Human Factors in Design, Engineering, and Computing, Vol. 238, 2026, 1045–1055
https://doi.org/10.54941/ahfe1008335

When More Becomes Too Much: Cognitive Support and Interpretive Constraint in Generative-AI-Assisted Reading of Chinese Classical Poetry

Rui Peng and Yuankai Lin

USC-SJTU Institute of Cultural and Creative Industry, Shanghai Jiao Tong University, Shanghai, China

ABSTRACT

Generative AI can enrich classical Chinese poetry reading through paraphrases, images and video, but these representations also shape how poems are presented, interpreted and understood by readers. We compared four reading packages in a within-participant study with 64 Chinese readers: poem only (P), poem plus AI paraphrase (PT), PT plus a four-panel image (PTI), and PT plus a silent video (PTV). Paraphrases improved perceived comprehension, vividness and engagement and reduced mental effort, while increasing perceived interpretive constraint by 0.75 points on a seven-point scale. Static imagery further increased vividness by 1.29 points and constraint by 1.22, alongside greater engagement and lower effort. Video increased vividness by 0.43 points and distraction by 1.40; the distraction pattern persisted in sensitivity and paired-response analyses. Objective comprehension accuracy was near ceiling, with no detectable differences in the focal comparisons. These findings reveal two costs of richer AI support: stronger perceived direction of interpretation and, for video, greater distraction. With static imagery, poems became more vivid while their meanings felt more predetermined. We therefore argue that literary AI interfaces should balance experiential benefits with interpretive freedom and attention, prioritizing optional support, explicit presentation of interpretive choices and reader-controlled playback.

Keywords: Chinese classical poetry, Generative artificial intelligence, Multimedia reading, Cognitive load, Interpretive constraint

INTRODUCTION

An AI paraphrase can make a classical Chinese poem more accessible while making one reading more prominent. Expanding sparse poetic language into prose or a scene selects relations, details and an emotional tone. These choices may help readers imagine the poem while making its meaning feel predetermined. Felt interpretive freedom therefore matters alongside ease of reading.

Cognitive Load Theory concerns the processing resources available for learning (Sweller, 1988). Coordinated verbal and pictorial representations can support learning, but their processing demands can exceed available capacity (Mayer and Moreno, 2003). Guidance also depends on prior knowledge

Received September 24, 2026; Revised November 2, 2026; Accepted November 12, 2026; Available online December 1, 2026

© 2026 The Authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

(Kalyuga et al., 2003). Animation comparisons are sensitive to informational equivalence, pacing and interactivity (Tversky et al., 2002); static illustrations have outperformed narrated animations in some instructional comparisons (Mayer et al., 2005). These findings caution against assuming that additional media improve poetry reading.

Poetry applications show benefits under particular learning conditions. An interactive multimodal application improved primary-school pupils' comprehension and intrinsic motivation relative to traditional instruction (Chen and Jamiat, 2023). LivePoem's AI-generated musical storyboards made learning more engaging and pleasant for non-native Chinese learners (Liang et al., 2025). How much interpretive room adult readers feel they retain remains less clear.

We studied 64 Chinese readers and eight Tang–Song seven-character quatrains. Each reader encountered four poems, one under each presentation package. We examined how adding an AI paraphrase, then a static image or silent video, changed experience, perceived interpretive constraint and distraction. Experience ratings, objective comprehension and exploratory AI-coded meaning-unit counts provided distinct outcomes. A study-specific constraint composite assessed readers' sense of being guided.

METHOD

PoetryTran generated modern Chinese paraphrases using DeepSeek-V4-Flash adapted with LoRA (Hu et al., 2022) on Alibaba Cloud Model Studio. The training corpus comprised 16,602 poem–paraphrase pairs, manually organized and edited from Tang poems in chinese-poetry (Chinese Poetry Contributors, n.d.). Figure 1 summarizes model adaptation and the study workflow; Appendix A provides the training settings and model comparison results.

