Scientific Research Publishing

Open Journal of Modern Linguistics, 2026, 16(2), 115-128
https://www.scirp.org/journal/ojml
ISSN Online: 2164-2834
ISSN Print: 2164-2818

Transformer-Based Automatic Item Generation for Course-Based Test Items: A Case Study of Translation Tasks in China's Context

Daohua Hu

School of Languages and Cultures (School of International Communication and Exchange), Shanghai University of Political Science and Law, Shanghai, China
Email: hudaohua01@126.com

How to cite this paper: Hu, D. H. (2026). Transformer-Based Automatic Item Generation for Course-Based Test Items: A Case Study of Translation Tasks in China's Context. *Open Journal of Modern Linguistics, 16,* 115-128.
https://doi.org/10.4236/ojml.2026.162009

Received: February 3, 2026
Accepted: March 16, 2026
Published: March 19, 2026

Copyright © 2026 by author(s) and Scientific Research Publishing Inc.
This work is licensed under the Creative Commons Attribution-NonCommercial International License (CC BY-NC 4.0).
http://creativecommons.org/licenses/by-nc/4.0/

Open Access

Abstract

In order to meet the rapidly increasing demand for item pools for large-scale assessments, automatic item generation (AIG) emerged about thirty years ago, using pre-programmed algorithms to automatically construct large numbers of test items with predictable item parameters. The rapid progress in natural language processing due to transformer networks has enabled large language models to handle a variety of natural language processing tasks, i.e., translation, text summarization, question answering, and writing text, at a level similar to humans. This study has carried out an empirical study on the human and transformer-based collaborative AIG framework for course-based item generation performances of several GenAI models for translation tasks of the English examination in China. The results show that: 1) Most GenAI models can successfully generate English-Chinese and Chinese-English sentence translation items. 2) Most GenAI models can generate both English-Chinese and Chinese-English text translation passages. 3) Readability of the generated passages is analyzed, and content validity of the generated sentence translation items and text translation passages is verified by subject matter experts. This study highlights that GenAI models help reduce teachers' burdens of repetitive and time-consuming human item writing tasks if handled properly.

Keywords

Course-Based Automatic Item Generation, Content Validity, DeepSeek, ERNIE, GenAI, Qwen, Readability, Translation Tasks

1. Introduction

Automatic item generation (AIG) is originally intended for large-scale assessments, aims to generate a large number of items for the item pool, and the generated items need automatic review before being put into use, so they are fairly suitable for test institutions. However, AIG for course-based assessments is rare.

The rapid progress in natural language processing has enabled large language models (LLMs) to handle a variety of natural language processing tasks (e.g., translation, text summarization, question answering, and writing text) at a level similar to humans (Sommer & Arendasy, 2025). Several researchers (e.g., Attali et al., 2022; Lee et al., 2023) have proposed including LLMs in the toolbox of test developers as a sub-variant of AIG. Generative artificial intelligence (GenAI) can generate a large number of test items according to the prompts input into the GenAI models, so it is feasible for course-based assessment item preparation (Song et al., 2025). To date, transformer-based AIG (TB-AIG) is feasible and greatly reduces the time needed to construct test items. Therefore, empirical research on the AIG capabilities of GenAI models for course-based assessments is of high significance. This study aims to investigate the AIG capabilities of certain GenAI models for translation tasks in China's context, i.e., sentence translation items and text translation passages.

2. Literature Review

2.1. Definition of AI and Its Application in Higher Education

The term artificial intelligence (AI) was coined in 1956 by John McCarthy, who used the term artificial intelligence for the first time (Russel & Norvig, 2010, p. 17).

> The study [of artificial intelligence] is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can, in principle, be so precisely described that a machine can be made to simulate it. An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves.

The application of AI in higher education can be traced back to the 1960s when early computer-assisted instruction systems were developed and employed in universities (Dhawan & Batra, 2021). In recent years, AI's potential in education has been increasingly recognized, and it has been adopted in various educational practices, such as in education (Chiu, 2023), higher education (Crompton & Burke, 2023), online education (Ouyang et al., 2022), etc.

2.2. AI Tools in Higher Education

The research concerning AI-based tools for teaching and learning in higher education has seen sustained exploration over the years (Crompton & Burke, 2023; Law, 2024; Zawacki-Richter et al., 2024).

The release of the ChatGPT LLM by OpenAI in 2022 particularly extended interest in AI to a broad civic audience and particularly to higher education institu-

tions. Kasneci et al. (2023) find that these comprehensive language models can serve as a supplement rather than a replacement for classroom instruction.

