SCIENTIFIC CULTURE

www.sci-cult.com

SCIENTIFIC CULTURE, Vol. 12, No. 2.1, (2026), pp. 770-780
Open Access. Online & Print

DOI: 10.5281/zenodo.122.12665

TASK-BASED EVALUATION OF AI TOOLS FOR CHINESE LANGUAGE EDUCATION: AN INTEGRATED AHP-TOPSIS APPROCH

Xin Liu[1] and Pitipong Yodmongkol[2*]

[1]College of Arts, Median and Technology, Chiang Mai University, Chiang Mai, Thailand
[2]College of Arts, Median and Technology, Chiang Mai University, Chiang Mai, Thailand

Received: 29/11/2025
Accepted: 09/12/2025

Corresponding Author: Pitipong Yodmongkol (pitipong.y@cmu.ac.th)

ABSTRACT

As artificial intelligence (AI) continues to transform education, selecting suitable tools for language learning has become increasingly complex. This study proposes a structured evaluation framework for Chinese language education by integrating the Analytic Hierarchy Process (AHP) and the Technique for Order Preference by Similarity to Ideal Solution (TOPSIS). Guided by Task-Based Language Teaching (TBLT), the Technology Acceptance Model (TAM), and the Innovation Diffusion Theory (IDT), six key evaluation criteria were established: Effectiveness, Usability, Interactivity, Adaptability, Feedback, and Cost. Data from 51 valid AHP questionnaires were analyzed to assess 10 AI-assisted language learning tools across six instructional tasks—grammar and vocabulary, listening, speaking, reading and Chinese character writing, writing skills, and integrated application. AHP identified the priority of each task-specific tool, while TOPSIS synthesized these results to generate an overall ranking. The findings show that Wordwall, Duolingo, and Mondly emerged as the most pedagogically versatile tools, offering balanced performance across multiple tasks. This integrated AHP–TOPSIS model provides a replicable and evidence-based framework for evaluating AI tools in education, supporting educators and policymakers in making informed, data-driven decisions about AI integration.

KEYWORDS: Artificial Intelligence (AI); Chinese Language Learning; Educational Technology Evaluation, Systematic Decision Making.

1. INTRODUCTION

In the digital age, language education has been transformed by the rapid development of information and communication technologies (ICTs), especially AI enhanced tools (Chen, 2024; Akram et al., 2022), which offer new ways to help learners in Chinese language learning, like giving timely feedback and adapting learning to individual needs (Zeng & Jiang, 2021; Liu et al., 2021).

However, despite policy support and growing awareness, Chinese language teachers find it hard to effectively integrate technology into instructional design, especially under task - based and communicative teaching paradigms that focus on interaction and context (Chen, 2024; Zeng & Jiang, 2021).

Research shows barriers at tool, teacher, and institutional levels (Akram et al., 2022; Balmes, 2022), and these are more obvious when teaching Chinese as a foreign language because of its unique linguistic and cultural aspects (Zeng & Jiang, 2021).

Moreover, boosting innovation in education requires a deeper pedagogical shift, not just in tools but in how technology reshapes teaching and learning relationships (Balmes, 2022; Hamzah et al., 2024).

Teachers are now expected to be learning facilitators rather than just knowledge transmitters, needing more autonomy, innovation, and collaboration (Kennedy, 2023).

But there's often a misalignment between technology design and real language learning tasks, which limits the pedagogical impact of many AI - based tools (Zeng & Jiang, 2021; Hamzah et al., 2024). Also, concerns about digital equity, over - reliance on automated systems, and lack of integration with curriculum objectives are common (Akram et al., 2022).

Given these challenges, there's a growing need for systematic, evidence - based ways to evaluate and select AI tools that support language learning goals while considering pedagogical and contextual realities.

As more AI tools emerge in education, educators and researchers find it harder to choose the most effective and appropriate ones for language instruction (Huang et al., 2023; Owan et al., 2023). Tools like automated writing evaluators, speech recognizers, and intelligent tutoring systems are everywhere, but the lack of standardized evaluation frameworks makes it difficult to assess their quality, alignment with learning objectives, and long - term impact (Alharbi, 2023; Berman et al., 2024).

Studies have found that AI tools vary greatly in design, functionality, and educational impact, leading to inconsistent learning outcomes (Danler et al., 2024; Lee & Lee, 2021).

Many tools have high usability, but their actual instructional value isn't always examined, showing the difference between usability and effectiveness (Berman et al., 2024).

This has led to more scholarly attention on developing rigorous, multi - dimensional evaluation strategies that consider context - specific needs, ethical issues, and pedagogical alignment (Owan et al., 2023).

To deal with these concerns, this study suggests a structured evaluation method based on the Analytic Hierarchy Process (AHP), which combines multiple evaluation criteria like pedagogical effectiveness, adaptability, interactivity, and user experience into a systematic decision - making framework.

