (0.01)	0.96	1.00	1.00	0.99 (0.01)	0.95	1.00	1.00
Question	0.95 (0.07)	0.76	1.00	1.00	0.98 (0.01)	0.95	0.98	1.00	0.99 (0.01)	0.95	1.00	1.00
Statement 1	0.99 (0.05)	0.67	1.00	1.00	1.00 (0.01)	0.95	1.00	1.00	1.00 (0.01)	0.93	1.00	1.00
Statement 2	0.98 (0.07)	0.60	1.00	1.00	1.00 (0.01)	0.92	1.00	1.00	1.00 (0.01)	0.94	1.00	1.00
Statement 3	0.99 (0.05)	0.67	1.00	1.00	1.00 (0.01)	0.97	1.00	1.00	1.00 (0.01)	0.93	1.00	1.00
Statement 4	0.98 (0.08)	0.52	1.00	1.00	1.00 (0.01)	0.94	1.00	1.00	0.99 (0.02)	0.89	1.00	1.00

3.2. Qualitative Review by Human Expert

With the initial forward translations generated by the LLM, we requested a Bulgarian professional translator to review the output. The expert provided item-level ratings and provided the final adapted versions for local test administration. Across 48 translated items, expert ratings showed complete agreement on age appropriateness and cultural sensitivity, with all items rated as highly satisfactory for both dimensions. For linguistic accuracy dimension, all 48 items were rated as neutral, with no items receiving satisfied or dissatisfied ratings. Table 3 summarizes the expert's qualitative evaluations and the resulting adaptations.

Table 3
Qualitative Comparison of LLM Translation and Expert's Adaptation

Category	Explanation	LLM Translation	Human Translation
Addressee Perspective	LLM addresses test-takers with a formal plural form, while the expert uses an informal singular form for a more direct tone.	• изберете • щракнете • прочетете	• избери • щракни • прочети
Lexical Precision	LLM tends to use simpler, more general vocabulary, whereas the expert favors more precise expressions.	• разгледаха • правят мед • учи умения • можем	• проучили • произвеждат мед • научат умения • позволява да
Instruction Clarity	LLM often expresses negation through full sentence forms, while the expert uses concise antonymic labels for clarity.	• Полезно/Не е полезно • Вярно/Не е вярно • Известно/Не е известно	• Полезно/Неполезно • Вярно/Невярно • Известно/Неизвестно

Specifically, as described in Table 3, a comparative analysis between the LLM's translation and the expert's adaptation reveals contrasts in terms of linguistic expression and stylistic conventions. First, while the expert chooses concise and accurate content words, the LLM phrases the sentence into more descriptive and explanatory ways. For example, in expert translation, expressions such as "**промените в околната среда**" (*environmental changes*) or "**произвеждат мед**" (*produce honey*) reflect precise language, whereas the LLM translations rely on more descriptive phrasings such as "**мястото се променя**" (*the place is changing*) or "**правят мед**" (*make honey*).

In addition, distinct stylistic variations appear in the interactive elements of the text. Regarding addressee perspective, the LLM employs a formal plural register such as "**изберете**", whereas the expert opts for an informal singular form, "**избери**" to establish a more direct tone. Similarly, regarding instruction clarity, the LLM often uses negations such as "**Не е вярно**" (*Not true*), while the expert prefers concise antonymic labels, "**Невярно**" (*False*) to ensure brevity and clarity. These translation differences are semantically identical, but they realized variations in the actual wording or sentence structure depending on the descriptiveness.

3.3. Performance Evaluation of LLM Translation

Lastly, as shown in Table 4, we evaluated the quality of the initial LLM-generated translations using the final human translation as the reference. Overall, the LLM outputs show strong alignment with the expert's adaptations across all items and their components.

