



Cognitive Load Assessment and Optimization in College Japanese Classrooms Assisted by Lightweight Multilingual Models

Sirui Wang¹, Zhan Wang^{2,*}

¹Guangdong University of Science and Technology, Dongguan, China

²School of Biomedical Engineering, Guangdong Medical University, China

*Corresponding author: Zhan Wang, School of Biomedical Engineering, Guangdong Medical University.

Email: 123069887@qq.com, wangzhan@gdmu.edu.cn

Abstract: With the gradual adoption of multilingual large language models in foreign language classrooms, the cognitive load that learners experience when using intelligent tools has attracted increasing attention. Taking college Japanese-major classrooms as the empirical setting, this study builds an intelligent foreign language teaching assistant system based on a lightweight distilled multilingual model, collects learners' subjective cognitive load data through questionnaires at three time points (pre-test, mid-point, and post-test of a four-week AI-assisted practice period), and evaluates the changes in cognitive load while learners use AI for Japanese language output practice. The results show that the lightweight multilingual model effectively reduces part of learners' memory retrieval load, but problems such as inaccurate model feedback introduce additional cognitive consumption, and that the magnitude of these changes differs between higher- and lower-proficiency learners. Accordingly, improvement countermeasures are proposed from two levels: model parameter optimization and adjustment of teaching interaction modes. This study provides a reference for the development and application of multilingual intelligent teaching tools for foreign language education.

Keywords: lightweight multilingual model; cognitive load; Japanese language teaching; AI-assisted foreign language teaching; knowledge distillation

Introduction

Generative artificial intelligence is rapidly reshaping the landscape of second language learning. Previous studies have pointed out that the influence of AI-mediated tools on foreign language learning cognitive load is a "redistribution" rather than a simple "reduction": AI offloads lower-level encoding demands, but raises the requirements on central executive functions for evaluation and integration. Cognitive Load Theory provides a structured framework for analyzing this phenomenon, distinguishing intrinsic load, extraneous load, and germane load.

Meanwhile, knowledge distillation provides a technical path for building lightweight teaching models suitable for real-time classroom interaction. Through a teacher-student framework, knowledge distillation transfers the feature representations ("knowledge") learned by a structurally complex, parameter-heavy teacher model to a lightweight student model, thereby significantly reducing inference latency while maintaining teaching quality. Previous studies have adopted a two-stage sequence-level knowledge distillation framework to transfer the translation capability of a large multilingual teacher model to a compact student model, with the teacher model being the 1.3-billion-parameter mBART50 and the student model a 400-million-parameter Transformer.

Existing research on AI-assisted foreign language teaching still focuses mainly on English, and attention to Japanese and other languages remains very limited. The Japanese writing system comprises three scripts—kanji, hiragana, and katakana—and the working memory burden of language output practice differs from English scenarios. Taking college Japanese-major classrooms as the empirical setting, this paper builds an intelligent teaching assistant system with a lightweight distilled multilingual model and examines its effects on learners' cognitive load.

Accordingly, the specific objectives of this study are threefold: (1) to build a lightweight multilingual teaching assistant system for Japanese output practice based on knowledge distillation; (2) to evaluate how learners' cognitive load changes across four dimensions during a four-week period of AI-assisted Japanese output practice, both overall and across proficiency levels; and (3) to identify the specific factors that increase cognitive load and to propose corresponding optimization countermeasures at the model level and the pedagogical level.

Research Design

2.1 Theoretical foundation and instrument adaptation. This study takes Sweller's Cognitive Load Theory as the core framework. Previous research has further refined AI-mediated cognitive load into four observable dimensions: prompt management, critical evaluation, integrative synthesis, and authorial core processing. That study proposed a "load shifting" effect—large language models reduce content generation load, but significantly increase the load of critical evaluation and prompt management. This study draws on this four-factor structure and makes adaptive revisions according to the characteristics of Japanese output practice. In adapting the instrument, the four factors were re-specified



in terms of the observable demands of Japanese output: the prompt-management items were reframed as strategy regulation load (five items) and the critical-evaluation items as feedback evaluation load (five items), whereas the two text-production factors were re-specified as memory retrieval load (four items, from integrative synthesis) and grammar processing load (four items, from authorial core processing), yielding four subscales with 18 items in total.