This way, we can help educators make transparent, evidence - based decisions about AI tool selection, improving the quality and coherence of technology integration in Chinese language education.

Based on this idea, this study uses AHP as a multi - criteria decision - making (MCDM) method. Unlike subjective or random tool selection, AHP allows for structured, pairwise comparison of evaluation factors, ensuring transparency and consistency in decision - making (Kubat & Gurkan, 2021).

To make the evaluation framework more complete, this study is based on three established theories. First, Task - Based Language Teaching (TBLT) provides a pedagogical foundation by emphasizing the importance of matching technological tools with communicative, task - oriented learning activities.

As González - Lloret (2014) says, integrating technology into TBLT frameworks improves learner engagement, interaction, and the authenticity of language use in digital environments. Second, the Technology Acceptance Model (TAM) helps evaluate usability and perceived usefulness, which are key factors influencing whether teachers and students are willing to adopt an AI tool. Recent applications of TAM in language learning show that these factors greatly predict the intention to use educational technology (Alfadda & Mahdi, 2021).

Third, Innovation Diffusion Theory (IDT) complements TAM by focusing on the broader social and institutional factors affecting adoption. Concepts like compatibility, trialability, and relative advantage in IDT are particularly relevant in educational settings where institutional norms and teaching culture affect how AI tools are perceived and used

(Lee, Hsieh, & Hsu, 2011).

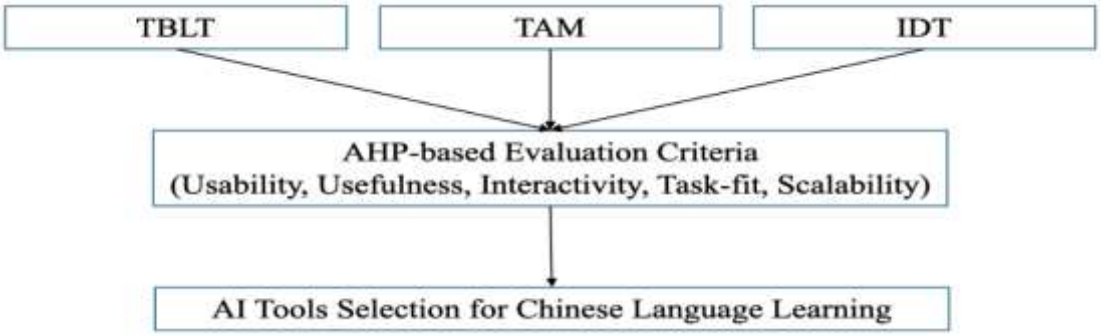

Figure 1: Research Conceptuall Framework.

These three frameworks (Figure. 1) together give a complete way to look at AI tools. We can see if they fit well with teaching, if people will accept them, and if they can be used widely in different institutions. And when we use the AHP model with these frameworks, we can make sure that both the teaching value and the practical usage in different situations are thought about when choosing AI tools.

Even though AI tools are becoming more common in education, the way people choose them is often not well - organized and not based on solid theories. Many educators and schools pick tools because they're popular, cheap, or seem new and cool. But they don't think enough about whether they match the teaching goals, are easy to use, or can be used long - term in a big way. This can lead to choosing tools that don't work well, students not being interested, and institutions wasting their resources.

So, the main goal of this study is to create a multi - criteria evaluation framework. This framework will help select AI tools that match specific tasks in Chinese language learning. By using ideas from teaching (TBLT), psychology (TAM), and culture and society (IDT) in an AHP - based model, the research wants to give a practical and theory - based method for choosing AI tools to Chinese language educators.

Based on this, the research question is:

How can educators in a clear and effective way evaluate and choose AI tools that best support different task - based learning goals in Chinese language education?

By answering this question, the study hopes to not only help test current AI tools better but also improve the ways AI is used in teaching second languages.

2. MATERIALS AND METHODS

2.1. Research Design

This study uses a multi - criteria decision - making (MCDM) method to assess and rank AI tools for Chinese language learning. As educational technology integration gets more complicated and depends a lot on the context, MCDM techniques are really helpful. They make decision - making across many criteria, which may even conflict with each other, more transparent and structured.

The core method used here is the Analytic Hierarchy Process (AHP). It's a well - known MCDM method that's good for comparing educational options based on both qualitative and quantitative judgments. AHP helps break down the decision - making problem into a hierarchy of criteria and alternatives. It allows decision - makers to assign relative weights through pairwise comparisons and check the consistency of these judgments.

To ensure the evaluation model is theoretically solid and relevant to the context, a mixed - method design was used. This included expert interviews to find out the evaluation criteria, AHP - based questionnaires to collect quantitative data, and hierarchical analysis to determine the final rankings of the AI tools being reviewed.