Table 4
Summary Statistics for LLM Translation Performance Relative to Expert References (N=48)

	METEOR				BERTScore				Semantic Similarity			
	Mean (SD)	Min	Median	Max	Mean (SD)	Min	Median	Max	Mean (SD)	Min	Median	Max
Passage	0.86 (0.11)	0.65	0.86	1.00	0.95 (0.04)	0.86	0.94	1.00	0.96 (0.04)	0.83	0.97	1.00
Question	0.80 (0.15)	0.21	0.80	1.00	0.94 (0.05)	0.69	0.94	1.00	0.95 (0.08)	0.47	0.97	1.00
Statement 1	0.90 (0.20)	0.19	1.00	1.00	0.97 (0.05)	0.83	1.00	1.00	0.97 (0.06)	0.72	1.00	1.00
Statement 2	0.91 (0.17)	0.36	1.00	1.00	0.98 (0.04)	0.85	1.00	1.00	0.97 (0.07)	0.63	0.99	1.00
Statement 3	0.93 (0.13)	0.49	0.99	1.00	0.98 (0.04)	0.87	1.00	1.00	0.98 (0.04)	0.78	1.00	1.00
Statement 4	0.89 (0.18)	0.32	1.00	1.00	0.97 (0.05)	0.85	1.00	1.00	0.97 (0.05)	0.81	1.00	1.00

Specifically, BERTScore and semantic similarity remain consistently high, indicating robust semantic correspondence. Although METEOR scores are slightly lower than the semantic indices, all averages exceed 0.80, which is generally considered satisfactory. The relatively lower scores in this metric are primarily due to stylistic adaptations made by the human expert, as summarized in Table 3. However, considering the high rating scores assigned by the expert, it can be confirmed that the overall translation performance remains stable and substantial.

4. Discussions and Implications

This study proposes a framework for LLM-powered multilingual translation of assessment items and applies it to English-to-Bulgarian translations, with the ultimate goal of preserving measurement equivalence. Based on comparisons with the original text through back-translation, the LLM achieved satisfactory linguistic accuracy and semantic preservation without the need for further re-translation iterations. Subsequent reviews by human expert confirmed that the quality of translation is overall acceptable, also showing high correspondence compared to human-adapted reference translations. Overall, these results suggest that the proposed framework yields stable translation performance that aligns with the expert's perspective without the use of fine-tuning techniques or extensive training data, and holds potential for implementation in actual testing contexts.

Notably, consistent differences were observed between LLM-based translations and those produced by human experts. The LLM produced more elaborate and often less formal translations, while human experts tended to maintain a more formal and precise style. These observations align with the findings of Kim and Shin [17], who showed that LLMs tend to elaborate when translating English items into Korean. Their study found that incorporating the context of the national curriculum into the prompts allows the LLM to better reflect the developmental levels of the learners. Without curriculum context, the model generates translations that resemble the content-word-focused style of human experts. Interestingly, Korean experts favored more elaborated structures with modifiers, suggesting that translation preferences may reflect unique national assessment traditions and item-writing conventions. After

all, whether such stylistic difference enhances test experiences or introduces unnecessary linguistic demands remains questionable. In this regard, future work can examine whether any stylistic differences between human and LLM translations affect the measurement invariance in actual testing contexts.

In addition, the data accumulated from each stage of the proposed framework would also offer a valuable foundation for future research on LLM-powered multilingual translations. Indeed, the framework makes it possible to observe stepwise improvements in translation quality, and the intermediate outputs generated throughout this process together with human expert's translations can serve as instruction-tuning data or reinforcement learning from human feedback (RLHF) data for developing advanced translation models [21]. This would help LLM-based translation evolve into a more sustainable and explainable system in automated evaluation research.

Moreover, the proposed framework explores the feasibility of an agentic AI [13] for translation system by organizing the translation process around goal-directed action, the invocation of linguistic quality evaluation tools, self-reflection based on comparisons among the source text, the translated output, and its back-translation, and a feedback loop that connects evaluation outcomes to translation strategy refinement and subsequent action modification. Although the feedback and re-translation stages were not implemented in our setting due to the stable performance of the initial translations, the importance of self-evaluation and feedback-driven refinement may vary across language pairs, and the applicability of fully agent approaches would need further exploration under human-governed principles.