As a phenomenally popular AI, DeepSeek has rapidly captured widespread attention through its unique technological innovations and promotion strategies, becoming a cornerstone of digital infrastructure for enterprises. People are increasingly finding themselves enveloped by its influence: their educational tools, daily necessities, professional dependencies, and social interactions all intersect with DeepSeek (Lu, 2025).

Baker and Smith (2019) approach educational AI tools from three different perspectives: a) learner-facing, b) teacher-facing, and c) system-facing AI in education (AIEd). Among these, teacher-facing systems are used to support the teacher and reduce his/her workload by automating tasks of administration, assessment, feedback, etc. Of the 138 research articles Crompton and Burke (2023) examined, only 17% of them focused on instructors, which is far from sufficient.

Assessment and evaluation were the most common uses of AIEd in higher education. Lin and Chen (2024) reported that ChatGPT can generate acceptable multiple-choice items based on the given reading materials. Lu et al. (2021) used natural language processing to create a system that automatically created highly realistic short-answer questions. Sayin and Gierl (2024) used OpenAI GPT to generate reading comprehension items, and the generated items produced a similar level of difficulty and yielded strong discrimination power. Law (2024) reviewed publications published between 2017 and July 2023 and highlighted several research gaps, including the need for more empirical studies to assess the effectiveness and impact of GenAI tools. The studies above demonstrate that AI technologies can be employed to generate various test items, but few studies have investigated AIG for translation tasks.

2.3. Feasibility of Transformer-Based AIG for Course-Based Item Writing

2.3.1. Lower Technical Barriers for Course-Based AIG

AIG was originally developed for large-scale assessments, and the creation and manipulating of item models and/or templates are needed, which is beyond the technical capabilities of ordinary teachers, especially language teachers who mostly major in liberal arts. The introduction of GenAI models in AIG can help teachers to overcome these technical barriers.

2.3.2. Course-Based Teachers Are Qualified Reviewers of Test Items

The generated test items by AIG need reviews by subject matter experts (SMEs), and the designers for course-based examinations are always the teachers of the courses, so they are qualified SMEs for the review of the generated test items of their courses.

2.3.3. Reducing Teachers' Burden of Test Item Writing

The syllabus, knowledge, and skills of a course are relatively stable, and the course teachers have to prepare test items semester after semester. By introducing GenAI

models, course teachers can generate many test items by providing proper prompts, which reduces their burden of repetitive tasks.

3. Methodology

3.1. Research Questions

RQ1: Whether the English-Chinese and Chinese-English sentence translation items generated by GenAI models reflect the differences in language, culture, and thinking patterns between English and Chinese, and how about their content validity?

RQ2: Whether the English-Chinese and Chinese-English text translation passages generated by GenAI models reflect the differences in language, culture, and thinking patterns between English and Chinese, and what about their content validity and readability?

3.2. AIG Tasks and Prompts of Translation Tasks

Listening, speaking, reading, writing, and translating/interpreting are basic skills of English as a foreign language (EFL) learners, among which translation is of vital significance for information input and output. Sentence- and text-translation (English-Chinese; Chinese-English) are the main test types. Thus, we put prompts into the seven GenAI models to check whether they can finish the specific generative tasks and how well their performances are.

Task 1: Generating English-Chinese Sentence Translation Items

Prompt 1: Based on a 10,000-word vocabulary base, generate 10 English-Chinese sentence translation items, paying attention to the differences in language, culture, and thinking patterns between English and Chinese. Reference translation and an explanation are required.

Task 2: Generating Chinese-English Sentence Translation Items

Prompt 2: It is the same as Prompt 1 except for the translation direction of Chinese-English.

Task 3: Generating English-Chinese Text Translation Passages

Prompt 3: Based on a 10,000-word vocabulary base, generate two English-Chinese text translation passages. The length is about 150 words, paying attention to the differences in language, culture, and thinking patterns between English and Chinese. Reference translation and an explanation are required.

Task 4: Generating Chinese-English Text Translation Passages

Prompt 4: It is the same as Prompt 3 except for the translation direction of Chinese-English.

3.3. GenAI Tools Used in This Research

Three kinds of GenAI models are used in this research, including one localized model Deepseek-r1:1.5b, four online models, DeepSeek R1, DeepSeek V3.1, Ernie Bot and QwQ-Plus; and three advanced models, including Qwen-VL, ChatPDF, and QWQ-Plus.