The research procedure was conducted in the following structured sequence:

- Establishing evaluation criteria: Six core evaluation dimensions were identified through literature review and expert consultation.
- Designing the pairwise comparison questionnaire: AHP matrices were constructed for each task category to collect participants' preferences among AI tools.
- Data collection: A total of 57 questionnaires were distributed, with 51 valid responses used for analysis after screening.
- Weight computation: Local and global weights of evaluation criteria and AI tools were calculated using eigenvector methods.
- Consistency analysis: A Consistency Ratio (CR) was computed for each matrix to ensure data reliability (acceptable threshold: CR < 0.10).
- Tool prioritization: Final priority rankings

were derived for AI tools within each learning task category and overall.

A Diagram Illustrating This Workflow Is Provided In Figure 2 To Support Reproducibility And Enhance Methodological Clarity.

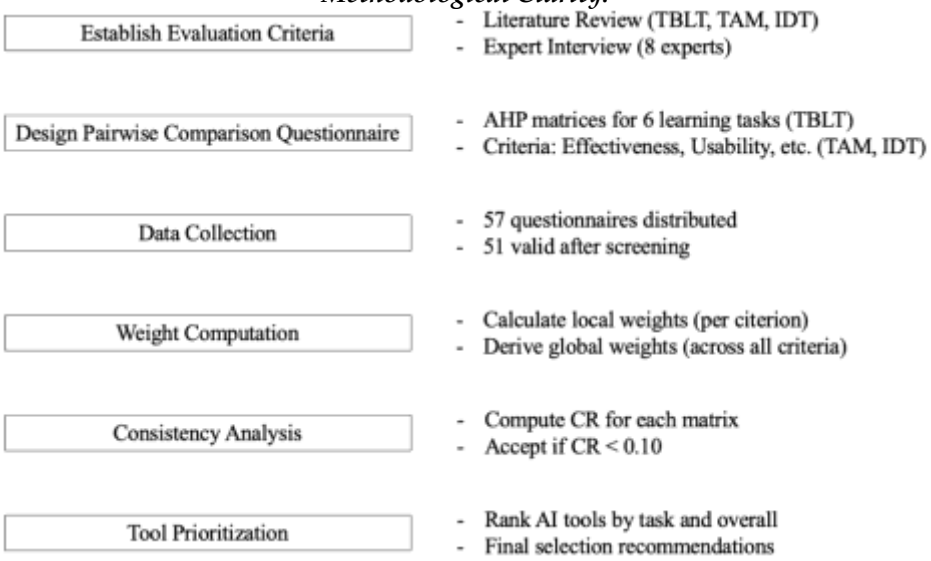

Figure 2: Workflow Of AHP-Based AI Tool Evaluation. This Diagram Illustrates The Multi-Phase Process Used In The Study, Including The Identification Of Evaluation Criteria, Design Of AHP Pairwise Comparison Questionnaires, Data Collection From Experts, Computation Of Weights And Consistency Ratios, And Final Integration Of AHP Results Into The TOPSIS Model.

This multi - phase design ensured methodological rigor and relevance. It helps educators and decision - makers make well - informed, evidence - based choices about AI integration in Chinese language education.

To make the evaluation more comprehensive and robust, the Technique for Order Preference by Similarity to Ideal Solution (TOPSIS) was added to the study. TOPSIS helps find the most suitable AI tools. It does this by looking for those closest to an ideal solution and farthest from a negative - ideal solution. It is especially helpful for ranking options when there is trade - offs among criteria. In this study, TOPSIS was applied to each tool's final AHP scores across six language learning tasks. The tasks and their weights are Grammar / Vocabulary (0.20), Listening (0.20), Speaking (0.15), Reading / Writing (0.15), Writing (0.15), and Integrated Application (0.15).

The combined AHP- TOPSIS approach supports both qualitative judgment and quantitative synthesis. It improves the overall reliability and practicality of choosing AI tools for Chinese language instruction.

2.2. Participants and Materials

To support the construction of the AHP evaluation framework, a panel of experts with diverse backgrounds was engaged in the initial phase. The group contributed rich experience in educational technology, Chinese language teaching, and the development of digital learning resources. Their collective expertise covered instructional design, digital learning environments, classroom pedagogy, and educational product innovation. Through semi-structured interviews and collaborative discussions, these experts played a vital role in identifying and refining the evaluation criteria. Their comprehensive insights ensured that the AHP model was closely aligned with pedagogical needs, user experience considerations, and the practical challenges of integrating new technologies into Chinese language instruction.

Then, 57 AHP-based questionnaires were sent out to a wider group of practitioners and researchers experienced in AI-supported language teaching. After a thorough screening for logical consistency and completeness, 51 valid responses were kept for analysis. The respondents were Chinese language instructors, curriculum developers, and postgraduate students in applied linguistics and educational technology.