Lastly, 22 of the 48 items used in this study were generated through an LLM-based AIG system [12], suggesting the feasibility of psychometric investigation into an end-to-end pipeline that integrates AIG with automatic translation. Therefore, this study establishes a foundation for examining how reliability, validity, and comparability can be ensured in AI-based measurement systems amid the growing integration of LLMs into automated assessment practices.

5. Limitations and Future Directions

Although this study demonstrated the promise of LLM-based translation through the proposed framework, there are several limitations that guide future research. To begin with, while LLM-based translations and subsequent expert reviews yielded promising results, these findings are insufficient to establish cross-lingual measurement equivalence, thus necessitating empirical psychometric comparison through pilot tests as the next step. Given the observed differences in translation style between LLM-based and human translations, future research should empirically investigate whether such differences yield implications for psychometric properties of assessment items, test fairness, and score comparability.

Furthermore, the generalizability of our findings is limited due to the involvement of only one human professional translator in Bulgarian, the relatively small set of 48 items in CT assessment, and the lack of consideration for various surface-level textual differences such as sentence length across individual items. Importantly, Bulgarian differs substantially from English in its grammatical and morphological structure in many ways. Therefore, when comparing these linguistic features, subsequent research should employ both common metrics and language-specific criteria for more detailed analysis.

Overall, this study demonstrates the potential of LLMs as tools for multilingual translation of assessment items. In an era of increasing use of LLMs and AI, these findings provide practical implications for developing responsible assessment practices based on human–AI collaboration aimed at ensuring fairness and comparability across countries and languages in cross-lingual assessment contexts.

Acknowledgments

This work was supported by the Ministry of Education of the Republic of Korea through the National Research Foundation of Korea (NRF-2025S1A5A8009125).

Declaration on Generative AI

The authors utilized ChatGPT and Gemini for the purpose of enhancing readability and checking grammar and spelling. All processes of the research and the writing of the manuscript were conducted entirely by the authors. The authors reviewed and edited the AI-generated suggestions as needed and take full responsibility for the content of the publication.

References

[1] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, E. Brynjolfsson, S. Buch, D. Card, R. Castellon, N. Chatterji, A. Chen, K. Creel, J. Q. Davis, D. Demszky, C. Donahue, M. Doumbouya, E. Durmus, S. Ermon, J. Etchemendy, K. Ethayarajh, L. Fei-Fei, C. Finn, T. Gale, L. Gillespie, K. Goel, N. Goodman, S. Grossman, N. Guha, T. Hashimoto, P. Henderson, J. Hewitt, D. E. Ho, J. Hong, K. Hsu, J. Huang, T. Icard, S. Jain, D. Jurafsky, P. Kalluri, S. Karamcheti, G. Keeling, F. Khani, O. Khattab, P. W. Koh, M. Krass, R. Krishna, R. Kuditipudi, A. Kumar, F. Ladhak, M. Lee, T. Lee, J. Leskovec, I. Levent, X. L. Li, X. Li, T. Ma, A. Malik, C. D. Manning, S. Mirchandani, E. Mitchell, Z. Munyikwa, S. Nair, A. Narayan, D. Narayanan, B. Newman, A. Nie, J. C. Niebles, H. Nilforoshan, J. Nyarko, G. Ogut, L. Orr, I. Papadimitriou, J. S. Park, C. Piech, E. Portelance, C. Potts, A. Raghunathan, R. Reich, H. Ren, F. Rong, Y. Roohani, C. Ruiz, J. Ryan, C. Ré, D. Sadigh, S. Sagawa, K. Santhanam, A. Shih, K. Srinivasan, A. Tamkin, R. Taori, A. W. Thomas, F. Tramèr, R. E. Wang, W. Wang, B. Wu, J. Wu, Y. Wu, S. M. Xie, M. Yasunaga, J. You, M. Zaharia, M. Zhang, T. Zhang, X. Zhang, Y. Zhang, L. Zheng, K. Zhou, P. Liang, On the opportunities and risks of foundation models, 2022. arXiv:2108.07258.

[2] J. Hao, A. A. von Davier, V. Yaneva, S. Lottridge, M. von Davier, D. J. Harris, Transforming assessment: The impacts and implications of large language models and generative AI, Educational Measurement: Issues and Practice (2024). doi:10.1111/emip.12602.