4. Results and Discussion

4.1. Theoretical Bases for the Content Validity of Translation Tasks

Depending on the purpose of testing, tests can generally be categorized into aptitude test, achievement test, diagnostic test, proficiency test, and exit test. Among these, achievement test is always used to assess the students' success in learning a foreign language, and it is usually directly related to a specific foreign language course. Therefore, it has been suggested that achievement test should be based on the specific course syllabus and teaching materials (Shu & Zhuang, 2008, p. 167).

According to testing theory, a test has content validity if its content constitutes a representative sample of the language skills, structures, etc., with which it is meant to be concerned (Hughes, 2003, p. 26). To ensure content validity, the skills or constructs to be tested are typically outlined in detail for test developers' reference (Shu and Zhuang, 2008, p. 170).

Based on the course syllabus of translation and some authoritative textbooks in China (Zhang, 2018; Feng & Chen, 2008; Qin & Wang, 2010), the English-Chinese translation strategies mainly include the selection, extension, and commendatory or derogatory meaning of words, conversion of parts of speech, amplification, repetition, omission, affirmation and negation; division and combination of sentences; and the Chinese-English translation strategies include equivalent translation, amplification, omission, combination translation, conversion of parts of speech, transformation of expression, commendatory or derogatory translation, translation of Chinese idioms, proverbs, and two-part allegorical sayings, subject prominence and topic prominence, passive and active voice, cohesion and coherence, etc. The operational definition of content validity for translation tasks refers to how well the task items cover all relevant parts of the construct of translation competence.

4.2. Course Teachers as SMEs for the Test Items

In order to review the content validity of the generated sentence translation items and text translation passages, two subject matter experts are invited. They hold Ph.D. degrees in linguistics and translation, respectively, with at least 20 years of teaching experience in universities related to translation and interpreting courses.

First, they are required to review the generated translation items and passages respectively according to the syllabus of English-Chinese and Chinese-English translation and interpreting courses. Second, they discuss and reach an agreement when the generated items are difficult to categorize. A checklist for RQ1 and RQ2 is as follows:

1) Does the test item reflect the difference between Chinese and English languages?

2) Does the test item reflect the difference between Chinese and English cultures?

3) Whether the test item reflects the differences in thinking-pattern features

between Chinese and English?

4) What specific translation strategy is used in the English-Chinese sentence translation item?

5) What specific translation strategy is used in the Chinese-English sentence translation item?

6) What are the text type and topic of the Chinese-English translation passage?

7) What is the text type and topic of the English-Chinese translation passage?

4.3. Content Validity of Sentence Translation Items

4.3.1. Content Validity of English-Chinese Sentence Translation Items

The generated English-Chinese sentence translation items by the seven GenAI models are listed in Table 1. As for task 1, Deepseek-r1:1.5b failed in the translation direction; what it generated were not English-Chinese but Chinese-English sentence translation items. The other six GenAI models completed task 1 fairly well, including a variety of topics, such as cultural image, technological term, legislative text, euphemism, etc. The translation skills mentioned above are employed in different translation items. The first English-Chinese sentence translation items generated by the six GenAI models are chosen at random and listed in Table 1 below.

Table 1. Generated English-Chinese sentence translation items.

Items/Topic GenAI Tools	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
Deepseek-r1:1.5b	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.
DeepSeek R1	Cultural difference	Slang	Passive voice	Participle structure	Idioms	Inverted structure	Metaphor	Idioms	Cultural comparison	Slang
DeepSeek V3	Cultural image	Tech term	Legislative text	Euphemism	Medical term	News Title	Literary rhetoric	Diplomacy	Philosophical term	Sports metaphor
ERNIE Bot	Tech	En-Ch difference	Lexical selection	Climate change	AI	Social media	En-Ch difference	E-books	Environmental protection	Internet
Qwen	Proverb	Idioms	Allusion	Allusion	Exaggerate	Thinking pattern	inverted Clausal	Idioms	Progressive relationship	Idioms
Qwen-VL	Idioms	Metaphor	Metaphor	Cultural difference	Idioms	Idioms	Allusion	Climate change	Slang	Allusion
QwQ-Plus	Circular economy	Educational thought	History	Arts	Social media	Economic policy	Life style	Human nature	Tech ethics	Globalization

(1) ST: The professor's lecture was such a white elephant that most students dozed off halfway through.

TT: 教授的讲座华而不实，大半学生听到一半就昏昏欲睡。(DeepSeek R1)

(2) ST: The arbitration award shall be final and binding on both parties.