To mirror the task - based approach in language learning, the evaluation spanned six distinct instructional task categories:

- Listening tasks (e.g., comprehension drills, audio-based input),
- Speaking tasks (e.g., pronunciation feedback, dialogue simulation),
- Reading tasks (e.g., vocabulary expansion,

skimming/scanning),
- Writing tasks (e.g., grammar correction, structure suggestion),
- Grammar and vocabulary practice (e.g., spaced repetition, game-based drills),
- Integrated application tasks (e.g., interactive platforms for multi-skill learning in real-world scenarios).

For each task category, a carefully curated list of AI tools was chosen, considering their availability, popularity among educators, and how well they matched the instructional goals. The selected tools included both general - purpose AI applications and specialized language - learning platforms, from automated feedback systems to intelligent speech analyzers.

After completing the AHP analysis, the final scores for each AI tool across the tasks were entered into the TOPSIS model. The importance of each task was reflected through pre - assigned weights. The TOPSIS analysis incorporated:
- Normalization of AHP results,
- Weighted normalization using task weights,
- Identification of positive and negative ideal solutions,
- Calculation of Euclidean distances to both ideals,
- Derivation of relative closeness scores (Ci) to rank tools.

2.3. *Evaluation Criteria and AHP Method*

The selection of evaluation criteria was cultivated from both theoretical and empirical sources. A comprehensive literature review identified key dimensions commonly used to assess the pedagogical value of AI tools in language education (e.g., Alharbi, 2023; Owan et al., 2023; Berman et al., 2024). These insights were further validated and contextualized through expert interviews with the eight panellists described in Section 2.2.

As a result, six evaluation criteria were finalized for inclusion in the AHP framework:
- Effectiveness – the extent to which the AI tool supports learning outcomes,
- Usability – the ease of use, interface intuitiveness, and overall accessibility,
- Interactivity – the level of learner-tool engagement and responsiveness,
- Adaptability – the tool's ability to personalize learning experiences,
- Feedback – the immediacy, quality, and usefulness of the feedback provided,
- Cost – the affordability and sustainability of using the tool.

These criteria were organized into a three-level AHP structure consisting of the overall goal (optimal AI tool selection), the evaluation criteria, and the AI tool alternatives under consideration (Figure 3).

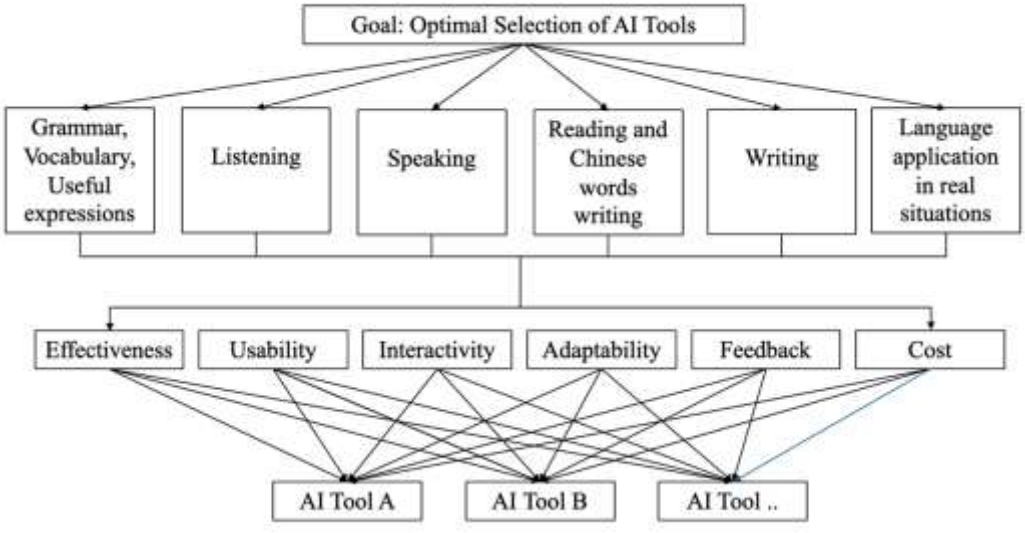

Figure 3: AHP Hierarchical Structure for AI Tool Evaluation.

To carry out the AHP analysis, this study followed the standard procedure proposed by Saaty (2008):
- Pairwise comparisons: Participants compared pairs of criteria using Saaty's 1–9 scale, where 1 indicates equal importance and 9 indicates extreme preference.
- Matrix construction: Each participant's input was used to build a reciprocal judgment

matrix.

- Weight calculation: The eigenvector method was applied to compute local and global weights.
- Consistency analysis: A Consistency Ratio (CR) was calculated for each matrix, with only those meeting the CR < 0.10 threshold included in the final analysis.
- Aggregation: The individual matrices were combined using the geometric mean method to establish a group consensus.

This structured approach ensured theoretical robustness and empirical validity in the selection process, enhancing the objectivity and transparency of the evaluation outcomes.