Four experts from the cultural and creative fields developed six evaluation dimensions: semantic accuracy, information completeness, handling of context and allusions, preservation of imagery and emotion, readability, and avoidance of unsupported additions. The automated judge qwen3.7-max-2026-06-08 scored each dimension from 0 to 5. The panel reviewed all 1,660 validation outputs to assess whether the automated scores aligned with expert judgments under the rubric, and approved the evaluation. Appendix A reports the automated scores; expert review assessed their consistency with human judgment.

Static images were generated with ChatGPT Images 2.0 and videos with Seedance 2.0 Fast, both at 1920 × 1080 pixels. Each image contained four panels, one per line of the quatrain; each video showed a continuous, ten-second silent scene. Style varied across poems. Images allowed free inspection and videos allowed replay, but no pausing or scrubbing (Figure 2).

Human Study Design and Procedure

Eight Tang–Song seven-character quatrains formed two sets, each containing two Tang and two Song poems. Participants read one set, encountering each

package once: poem only (P), poem plus paraphrase (PT), PT plus static imagery (PTI), and PT plus silent video (PTV). Four questionnaire versions varied poem–presentation pairings and order; analyses followed actual administrations.

Adults recruited through campus course announcements and an open call participated online on 23 August 2026 using their own devices; display size and viewing conditions were unrecorded. Online responses were authoritative; printed copies supported manual checking. S01 and S03 each contained 32 participants, equally divided between general-reader and culturally engaged strata. Removing S02, a source-file duplicate of S01, left 64 distinct response profiles and 256 tasks (Table 1). S01 was a nested sensitivity subset, not an independent sample.

Participants wrote interpretations before answering comprehension questions and rating experience. Instructions prohibited searching or additional AI assistance; compliance was unverified. All 64 completed four tasks in approximately 10–15 minutes.

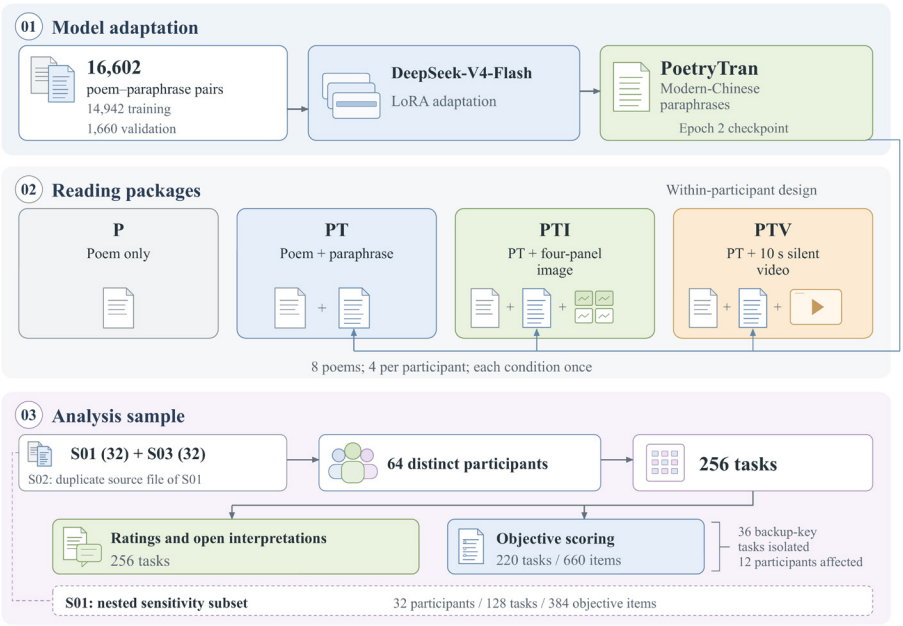


Figure 1: Study pipeline: model adaptation, the four presentations, and the analysis sample.

This study used an anonymous questionnaire. No personally identifiable or sensitive personal information was collected, and the study involved no procedures harmful to participants. Under Article 32 of China's Measures for the Ethical Review of Life Science and Medical Research Involving Humans (2023), the study was eligible for exemption from ethics review as research using anonymized information. All participants were informed of the study purpose, data anonymity and the voluntary nature of participation before taking part.