Specifically, the original items, which were developed for English L2 writing, were reworded to match Japanese output practice: expressions referring to "writing in English" were replaced with "composing Japanese sentences and short texts," and the item wording was adapted to salient features of Japanese output, such as evaluating feedback on honorific usage and particle selection. The adapted items were reviewed by two university Japanese instructors for content validity, and they were pilot-tested with 12 second-year Japanese majors in a previous teaching cycle before formal administration; on the basis of the pilot feedback, the wording of two items that learners found ambiguous was revised.

2.2 Construction of the lightweight multilingual teaching system. The system is based on a knowledge distillation architecture. The teacher model is a large pre-trained model with multilingual understanding and generation capabilities (approximately 1.3 billion parameters); it provides two kinds of supervision—reference translations used as hard labels, and its output distributions used as soft labels. The student model is a compact Transformer optimized for Japanese teaching scenarios, with a parameter size of approximately 130 million, about one tenth of that of the teacher model, so that it can respond in real time on ordinary classroom computers without dedicated GPU servers. In distillation training, the total loss is a weighted combination of cross-entropy loss on hard labels and KL divergence between the teacher and student output distributions ($L = L_{CE} + \lambda \cdot L_{KD}$).

The weighting factor λ was set to 0.3 based on preliminary experiments on a held-out validation set: smaller values underused the knowledge of the teacher model, whereas larger values over-smoothed the student output distribution and reduced the precision of grammar feedback, and $\lambda = 0.3$ achieved the best trade-off between translation quality and feedback accuracy. In a preliminary comparison across $\lambda \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ on the validation set, $\lambda = 0.3$ yielded the lowest validation loss and the highest grammar-correction accuracy, and this setting was therefore retained for the full training run.

The system consists of three core modules—a multilingual capability evaluation module, a data generation engine, and a quality feedback loop—and provides learners with Japanese output practice support, including immediate grammar correction, expression suggestions, and writing feedback. Previous research on knowledge distillation in education has pointed out that transferring the capabilities of a large teacher model to a lightweight student model can provide high-quality AI services for educational scenarios without requiring powerful hardware support.

In terms of overall model performance, on a held-out test set of 1,200 Japanese sentences drawn from the course materials, the student model achieved 33.4 BLEU and 81.2% grammar-correction accuracy, compared with 36.8 BLEU and 87.6% for the teacher model; that is, it retained about 90% of the teacher model's translation quality while using roughly one tenth of its parameters. The student model generated feedback with an average latency of 0.9 seconds per response on an ordinary classroom computer, which made real-time interaction in class feasible.

2.3 Participants and data collection. The participants were 58 second-year Japanese majors at a university in China (41 female, 17 male; aged 19 to 21 years, $M = 20.2$, $SD = 0.7$), all native speakers of Chinese, whose Japanese proficiency was approximately at JLPT N4 to N3 level at the beginning of the study. All participants were enrolled in the same course and used the system as part of their regular coursework over four weeks. Because the study introduced no intervention beyond regular classroom practice and collected only anonymized questionnaire and interview data, no formal institutional ethics approval was obtained; the study protocol was instead reviewed and approved by the department and the course instructor, and all students were informed of the purpose of the study and provided informed consent before data collection. Participation was voluntary and anonymous, and course grades were not affected by participation.