2.4. Data Collection and Analysis

The AHP analysis was conducted using data gathered from a meticulously designed structured questionnaire. It aimed to capture pairwise comparisons of AI tools across each language learning task category. Participants evaluated tool alternatives based on the six predefined AHP criteria, utilizing the standard nine-point scale of relative importance.

Out of the 57 questionnaires distributed, 51 were deemed valid after a rigorous two-step screening process. The first step weeded out incomplete submissions and those with substantial missing data. The second step excluded responses with high internal inconsistency, identified by a Consistency Ratio (CR) exceeding the 0.10 threshold (Saaty, 2008).

To analyse the collected data, Microsoft Excel and SPSS Statistics were used in tandem:

- Reciprocal matrices were manually constructed in Excel based on participants' pairwise comparison data,
- Local weights were calculated using the approximation method through column normalization and row averaging,
- Consistency Ratios were computed manually in Excel using standard AHP formulas,
- SPSS was used to conduct descriptive statistical analysis (e.g., mean, standard deviation) to identify patterns in participant responses,
- Aggregated weights were calculated using the geometric mean method, and final rankings of AI tools were determined within and across task categories.

While specialized AHP software might offer automation and visualization, the combination of Excel and SPSS used in this study provided a transparent, replicable, and statistically rigorous method for conducting the AHP analysis. This approach allowed for full control over the data processing and ensured that the analysis was both clear and verifiable.

To further refine the prioritization of AI tools, the Technique for Order Preference by Similarity to Ideal Solution (TOPSIS) was applied using the final AHP scores for each task category. The TOPSIS procedure followed these steps:

- Normalization of AHP Scores: Each tool's AHP score across the six task categories was normalized to ensure comparability. This step adjusted the scores to account for differences in scale and range.
- Weight Assignment: Predefined weights for each task category were applied to reflect their relative importance. These weights, derived from the expert consultation phase, ensured that the final ranking aligned with the overall goals of Chinese language instruction.
- Ideal and Negative-Ideal Solutions: The ideal solution was defined as the combination of the highest scores across all task categories, while the negative-ideal solution was the combination of the lowest scores. The distance of each tool's score from these two solutions was calculated.
- Closeness Coefficient Calculation: The closeness coefficient for each tool was determined by calculating the ratio of the distance from the negative-ideal solution to the distance from the ideal solution. This coefficient provided a measure of how close each tool was to the ideal solution relative to the negative-ideal solution.
- Ranking of Tools: The tools were ranked based on their closeness coefficients. The tool with the highest closeness coefficient was considered the most suitable, while the tool with the lowest coefficient was deemed the least suitable. This ranking provided a clear and objective order of preference for the AI tools in the context of Chinese language education.

This integrated approach enabled a more nuanced ranking system that accounts for pedagogical priorities and real-world instructional demands, offering a robust foundation for data-driven AI tool selection in language education.

3. RESEARCH RESULTS

3.1. AHP Expert Evaluation Results

To figure out how important each evaluation

criterion is for picking AI tools in Chinese language education, the Analytic Hierarchy Process (AHP) was used. We had six key criteria: Effectiveness, Usability, Interactivity, Adaptability, Feedback, and Cost.

Fifty-one experts, including language educators, AI developers, and instructional designers, compared each criterion with the others to see how important they are relative to one another. They used Saaty's 1-9 scale, where 1 means two criteria are equally important, and 9 means one criterion is a lot more important than the other.

Each expert compared the six criteria using Saaty's 1–9 scale. Their judgments were compiled into reciprocal matrices A=[aij], where:

$$a_{ij} = \text{importance of criterion } i \text{ relative to } j, \quad a_{ji} = \frac{1}{a_{ij}}, \quad a_{ii} = 1$$

All individual matrices were aggregated using the geometric mean method to produce a group judgment matrix. From this matrix, weights were derived through the following steps.

Step 1: Column Normalization

Each entry in the matrix was normalized column-wise:

$$a'_{ij} = \frac{a_{ij}}{\sum_{i=1}^{n} a_{ij}}, \quad n = 6$$

This ensures that the sum of each column is equal to 1. For example, if the original column sum for Usability was 15.235, and a_{ij}=2.563, then:

$$a'_{ij} = \frac{2.563}{15.235} = 0.1682$$

Step 2: Deriving the Priority Vector (Weights)

Next, each row in the normalized matrix was averaged to obtain the approximate eigenvector w, where:

$$w_i = \frac{1}{n} \sum_{j=1}^{n} a'_{ij}, \quad i = 1, 2, \ldots, 6$$

The final calculated average weights (Figure. 4) were:

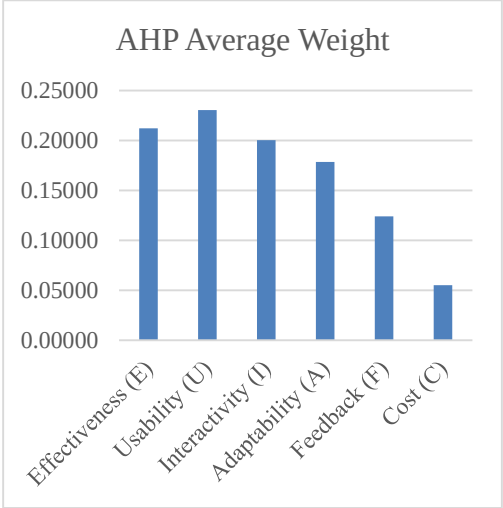

Figure 4: Average Weights Of Six Evaluation Criteria Derived From Expert Judgments Using The Analytic Hierarchy Process (Ahp). Usability And Effectiveness Were Rated As The Most Influential Criteria For Selecting Ai Tools, Whereas Cost Had The Lowest Relative Weight.

These results indicate that Usability and Effectiveness were perceived as the most influential criteria, while Cost had the least impact on decision-making.

Step 3: Weighted Sum Vector and λ_{max} Approximation

To evaluate consistency, the weighted sum vector A_w was calculated by multiplying the original matrix A with the priority vector w. Then, each element of the product was divided by the corresponding weight:

$$\lambda_i = \frac{(Aw)_i}{w_i}$$

An example for one criterion:

$(A_w)_1 = 1.3947$, $w_1 = 0.23022$

$$\lambda_1 = \frac{1.3947}{0.23022} \approx 6.059$$

This was repeated for all six criteria, and the average value was used to approximate

$$\lambda_{max} = \frac{1}{n}\sum_{i=1}^{n} \lambda_i = \frac{6.1735 + 6.1512 + 6.2201 + 6.1238 + 6.1827 + 6.1965}{6} \approx 6.174$$

Step 4: Consistency Index (CI)

The Consistency Index (CI) was then calculated as:

$$CI = \frac{\lambda_{max} - n}{n - 1} = \frac{6.174 - 6}{5} = \frac{0.174}{5} = 0.0349$$

Step 5: Consistency Ratio (CR).

Finally, the Consistency Ratio (CR) was computed using the Random Index (RI) value for

n=6n = 6n=6, which is 1.24:

$$CR = \frac{CI}{RI} = \frac{0.0349}{1.24} \approx 0.0281$$

The since the resulting CR was well below the recommended threshold of 0.10, the consistency of expert judgments was confirmable. Therefore, the derived weights are valid and reliable for further analysis.

This consistent and well-validated priority structure provided the foundation for the subsequent evaluation of AI tools within each language learning task. The weights were used to aggregate tool performances across criteria in both the AHP and TOPSIS analyses that followed.

4. AI TOOL EVALUATION BY LANGUAGE TASK

Based on the AHP-derived weights presented in Section 3.1, ten AI tools were evaluated across six core language learning tasks:

- Grammar and Vocabulary,
- Listening,
- Speaking,
- Reading and Chinese Word Writing,
- Writing
- Language Application in Real Situations.

Each task was treated as a separate decision-making context under the AHP framework.

The final AHP score of each tool iii in task ttt was computed using the following weighted summation formula:

$$\text{Score}_i^{(t)} = \sum_{j-1}^{6} w_j \cdot p_{ij}^{(t)}$$

Where:

w_j is the global weight of criterion j, derived from the AHP consistency-verified matrix,

$P(t)_{ij}$ is the local priority score of tools *i* on criterion *j* under task *t*.

All local priority values were obtained by normalizing the raw expert-assigned ratings for each tool per criterion within the task, and the global scores reflect the weighted utility of each tool in the given pedagogical context.

Table 1: Weighted AHP Scores Of Ten AI Tools Across Six Language-Learning Tasks, Showing Their Relative Strengths In Grammar, Listening, Speaking, Reading/Writing, And Integrated Applications.

AI Tools	AHP					
	Grammar, Vocabulary, Useful expressions	Listening	Speaking	Reading and Chinese words writing	Writing	Language application in real situations
Kahoot!	4.5422882	2.1800578	2.1612046	2.2627447	2.29265	2.1931966
Grammarly	2.2346188	2.2685477	2.2559481	2.164587	4.5232	2.1596781
iChineseReader	2.2056377	2.2360309	2.237477	4.4801802	2.1674382	2.1894274
Mondly	2.2093544	4.4411932	4.4923494	2.1796566	2.1799952	4.514126
Quizlet	4.4869242	2.216862	2.2344364	2.2552358	4.5081416	2.1663633

This Table I weighted aggregation ensures a comprehensive evaluation of tools, considering both their raw performance and the relative importance of each criterion.