TT: 仲裁裁决应是终局的，对双方均具有约束力。(DeepSeek V3)

(3) ST: The revolution in technology has led to a surge in remote work opportunities.

TT: 技术领域的革命已经导致远程工作机会激增。(ERNIE Bot)

(4) ST: Time is money.
TT: 一寸光阴一寸金。(Qwen)
(5) ST: He's a real night owl — still working at 2 a.m.
TT: 他是个十足的夜猫子，凌晨两点还在工作(Qwen-VL)
(6) ST: The circular economy aims to close the loop between production and consumption by reusing resources indefinitely.
TT: 循环经济的目标是通过无限循环利用资源，实现生产与消费之间的闭环。((Qwen-Plus)

In example 1, "white elephant" means "a possession that is useless or troublesome, especially one that is expensive to maintain or difficult to dispose of" (Siefring, 2004, p. 311), and there is no such cultural image in Chinese, so it is translated into Chinese "华而不实 hua er bu shi" by liberal translation.

In example 2, the usage of the modal verb shall in legislative text is the key, which means "1. has a duty to; more broadly, is required to." and "This is the mandatory sense that drafters typically intend and that courts typically uphold." (Garner, 2009, p. 1499) The modal verb shall is translated into Chinese "应 ying", showing its mandatory force.

In example 3, the noun surge means "a sudden increase in amount or number", and was translated into the Chinese verb "激增 ji zeng", which employed the conversion of parts of speech.

In example 4, the proverb "Time is money" means "time is a valuable resource, therefore it's better to do things as quickly as possible." (Siefring, 2004, p. 293), while its Chinese translation is an equivalent Chinese proverb "一寸光阴一寸金 yi cun guang yin yi cun jin", with a metaphor in it. What's more, the adapted translation was employed in example (4).

In example 5, the phrase "night owl" means "a person who enjoys staying up late at night", and its Chinese equivalent "夜猫子 ye mao zi" has the same meaning, which illustrates the cultural universals of animal metaphor in this respect.

In example 6, the logic in English is from aim to means, while the logic in Chinese is from aim to result via means. Therefore, the sentence order was adjusted in the English-Chinese translation process.

From the examples above, we can conclude that the content validity of the English-Chinese sentence translation items is justified.

4.3.2. Content Validity of Chinese-English Sentence Translation Items

As for task 2, Deepseek-r1:1.5b and ERNIE Bot failed. What the former generated were only some Chinese words or phrases as the source text items, with full Chinese translations. What the latter generated was "Please translate… (words or phrases) into English." The other five models completed task 2 fairly well, including a variety of topics, such as cultural images, technological terms, legislative texts, euphemisms, etc., whose topics are listed in **Table 2**. The translation skills mentioned above are employed in different translation items. The first Chinese-English sentence translation items generated by the five GenAI models are chosen and

listed as follows:

(7) ST: 塞翁失马，焉知非福。
TT: When the old man of Sai lost his horse, who could have known it was not a blessing in disguise? (DeepSeek R1)

In example 7, "塞翁失马 sai weng shi ma" is a Chinese allusion. The English translation retains the core image of "塞翁 sai weng (the old man on the frontier)" and "马 ma (his horse)", and then a Western expression, "a blessing in disguise", was employed to reveal its true meaning. The literal translation strategy was used by DeepSeek R1, but it mistranslated the word "塞 sai" as a surname.

(8) ST：这位作家是文坛常青树，笔耕不辍六十载。
TT: This writer is a literary evergreen who has kept writing diligently for sixty years. (DeepSeek V3.1)

In example 8, "常青树 chang qing shu" is a literary allusion, meaning that some writer is still popular even when he or she is fairly old. Chinese and English share this literary allusion, so the literal translation was employed by DeepSeek V3.1, which vividly maintains the plant metaphor.

(9) ST: 他虽然年纪大了，但精神矍铄，每天坚持晨跑五公里。
TT: Although he is advanced in age, he remains mentally and physically vigorous, jogging five kilometers every morning without fail. (Qwen)

In example 9, the Chinese words "年纪大了 nian ji da le" cannot be translated into "old", which is possibly derogatory, so the expression "advanced in age" was employed; and there is no equivalent expression for the Chinese four-character idiom "精神矍铄 jing shen jue shuo", so it was paraphrased as "mentally and physically vigorous".