Cognitive load was measured with a questionnaire based on the CL-AI-L2W scale (18 items, four-factor structure, 7-point Likert scale), adapted as described in Section 2.1. The internal consistency of the adapted scale was acceptable, with Cronbach's $\alpha = .82$ for memory retrieval, $.80$ for grammar processing, $.88$ for feedback evaluation, and $.84$ for strategy regulation, and $\alpha = .90$ for the overall scale. Data were collected at three time points: the pre-test before the AI-assisted practice, the mid-point of the four-week practice phase, and the post-test. Paired-samples t-tests (two-tailed) were used to compare pre- and post-test scores on each dimension, and independent-samples t-tests were used for between-group comparisons, with the significance level set at $p < .05$.

At the pre-test, learners were divided into a high-proficiency group and a low-proficiency group by a median split of their pre-test Japanese output scores; the high-proficiency group comprised 28 students and the low-proficiency group 27 students, and three learners whose scores fell exactly at the median were excluded from the between-group comparison. In addition to the scaled items, the post-test questionnaire contained open-ended questions asking learners to describe in detail the situations in which the AI system increased their cognitive burden; 52 of the 58 learners provided valid open-ended responses. Semi-structured interviews were then conducted with 12 learners, purposively sampled from both proficiency groups (six from each group), following a protocol covering feedback accuracy, feedback volume, and prompt-writing experience; each interview lasted approximately 30 minutes (range 22–38 minutes) and was audio-recorded and transcribed verbatim. The open-ended responses and interview transcripts were analyzed through thematic content analysis: two researchers coded the data independently, resolved discrepancies through discussion, and consolidated the initial codes into the themes reported in Section 3.3, achieving an inter-rater agreement of Cohen's $\kappa = .84$.

Results

3.1 Overall trends in cognitive load changes. Questionnaire data analysis showed that after four weeks of AI-assisted practice, learners exhibited differentiated change trends across the four cognitive load dimensions, as shown in Table 1:

Cognitive load dimension	Pre-test M (SD)	Post-test M (SD)
Memory retrieval load	4.82 (1.13)	3.67 (1.21)
Grammar processing load	4.95 (1.08)	4.12 (1.25)
Feedback evaluation load	4.31 (1.19)	5.08 (1.34)
Strategy regulation load	4.56 (1.22)	4.38 (1.31)

Table 1: Cognitive load scores before and after AI-assisted practice (M (SD))

Data collected at the mid-point of the practice phase showed that the mean scores of the four dimensions were 4.06 (SD = 1.18), 4.47 (SD = 1.19), 4.55 (SD = 1.27), and 4.49 (SD = 1.29) for memory retrieval, grammar processing, feedback evaluation, and strategy regulation, respectively. These mid-stage results indicate that the reductions in memory retrieval load and grammar processing load had already emerged by the mid-point, whereas the increase in feedback evaluation load accumulated gradually over the practice period, and strategy regulation load remained essentially unchanged throughout.

In the memory retrieval load dimension, learners' self-rated scores decreased from 4.82 (SD = 1.13) to 3.67 (SD = 1.21), a decrease of 23.9% ($p < .001$), the largest change among the four dimensions. Previous studies have similarly found that AI-assisted translation helps reduce surface-level difficulties in vocabulary selection and grammatical accuracy. In the grammar processing load dimension, the score decreased from 4.95 (SD = 1.08) to 4.12 (SD = 1.25), a decrease of 16.8% ($p < 0.05$).

In contrast, the feedback evaluation load dimension showed an upward trend, rising from 4.31 (SD = 1.19) to 5.08 (SD = 1.34), an increase of 17.9% ($p < .001$). This finding is consistent with previous research: AI-assisted translation tools do not eliminate mental effort, and students still need to carefully verify, evaluate, and revise AI-generated output. The change in the strategy regulation load dimension did not reach significance (4.56 $\rightarrow$ 4.38), which may be related to the insufficient design of metacognitive support functions in the system.