Kahoot!, Quizlet, and Wordwall excelled in grammar and vocabulary tasks, indicating their strength in interactive, form - focused instruction. For listening tasks, Wordwall and Duolingo stood out due to their robust support for audio input and comprehension feedback. Speaking tasks were best supported by Mondly, Duolingo, and Speechace, which offer pronunciation assessment and conversational simulations. In reading and Chinese word writing tasks, Wordwall, Speechace, and iChineseReader scored highly, thanks to their vocabulary scaffolding and character - writing aids. For writing tasks, Quizlet, LangCorrect, and Grammarly performed strongly, with features like grammar correction, structured prompts, and handwriting support. Finally, in integrated application tasks, Duolingo, Wordwall, and Mondly showed the highest scores, aligning well with real - world language use.

4.1. Integrated Evaluation via TOPSIS

To create a comprehensive ranking of AI tools across all six task-based dimensions, this study used the Technique for Order Preference by Similarity to Ideal Solution (TOPSIS). TOPSIS is a well-recognized multi-criteria decision-making (MCDM) method. It evaluates tools based on how close they are to an ideal solution, considering all criteria weighted by

their pedagogical importance. Unlike AHP, which evaluates tools within a specific task, TOPSIS provides a unified ranking.

The integrated TOPSIS analysis followed a standard five-step procedure:

Step 1: Construct the decision matrix D=[xij], where xij is the AHP score of tools I on task dimension j.

Step 2: Normalize the decision matrix using vector normalization:

Step 3: Compute the weighted normalized matrix:

Where wj is the predefined weight for task j. In this study, the weights are as follows:

- Grammar and Vocabulary = 0.20
- Listening = 0.20
- Speaking = 0.15
- Reading and Word Writing = 0.15
- Writing = 0.15
- Language Application = 0.15

Identify the positive ideal solution (PIS) A+ and negative ideal solution (NIS) A−:

Calculate the Euclidean distance from PIS and NIS for each tool:

Compute the relative closeness to the ideal solution:

The final TOPSIS score Ci∈[0,1] indicates the relative performance of each AI tool, with values closer to 1 representing better overall suitability across all language tasks.

Figure 5. Overall performance of ten AI tools based on integrated AHP–TOPSIS analysis. Wordwall, Duolingo, and Mondly achieved the highest closeness coefficients, indicating superior adaptability across multiple language learning tasks.

The TOPSIS analysis results are summed up in Figure 5. Wordwall got the highest overall score (0.6375), showing it works really well for many teaching tasks. Duolingo (0.4994) and Mondly (0.4955) came next, both being good for productive skills and integrated language use. Quizlet was fourth (0.4240), mainly because it's great for vocabulary and writing. Kahoot! (0.3619) came after, thanks to its gamified grammar teaching. In the middle to lower end, iChineseReader (0.3017) and Speechace (0.2826) did moderately well, each being more useful for reading or pronunciation tasks. Skritter (0.2693), LangCorrect (0.2688), and Grammarly (0.2685) rounded out the rankings, suggesting they focus on narrower tasks or don't integrate as well into broader teaching needs. The results show that tools with more balanced skills, especially in receptive and integrated tasks, usually got higher TOPSIS scores. It also indicates that while some tools are great for individual tasks, they might not do as well when looked at from a comprehensive teaching design viewpoint.

5. DISCUSSION

The use of TOPSIS enables educators to make more comprehensive and balanced decisions. It allows them to consider the overall performance of each tool across diverse instructional tasks. Building on these results, this section discusses the key findings in relation to pedagogical applications, theoretical frameworks, practical implications, limitations, and directions for future research. These findings confirm that AI tools vary significantly in their task-based pedagogical value. While tools like Wordwall, Duolingo, and Mondly demonstrated strong cross-task adaptability and high overall performance in the TOPSIS analysis, others such as LangCorrect and Grammarly provided more specialized support in specific skill areas. These distinctions underscore the importance of aligning AI tool selection with instructional goals, task types, and user needs. Moreover, the observed results reinforce theoretical insights from TAM, TBLT, and IDT, validating that perceived usefulness, task alignment, and innovation characteristics jointly influence expert preference and predicted adoption. By integrating AHP and TOPSIS, this study not only provides a replicable framework for evidence-based AI tool selection but also bridges the gap between theoretical constructs and classroom decision-making.

These outcomes pave the way for the following conclusion, offering practical recommendations and reflecting on the broader educational implications of AI integration in Chinese language instruction. This research offers a clear evaluation model for selecting AI tools based on task alignment and usability. It helps educators move from informal tool adoption to more structured, evidence-based decisions. Task-specific rankings provide practical guidance—for example, Speechace for speaking and listening, or Wordwall for grammar. The framework also supports professional development by clarifying what makes a tool effective, guiding teachers in making informed decisions. Additionally, the model can inform institutional procurement, helping align technology choices with instructional needs and policy goals in diverse contexts like Thai secondary schools.

This study relied on expert ratings, which, despite consistency checks, may carry individual bias. The lack of student input limits understanding of actual classroom impact. The sample size, though adequate for AHP analysis, may not reflect wider learner

demographics or educational settings. Some tools also serve multiple functions, making strict task categorization challenging. As AI tools evolve rapidly, findings may need updating to remain relevant. The framework should be periodically reviewed to reflect changes in technology and pedagogy.