(10) ST：他这人外强中干，一遇压力就原形毕露。
TT: He's all bark and no bite; under pressure, his true colors show. (Qwen-VL)

In example 10, the literal translation for the two Chinese four-character idioms "外强中干 wai qiang zhong gan" and "原形毕露 yuan xing bi lu" would be somewhat stiff, i.e., "be outwardly strong but inwardly weak" for the former (Wang et al., 1996: p. 401) and "be revealed for what one is" for the latter (Wang et al., 1996: p. 535), so liberal translation was employed by Qwen-VL.

(11) ST：中秋节吃月饼的习俗象征团圆和丰收。
TT: The Mid-Autumn Festival tradition of eating mooncakes symbolizes family reunion and a bountiful harvest. (QwQ-Plus)

In example 11, the traditional Chinese festival "中秋节 zhong qiu jie" symbolizes "family reunion", but not "bountiful harvest". The QwQ-Plu made a mistake here, which is called the hallucination of GenAI tools (Wang & Zhang, 2024).

Based on the analysis above, the content validity of the Chinese-English sentence translation items is also justified.

Table 2. Generated Chinese-English sentence translation items.

Items/Topic GenAI Tools	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
Deepseek-r1:1.5b	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.
DeepSeek R1	Literary allusion	Diplomacy	Business Negotiation	Tech Report	TCM Culture	Ancient Poems	Legislative article	Environmental Policy	Philosophy	News report
DeepSeek V3	Allusion	Poems	Political Terms	Metaphor	Political rhetoric	Allusion	Diplomacy	Philosophical thought	Economic policy	Traditional rituals
ERNIE Bot	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.
Qwen	Idioms	Cultural difference	Structural difference	Allusion	Cultural difference	Metaphor	Terminology	Linguistic difference	Thinking pattern	Ecological civilization
Qwen-VL	Idioms	Idioms	Allusion	Cultural comparison	Body metaphor	Political affairs	Idioms	Political affairs	Poems	Idioms
QwQ-Plus	Cultural customs	Idioms	Political terms	Tech terms	Legislative article	Literary metaphor	Business terms	Environmental terms	Philosophical terms	Chinese metaphor

4.4. Content Validity of the Generated Translation Passages

Undergraduate students majoring in translation and interpreting, after professional training, are capable of handling the translation of comprehensive texts in current affairs and news, as well as professional texts of general difficulty in fields such as politics, economy, society, and culture (Zhong & Zhao, 2015). Therefore, this section will check the content validity of the generated passages for English-Chinese and Chinese-English text translation, respectively, in terms of topics, discourse classification, and text readability. In order to assess text readability within a uniform framework, the source texts of task 3 and the target texts of task 4 are analyzed.

4.4.1. Topics, Discourse Classification, Statistics, Readability, and Content Validity of the Generated English-Chinese Translation Passages

After inputting prompt 3 into the GenAI models, the topics, discourse classification, passage statistics, and readability details of the generated passages are listed in Table 3.

A variety of topics are generated, including political philosophy, AI ethics, urban loneliness, digitization of traditional crafts, etc., which cover the comprehensive texts and professional texts required, and the GenAI can generate texts on various topics.

Discourse classification mainly consists of argumentation, exposition, narration, and description (Humes, 1983). The generated passages are all argumentation except text 7, and most of the generated argumentative passages include elements of claim, evidence, and conclusion (Rottenberg & Winchell, 2021, pp. 667-8).

Example (12):

[1] The rapid development of generative AI has sparked intense ethical

debates [2]. While these systems demonstrate remarkable creativity—producing original art or composing music—their "black box" nature raises concerns about accountability [3]. When an AI-generated image violates copyright laws, who bears responsibility [4]? The programmer, the user, or the algorithm itself [5]? This dilemma is compounded by cultural differences: Western frameworks emphasize individual liability, whereas Eastern philosophies often view responsibility as collective [6]. Furthermore, the anthropomorphic tendency to describe AI as "learning" or "thinking" obscures its mechanistic essence, potentially misleading the public [7]. Resolving these issues requires not only technical transparency but also cross-cultural dialogue to redefine ethical boundaries in the digital age. (DeepSeek R1)

In example 12, sentence 1 constitutes the claim, sentences 2 - 4, 5, and 6 are the evidence, and sentence 7 the conclusion.