Table 1: Participant characteristics (N = 64).

Characteristic	Category	n (%)
Reader stratum	General / culturally engaged	32 / 32 (50 / 50)
Age (years)	18–30	19 (29.7)
	31–45	19 (29.7)
	46 and older	26 (40.6)
Gender	Male / female	32 / 32 (50 / 50)
Education	High school or below	13 (20.3)
	Bachelor	29 (45.3)
	Master or higher	22 (34.4)

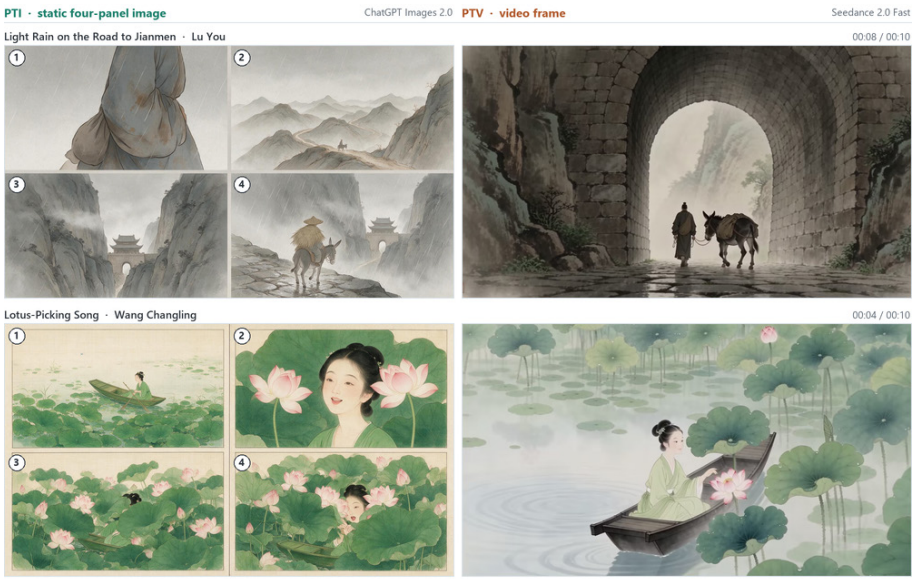


Figure 2: Example stimuli for two of the eight poems. Each PTI image had four panels, one per line of the quatrain (numbered 1–4); each PTV clip was a ten-second silent video of one continuous scene.

Measures and Coding

Objective comprehension comprised three four-option questions per task. Excluding 36 backup-key tasks from 12 participants left 220 tasks and 660 binary responses; their other eligible tasks, including Lotus-Picking Song, remained. Ratings and open-text analysis retained all 256 tasks.

Five seven-point agreement items (1 = strongly disagree; 7 = strongly agree) measured perceived comprehension, imagery vividness, emotional engagement, mental effort and distraction. The study-specific three-item interpretive-constraint composite averaged fixed meaning, reduced interpretive space and cue guidance; a two-item version omitted cue guidance. Internal consistency concerned these composites, not single-item ratings. Separately rated five-point poem familiarity was mean-centered.

Two AI-assisted coders independently counted distinct, textually supported meaning units beyond literal scene description in all 256 interpretations.

Participant, batch, order, reader-stratum and condition metadata were masked, although content could reveal modality. Exact duplicates received identical annotations within each coder. Only counts were linked to the analysis ledger by normalized exact text; earlier alternative-reading and cue-adoption labels were excluded. Counts were modeled separately without adjudication. Agreement used ICC(A,1), a 2,000-resample participant-cluster bootstrap interval, and exact agreement. This exploratory coding lacked human-expert validation.

Statistical Analysis

R models used lme4 (Bates et al., 2015) and emmeans for marginal means and contrasts. Item-level binomial correctness models included condition, question, centered familiarity, categorical task position, poem and batch as fixed effects and a participant random intercept; a random-poem specification checked sensitivity. Linear mixed rating models used the same task-level terms and participant intercepts; each coder's counts used a separate Poisson mixed model. Correctness and three-item constraint were primary for this analysis; two-item constraint and five experience ratings were secondary. Analyses were not preregistered.