Future studies could explore learner perspectives, capturing engagement, motivation, and user experience alongside expert judgment. Combining AHP with system logs or behavioral data would enhance objectivity. Long-term studies could examine sustained tool effectiveness over time, especially as features and user familiarity change. Evaluations in varied linguistic and cultural settings would test the framework's adaptability. Emerging themes such as ethical use, privacy, and digital well-being should be integrated into future evaluation models to ensure tools are not only effective but also equitable and responsible.

In summary, despite the methodological rigor of this study, several limitations should be acknowledged. The evaluation relied mainly on expert judgments, which may introduce subjective bias. The sample size, though sufficient for AHP analysis, limits generalizability. The rapid evolution of AI tools may also affect the long-term relevance of the findings. Furthermore, functional overlap among tools can blur task distinctions. Recognizing these potential methodological constraints helps contextualize the findings and informs future efforts to refine and validate the proposed AHP–TOPSIS framework across broader educational settings.

Author Contribution: Liu Xin was responsible for data collection, AHP and TOPSIS analysis, visualization of results, and drafting the initial version of the manuscript. Pitipong conceptualized the study, designed the methodological framework, and contributed to theoretical guidance and critical revisions. Both authors reviewed and approved the final version of the paper.

Acknowledgements: This research was supported by the President's Scholarship from Chiang Mai University. And all the participants in this research.

REFERENCES

Akram, H., Abdelrady, A. H., Al-Adwan, A. S., & Ramzan, M. (2022). Teachers' perceptions of technology integration in teaching-learning practices: A systematic review. Frontiers in Psychology, 13, 920317.

Alfadda, H. A., & Mahdi, H. S. (2021). Measuring students' use of Zoom application in language course based on the Technology Acceptance Model (TAM). Journal of Psycholinguistic Research, 50(4), 883–900.

Alharbi, H. (2023). AI in the foreign language classroom: A pedagogical overview of automated writing assistance tools. Education Research International, 2023, Article ID 4253331.

Balmes, S. R. (2022). Technology integration and transformative innovation in education. International Journal of Research Publications, 106(1), 204–208.

Berman, G., Goyal, N., & Madaio, M. (2024). A scoping study of evaluation practices for responsible AI tools: Steps towards effectiveness evaluations. CHI Conference on Human Factors in Computing Systems, 1–24.

Chen, W. (2024). Advancements and challenges in technology-enhanced foreign language education in China. ERSS, 23, 52–64.

Danler, M., Hackl, W. O., Neururer, S. B., & Pfeifer, B. (2024). Quality and effectiveness of AI tools for students and researchers. Studies in Health Technology and Informatics, 313, 203–210. https://doi.org/10.3233/SHTI240038

González-Lloret, M. (2014). Tasks and technology in language learning: Exposure, affordance, and engagement. Language Learning & Technology, 18(1), 1–17.

Hamzah, F., Abdullah, A. H., & Ma, W. (2024). Advancing education through technology integration, innovative pedagogies and emerging trends: A systematic literature review. Journal of Advanced Research in Applied Sciences and Engineering Technology, 41(1), 44–63.

Huang, X., Zou, D., Cheng, G., Chen, X., & Xie, H. (2023). Trends, research issues and applications of artificial intelligence in language education. Educational Technology & Society, 26(1), 112–131.

Kennedy, G. M. (2023). Challenges of ICT integration in teachers' education: A case study of the University of Liberia. International Journal of Social Science and Education Research Studies, 3(5), 860–870.

Kubat, U., & Gurkan, B. (2021). An AHP-based approach for evaluating educational software. Education and Information Technologies, 26, 7899–7914.

Lee, H. S., & Lee, J. (2021). Applying artificial intelligence in physical education and future perspectives.

Sustainability, 13(1), 351.
Lee, Y. H., Hsieh, Y. C., & Hsu, C. N. (2011). Adding Innovation Diffusion Theory to the Technology Acceptance Model: Supporting employees' intentions to use e-learning systems. Educational Technology & Society, 14(4), 124–137.
Liu, Z., Liu, S., Xu, Y., & Wang, Q. (2021). ICT-supported collaborative learning and cognitive development. Frontiers in Psychology, 12.
Owan, V. J., Abang, K. B., Idika, D. O., Etta, E. O., & Bassey, B. A. (2023). Exploring the potential of artificial intelligence tools in educational measurement and assessment. EURASIA Journal of Mathematics, Science and Technology Education, 19(8), em2307.
Saaty, T. L. (2008). Decision making with the analytic hierarchy process. International Journal of Services Sciences, 1(1), 83–98.
Zeng, Y., & Jiang, W. (2021). Barriers to technology integration into teaching Chinese as a foreign language: A case study of Australian secondary schools. World Journal of Education, 11(5), 17–27.