Passage statistics cover length, number of sentences, words per sentence, and the percent of difficult words, respectively. Nine texts have more than 100 words; the other five texts have fewer than 100 words, especially texts 1 - 2 and texts 7 - 8, which have only about 30 or 50 words, respectively. Among them, nine texts have at least seven sentences, which are conducive to the complete structure of an argumentation, except for texts 1 - 2 and texts 7 - 8. The percent of difficult words ranges from 30% to 40%, except for texts 1 - 2 and texts 7 - 8.

As for text readability, Flesch Reading Ease is discussed (Xu, 2024). Flesch Reading Ease scores text readability on a 100-point scale, and there are seven levels of text difficulty: 0 - 29 (very difficult), 30 - 49 (difficult), 50 - 59 (fairly difficult), 60 - 69 (standard), 70 - 79 (fairly easy), 80 - 89 (easy), and 90 - 100 (very easy). According to the Flesch Reading Ease scores, texts 3 - 6 and texts 9 - 13 are very difficult, and text 14 is difficult.

Through the analysis of topics, discourse classification, statistics, and readability, we can conclude that the content validity of the generated English-Chinese translation passages is justified.

Table 3. Generated English-Chinese text translation passages.

Items GenAImodels	Text	Topic	Discourse Classification	Passage Statistics				Text Readability
				Total words	No. of Sentences	Words Per Sentence	Difficult Words (%)	Flesch Reading Ease
Deepseek-r1:1.5b	T1	Tech and Environment Protection	Argumentation	37	2	18.5	26.3	25
	T2	Political philosophy	Argumentation	30	2	15	43.3	0
DeepSeek R1	T3	AI Ethics	Argumentation	124	7	17.7	36.3	0.5
	T4	Cultural values shape pedagogy.	Argumentation	113	7	16.1	33.1	5
DeepSeek V3	T5	Tech Ethics	Argumentation	106	7	15.1	37.3	3.7
	T6	*Negotiating joint ventures*	Argumentation	89	5	17.8	34.4	12.6

Continued

ERNIE Bot	T7	Child's fascination with nature	Narration	50	3	16.7	14	69.8
	T8	Technology affect	Argumentation	57	4	14.3	22.8	51.6
Qwen	T9	Urban Loneliness	Argumentation	128	8	16	37.3	1.2
	T10	Cultural Perceptions of Time	Argumentation	126	8	15.8	33.6	14.7
Qwen-VL	T11	Urban Loneliness	Argumentation	128	9	14.2	37.3	3.8
	T12	Cultural Perceptions of Time	Argumentation	126	8	15.8	33.6	14.7
QwQ-Plus	T13	Digitization of traditional crafts	Argumentation	138	8	17.3	32.9	17.8
	T14	Social media and adolescents' self-identity	Argumentation	133	9	14.8	29.9	32.8

4.4.2. Topics, Discourse Classification, Statistics, Readability, and Content Validity of the Generated Chinese-English Translation Passages

After inputting prompt 4 into the GenAI models, the topics, discourse classification, passage statistics, and readability details of the generated passages are listed in Table 4.

A variety of topics are generated, including sharing bikes, the Dragon Boat Festival, AI in medical care, urbanization, and traditional villages, etc., which cover the comprehensive and professional texts required, and the GenAI tools can generate texts on various topics.

As for the discourse classification, the generated passages are all argumentative.

As for the passage statistics, ten texts are more than 100 words, and the other two texts are less than 100 words, except texts 1 - 2. Among them, ten texts have at least 6 sentences, which helps to guarantee the complete structure of an argumentation, except texts 7 - 8. The percentage of difficult words ranges roughly from 20% to 35%.

As for text readability, Flesch Reading Ease is discussed (Xu, 2024). According to the Flesch Reading Ease scores in Table 4, texts 3 - 6 and texts 9 - 11 are very difficult, and texts 12 - 14 are difficult.

Through the analysis of topics, discourse classification, and readability, we can conclude that the content validity of the generated Chinese-English translation passages is justified.

4.5. Summary of the Generated Results by GenAI Models

Summarizing the generated results from Tables 1-4, we can see that DeepSeek V3, DeepSeek R1, Qwen, Qwen-VL, and QwQ-Plus completed the four generating tasks fairly well. More than ninety percent of the test items can be used directly, with only some minor corrections needed, but they are models for general purposes, so more models for vertical sectors, i.e., higher education, should be developed for higher quality generating results.

Table 4. Generated Chinese-English text translation passages.