[3] A. A. von Davier, J. Burstein, AI in the assessment ecosystem: A human–centered AI perspective, in: Artificial Intelligence in Education: The Intersection of Technology and Pedagogy, volume 261 of *Intelligent Systems Reference Library*, Springer, Cham, 2024. doi:10.1007/978-3-031-71232-6_6.

[4] A. A. von Davier, R. J. Mislevy, J. Hao, Computational Psychometrics: New Methodologies for a New Generation of Digital Learning and Assessment, Springer, Cham, Switzerland, 2021. doi:10.1007/978-3-030-74394-9.

[5] K. Ercikan, H.-H. Por, Comparability in multilingual and multicultural assessment contexts, Comparability of large-scale educational assessments: Issues and recommendations (2020) 205–225.

[6] M. D. Shermis, International applications of automated scoring, in: Handbook of Automated Scoring, Chapman and Hall/CRC, 2020, pp. 113–132.

[7] J. Y. Jung, L. Tyack, M. von Davier, Combining machine translation and automated scoring in international large-scale assessments, Large-scale Assessments in Education 12 (2024). doi:10.1186/s40536-024-00199-7.

[8] A. Grisay, J. H. A. L. De Jong, E. Gebhardt, A. Berezner, B. Halleux-Monseur, Translation equivalence across PISA countries, Journal of Applied Measurement 8 (2007) 249–266.

[9] International Test Commission, ITC guidelines for translating and adapting tests (second edition), International Journal of Testing 18 (2018) 101–134. doi:10.1080/15305058.2017.1398166.

[10] K. S. Kalyan, A survey of GPT-3 family large language models including ChatGPT and GPT-4, 2023. arXiv:2310.12321.

[11] D. Ataman, A. Birch, N. Habash, M. Federico, P. Koehn, K. Cho, Machine translation in the era of large language models: A survey of historical and emerging problems, Information 16 (2025) 723. doi:10.3390/info16090723.

[12] H. J. Shin, S. Li, J. H. Ryoo, A. von Davier, T. Lubart, S. Khalil, The nature and measure of critical thinking: The PACIER framework and assessment, Journal of Intelligence 13 (2025). doi:10.3390/jintelligence13090113.

[13] A. Plaat, M. Van Duijn, N. Van Stein, M. Preuss, P. Van der Putten, K. J. Batenburg, Agentic large language models, a survey, Journal of Artificial Intelligence Research 84 (2025). doi:10.1613/jair.1.18675.

[14] OpenAI, GPT-5 System Card, Technical Report, OpenAI, 2025. URL: https://cdn.openai.com/gpt-5-system-card.pdf.

[15] M. Tars, T. Purason, A. Tättar, Teaching unseen low-resource languages to large translation models, in: Proceedings of the Seventh Conference on Machine Translation (WMT), Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid), 2022, pp. 375–380. URL: https://aclanthology.org/2022.wmt-1.33/.

[16] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, I. Stoica, Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena, 2023. arXiv:2306.05685.

[17] E. Kim, H. J. Shin, Leveraging large language model for automatic translation of educational content: Exploring the effectiveness of curriculum-aware prompt engineering, Korean Journal of Educational Research 63 (2025) 427–462. doi:10.30916/KERA.63.2.427.

[18] A. Mukherjee, M. Shrivastava, Lost in translation? Found in evaluation: A comprehensive survey on sentence-level translation evaluation, ACM Comput. Surv. 58 (2025). doi:10.1145/3735970.

[19] S. Banerjee, A. Lavie, METEOR: An automatic metric for MT evaluation with improved correlation with human judgments, in: Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, Association for Computational Linguistics, Ann Arbor, Michigan, 2005, pp. 65–72. URL: https://aclanthology.org/W05-0909/.

[20] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, BERTScore: Evaluating text generation with BERT, 2020. arXiv:1904.09675.

[21] P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, D. Amodei, Deep reinforcement learning from human preferences, in: Advances in Neural Information Processing Systems, 2017, pp. 4299–4307.