3.2 Differences between learners of different proficiency levels. Learners were divided into a high-proficiency group and a low-proficiency group according to pre-test Japanese output ability for comparative analysis. The increase in feedback evaluation load in the high-proficiency group (+0.42) was significantly lower than that in the low-proficiency group (+1.15, $p < 0.01$). This result echoes previous findings: ChatGPT feedback guides learners to notice various linguistic forms, but the depth of attention depends on learners' familiarity with these features, and the usefulness of AI feedback depends on how learners understand and evaluate the suggestions based on their existing knowledge. Learners in the low-proficiency group, with relatively weak language foundations, consumed more cognitive resources in the process of evaluating feedback. Given the limited subgroup sizes, this between-group result should be interpreted with appropriate caution (see the Conclusion).

3.3 Identification of factors increasing cognitive load. Through the analysis of open-ended questionnaire responses ($n = 52$) and interview records ($n = 12$), this study identified three types of AI-assisted factors that increase cognitive load.

First, inaccurate feedback or over-correction. Some learners reported that the feedback provided by the system on honorific usage and particle selection was inconsistent with classroom teaching. Previous studies have pointed out that inaccurate correction suggestions still exist in AI outputs, and that to maximize their effectiveness, appropriate prompt design and the cultivation of learners' AI literacy are indispensable. Second, feedback information overload. Multiple learners mentioned that the system annotated problems at multiple levels—vocabulary, grammar, and discourse structure—in a single feedback session. Third, prompt management burden. Some learners said that in the early stage they did not know how to effectively describe their practice needs to the system. Previous studies have similarly found that EFL students reported high cognitive load in writing tasks, especially during the prompt engineering stage.

Discussion and Countermeasures

4.1 Redistribution of cognitive load and implications for teaching design. The results of this study show that the lightweight multilingual model presents a cognitive load change pattern of "declining with rising" in Japanese output practice. Some scholars have proposed the concept of the "cognitive paradox": if students are not good at critical evaluation or prompt management, or if the system essentially completes the work without requiring evaluation, the total cognitive load may fall below the threshold required for germane processing, causing the task to become too easy and schema construction to stagnate. This means that teachers need to actively intervene in the AI-assisted learning process, so that feedback evaluation load is transformed into germane load rather than extraneous load.

4.2 Countermeasures for model parameter optimization. First, a teaching-oriented filtering mechanism should be introduced in the distillation training stage. The current quality evaluation module focuses mainly on semantic accuracy and language fluency, with insufficient weight on teaching suitability. It is suggested that a "teaching appropriateness" dimension be added in the data screening stage to filter out feedback samples that over-correct or exceed learners' level. Studies have pointed out that most pre-trained models are over-optimized on English data and have significant weaknesses in non-English capabilities. It is suggested that domain-adaptive training on Japanese teaching corpora be introduced during distillation to enhance the model's accuracy in Japanese grammar correction and expression suggestions. Second, a dynamic quality feedback loop should be established, incorporating high-quality feedback samples into the continuous training dataset.

4.3 Adjustment of teaching interaction modes. First, a staged feedback strategy should be implemented. Previous research has proposed the teaching idea of "hints before corrections": hints are first provided to guide learners to discover

problems independently, and correction suggestions are given afterwards. Specifically, feedback can be divided into three levels: the first level provides problem hints, the second level gives reference directions for revision, and the third level presents complete optimized expressions, so that learners obtain feedback layer by layer to avoid information overload. Second, prompt support functions should be optimized by providing simplified input interfaces and prompt templates for lower-proficiency learners. Finally, a teacher-mediated AI-assisted teaching process should be established, forming a collaborative teaching model of "AI providing immediate feedback + teachers providing in-depth guidance".

Conclusion

Taking college Japanese-major classrooms as the empirical setting, this study evaluated the changes in cognitive load in Japanese output practice assisted by a lightweight multilingual model. The main findings are as follows: the model effectively reduced memory retrieval load and grammar processing load, while feedback evaluation load increased, showing the characteristic of cognitive load redistribution; there were significant differences in cognitive load changes among learners of different proficiency levels; and inaccurate feedback, feedback information overload, and prompt management burden were the main factors increasing cognitive load. Based on these findings, improvement countermeasures were proposed from two levels: model parameter optimization and adjustment of teaching interaction modes.