Items GenAImodels	Text	Topic	Discourse Classification	Passage Statistics				Text Readability
				Total words	No. of Sentences	Words Per Sentence	Difficult Words (%)	Flesch Reading Ease
Deepseek-r1:1.5b	T1	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.
	T2	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.
DeepSeek R1	T3	Sharing bike	Argumentation	117	7	16.7	23.7	27.1
	T4	Modernization of TCM	Argumentation	124	6	20.7	34.2	15.3
DeepSeek V3	T5	Dragon Boat Festival	Argumentation	142	6	23.7	29	20.7
	T6	Double Reduction Policy	Argumentation	138	7	19.7	25.7	28
ERNIE Bot	T7	AI in medical care	Argumentation	69	4	17.3	36.2	10.1
	T8	Green travel modes	Argumentation	77	4	19.3	32.5	14.8
Qwen	T9	Urbanization and traditional villages	Argumentation	148	8	18.5	26.5	30.1
	T10	Face and personal boundaries	Argumentation	125	9	13.9	33.3	15.1
Qwen-VL	T11	Filial piety	Argumentation	129	8	16.1	32.1	13.3
	T12	Slow living	Argumentation	136	8	17	18.4	42.6
QwQ-Plus	T13	Change of Spring Festival	Argumentation	137	6	22.8	27	31.8
	T14	AI in education	Argumentation	124	8	15.5	19	38.6

Deepseek-r1:1.5b is a lightweight model for decoder-only, and it is good at generating texts in Chinese, so it can explain to some degree why it failed the four tasks of translation questions.

Li Yanhong, the CEO of Baidu, put it, "Baidu's Ernie large model is a highly localized large language model in the Chinese market. This means that the Ernie that Baidu is currently developing will be more suitable for the Chinese language and the Chinese market than models developed abroad" (Yuan, 2023). Therefore, it does not excel at text generation in English, and it failed tasks 2 - 4 of translation questions.

5. Conclusion

This article has conducted empirical research on the AIG capability of seven GenAI models for four course-based translation tasks. The results show that five GenAI models can successfully complete the four tasks, while two models failed in all four tasks and three tasks respectively, which demonstrates that most of the investigated GenAI models can be used for course-based AIG in China's context if handled properly. With the advancements of LLMs and GenAI, they can be employed for course-based AIG test items, with course teachers as SMEs who review the factual information, content validity, and bias issues of the generated test items

or passages.

This research focuses only on AIG for course-based translation tasks, and other tasks (i.e., Listening Comprehension, Vocabulary and Structure, Reading Comprehension, Cloze, and Writing) of the English examination should be further investigated in the future to verify the applicability of GenAI models. From the history of technology development in education, the appropriate attitude is to embrace GenAI warmly, cultivate teachers and learners with GenAI competency, and guide their proper use.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

Attali, Y., Runge, A., LaFlair, G. T., Yancey, K., Goodwin, S., Park, Y. et al. (2022). The Interactive Reading Task: Transformer-Based Automatic Item Generation. *Frontiers in Artificial Intelligence, 5,* Article 903077. https://doi.org/10.3389/frai.2022.903077

Baker, T., & Smith, L. (2019). *Educ-AI-Tion Rebooted? Exploring the Future of Artificial Intelligence in Schools and Colleges.* Nesta Foundation Website. https://media.nesta.org.uk/documents/Future_of_AI_and_education_v5_WEB.pdf

Chiu, T. K. F. (2023). The Impact of Generative AI (GenAI) on Practices, Policies and Research Direction in Education: A Case of ChatGPT and Midjourney. *Interactive Learning Environments, 32,* 6187-6203. https://doi.org/10.1080/10494820.2023.2253861

Crompton, H., & Burke, D. (2023). Artificial Intelligence in Higher Education: The State of the Field. *International Journal of Educational Technology in Higher Education, 20,* Article No. 22. https://doi.org/10.1186/s41239-023-00392-8

Dhawan, S., & Batra, G. (2021). *Artificial Intelligence in Higher Education: Promises, Perils, and Perspectives.* ResearchGate. https://www.researchgate.net/publication/348910302

Feng, Q. H., & Chen, K. F. (2008). *An Elementary Coursebook on Chinese-English Translation.* Higher Education Press.

Garner, B. A. (2009). *Black's Law Dictionary* (9th ed., p. 1499). WEST.

Hughes, A. (2003). *Testing for Language Teachers* (2nd ed.). Cambridge University Press.

Humes, A. (1983). *Discourse Type and Composition Research.* Southwest Regional Laboratory Working Paper, WP 2-83 /01.

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F. et al. (2023). ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education. *Learning and Individual Differences, 103,* Article ID: 102274. https://doi.org/10.1016/j.lindif.2023.102274

Law, L. (2024). Application of Generative Artificial Intelligence (GenAI) in Language Teaching and Learning: A Scoping Literature Review. *Computers and Education Open, 6,* Article ID: 100174. https://doi.org/10.1016/j.caeo.2024.100174

Lee, P., Fyffe, S., Son, M., Jia, Z., & Yao, Z. (2023). A Paradigm Shift from "Human Writing" to "Machine Generation" in Personality Test Development: An Application of State-Of-The-Art Natural Language Processing. *Journal of Business and Psychology, 38,* 163-190. https://doi.org/10.1007/s10869-022-09864-6

Lin, Z., & Chen, H. (2024). Investigating the Capability of ChatGPT for Generating Multi-

ple-Choice Reading Comprehension Items. *System, 123,* Article ID: 103344. https://doi.org/10.1016/j.system.2024.103344

Lu, D. K. (2025). Subversion and Reconstruction: The "Butterfly Effect" in Education Triggered by DeepSeek and Measures of Response. *Journal of Xinjiang Normal University, 46,* 144-152.

Lu, O. H. T., Huang, A. Y. Q., Tsai, D. C. L., & Yang, S. J. H. (2021). Expert-Authored and Machine-Generated Short-Answer Questions for Assessing Students' Learning Performance. *Educational Technology & Society, 24,* 159-173.

Ouyang, F., Zheng, L., & Jiao, P. (2022). Artificial Intelligence in Online Higher Education: A Systematic Review of Empirical Research from 2011 to 2020. *Education and Information Technologies, 27,* 7893-7925. https://doi.org/10.1007/s10639-022-10925-9

Qin, H. W., & Wang, K. F. (2010). *Comparison and Translation Between English and Chinese.* Foreign Language Teaching and Research Press.

Rottenberg, A. T., & Winchell, D. H. (2021). *The Structure of Argument* (10th ed., pp. 667-668). Bedford/St. Martin's.

Russel, S., & Norvig, P. (2010). *Artificial Intelligence—A Modern Approach.* Pearson Education.

Sayin, A., & Gierl, M. (2024). Using OpenAI GPT to Generate Reading Comprehension Items. *Educational Measurement: Issues and Practice, 43,* 5-18. https://doi.org/10.1111/emip.12590

Shu, D. F., & Zhuang, Z. X. (2008). *Modern Foreign Language Teaching: Theory, Practice and Method (Revised Edition).* Shanghai Foreign Language Education Press.

Siefring, J. (2004). *The Oxford Dictionary of Idioms* (2nd ed.). Oxford University Press.

Sommer, M., & Arendasy, M. (2025). Automatic- and Transformer-Based Automatic Item Generation: A Critical Review. *Journal of Intelligence, 13,* Article 102. https://doi.org/10.3390/jintelligence13080102

Song, Y. S., Du, J. L., & Zheng, Q. H. (2025). Automatic Item Generation for Educational Assessments: A Systematic Literature Review. *Interactive Learning Environments, 33,* 1-20.

Wang, D. F. et al. (1996). *A Chinese-English Dictionary of Idioms.* Sichuan People's Publishing House.

Wang, Y., & Zhang, Z. (2024). The Worries and Solutions to ChatGPT AI Translation. *Chinese Translators Journal, No. 2,* 95-102.

Xu, J. J. (2024). *BFSU Readability Analyzer 3.* https://corpus.bfsu.edu.cn

Yuan, C. X. (2023). *Baidu's Net Profit Increased by 10% Last Year, Betting AI Era, Planning to Integrate Multiple Mainstream Businesses with Ernie Bot.* Securities Daily. http://www.zqrb.cn/gscy/qiyexinxi/2023-02-23/A1677080208200.html

Zawacki-Richter, O., Bai, J. Y. H., Lee, K., Slagter van Tryon, P. J., & Prinsloo, P. (2024). New Advances in Artificial Intelligence Applications in Higher Education? *International Journal of Educational Technology in Higher Education, 21,* Article No. 32. https://doi.org/10.1186/s41239-024-00464-3

Zhang, P. J. (2018). *A Course in English-Chinese Translation (Revised Edition).* Shanghai Foreign Language Education Press.

Zhong, W. H., & Zhao, J. F. (2015). Interpretation of Key Points in the National Standards for Teaching Quality of Bachelors' Translation and Interpreting Programs. *Foreign Language Teaching and Research (Bimonthly), 47,* 289-296.