Several limitations should be acknowledged when interpreting these findings. First, the study adopted a single-group pre-test–post-test design without a control group. Because the system was integrated into a regular credit-bearing course, withholding AI support from a randomized control group for four weeks was pedagogically impractical and would have been unfair to the students concerned; consequently, the observed changes cannot be attributed to the AI system alone. Second, all cognitive load data were self-reported, and self-report measures may be subject to social desirability and recall biases. Third, the sample of 58 students from a single university is relatively modest; no a priori power analysis was conducted, the proficiency subgroups were small, and the reported effect sizes should therefore be interpreted with caution and generalized carefully to other institutions and learner populations. Fourth, the four-week research period was short. Future research should employ multi-site samples with larger, a priori powered designs, include control or comparison conditions, combine behavioral data and physiological indicators for multimodal cognitive load assessment, and track learners over longer periods.

Acknowledgments: *The authors thank the students who participated in this study, as well as the colleagues who provided valuable comments on the questionnaire design and data analysis. This study is supported by the 2024 Quality Engineering Project of Guangdong University of Science and Technology (Grant No. GKZLGC2024376).*

REFERENCES

- [1]. Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12(2), 257-285.
- [2]. Yao, G., & Fan, L. (2025). Cognitive load scale for AI-assisted L2 writing: Scale development and validation. *Frontiers in Psychology*, 16, 1666974.
- [3]. Mannahali, M., & Malik, A. R. (2026). Exploring AI-assisted translation and cognitive load in L2 learning: Evidence from Indonesian university students. *Journal of Language Teaching and Research*, 17(5), 1-15.
- [4]. Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. *NIPS Deep Learning and Representation Learning Workshop*.
- [5]. Zhuang, R., & Zhang, W. (2025). Research on the construction of a small-sample foreign language listening comprehension model based on knowledge distillation. *Information Technology and Informatization*, (7), 153-156. (in Chinese)
- [6]. Lai, V. D., Ngo, N. T., Veyseh, A. P. B., Man, H., Dernoncourt, F., Bui, T., & Nguyen, T. H. (2023). ChatGPT beyond English: Towards a comprehensive evaluation of large language models in multilingual learning. *Findings of the Association for Computational Linguistics: EMNLP 2023*, 13171-13189.
- [7]. Zergui, A., & Adder, Z. (2026). Generative artificial intelligence and reducing the cognitive load in language learning: A study in the Algerian educational environment. *Revue Algérienne des Sciences du Langage*, 11(1), 296-310.
- [8]. Li, C., & Alias, A. (2025). Generative artificial intelligence and cognitive load in second language learning: A narrative review. *International Journal of Academic Research in Progressive Education and Development*, 14(4).
- [9]. Bhurle, M., Badhe, V., & Rajendran, R. (2025). Investigating the impact of AI-generated mnemonic illustrations on learning Japanese kanji. *Proceedings of the 33rd International Conference on Computers in Education (ICCE 2025)*.
- [10]. Woo, D. J., Wang, D., Guo, K., & Susanto, H. (2024). Teaching EFL students to write with ChatGPT: Students' motivation to learn, cognitive load, and satisfaction with the learning process. *Education and Information Technologies*, 29(18), 24963-24990.
- [11]. Hirata, M. (2025). Evaluating the use of a chatbot in an elementary Japanese course: Benefits, challenges, and student feedback. *Proceedings of the 13th International Symposium on Japanese Language Education and Japanese Studies*, 1, 12-38.
- [12]. Kabe, Y. (2025). The use of generative AI for Japanese language learning: Encouraging learners' engagement through practical applications in an intermediate-level course. *The Journal of Japanese Language Education Methods*, 32(1), 10-11.