Focal contrasts were PT–P, PTI–PT and PTV–PT; PTV–PTI was secondary. Two-sided tests used pointwise 95% Wald intervals. Holm correction covered three focal tests per outcome, six across primary outcomes and 15 across experience ratings. Raw paired rating differences used 5,000 participant-level bootstrap resamples. Models were repeated in S01. Exploratory condition-by-stratum and condition-by-familiarity interactions used separate maximum-likelihood likelihood-ratio comparisons retaining moderator main effects; main rating estimates used restricted maximum likelihood. Holm correction covered seven outcomes separately for each moderator and categorical task-position tests. Distraction checks used ordinal and binary mixed models and paired Wilcoxon testing. Fixed-standard-error sensitivity thresholds used $(z_{0.975} + z_{0.80})SE$ at unadjusted two-sided $\alpha = .05$, exponentiating log-odds thresholds: retrospective local precision approximations, not prospective power estimates.

RESULTS

Analysis Aample and Objective Accuracy

The 64 participants contributed 256 tasks, with 64 per condition for ratings and open interpretations and 55 standard-key tasks (165 items) per condition for accuracy (Table 2). The three-item constraint composite had raw $\alpha = .859$ and within-person $\alpha = .858$; corresponding two-item values were .901 and .894.

Objective accuracy was near ceiling, ranging from 91.5% to 95.2%. No focal contrast was detected (primary-family Holm $p \geq .744$; Table 3). The accuracy intervals were wide. Random-poem estimates were similar (ORs = 2.04, 0.75 and 0.92 for PT–P, PTI–PT and PTV–PT; within-outcome Holm $p \geq .749$). Participant-intercept variance reached zero in objective and count models, producing singular fits; no additional optimizer or Hessian failure was recorded.

Table 2: Raw outcomes by observed presentation condition.

Outcome	P	PT	PTI	PTV
Objective accuracy	91.5% (n=55)	95.2% (n=55)	93.9% (n=55)	94.5% (n=55)
Interpretive constraint	2.52 (0.44)	3.32 (0.44)	4.51 (0.55)	3.58 (0.68)
Perceived comprehension	4.77 (0.83)	5.31 (0.53)	5.47 (0.53)	5.16 (0.65)
Imagery vividness	3.98 (0.93)	4.48 (0.56)	5.67 (0.76)	4.80 (1.04)
Emotional engagement	4.05 (0.93)	4.44 (0.69)	4.84 (0.76)	4.47 (0.82)
Mental effort	3.16 (1.07)	2.52 (0.76)	2.17 (0.38)	2.72 (0.79)
Distraction	2.02 (0.12)	2.00 (0.00)	2.00 (0.00)	3.44 (1.47)

Note. Ratings are mean (SD), on 1–7 scales, from 64 tasks per condition. Accuracy uses 55 standard-key tasks (165 item responses) per condition.

Experience Gains and Perceived Interpretive Constraint

Adjusted contrasts showed that paraphrase support improved reported experience relative to the poem alone: perceived comprehension increased by 0.53 points, vividness by 0.45 and engagement by 0.39, while effort fell by 0.61 (all experience-family Holm $p \leq .002$). Adding static imagery increased vividness by 1.29 points and engagement by 0.45, and reduced effort by 0.42 (all $p < .001$). Its perceived-comprehension contrast did not survive correction ($p = .201$).

These experiential gains accompanied higher perceived interpretive constraint. The composite increased by 0.75 points from P to PT and by 1.22 from PT to PTI. Video produced a smaller increase over PT, 0.30 (all primary-family Holm $p < .001$). PTV was 0.92 points below PTI [−1.062, −0.782].

Table 3: Adjusted focal contrasts with 95% confidence intervals.

Outcome	PT–P	PTI–PT	PTV–PT
Objective accuracy	OR 2.12 [0.59, 7.60]	OR 0.74 [0.20, 2.79]	OR 0.92 [0.25, 3.31]
Interpretive constraint	0.75 [0.61, 0.90]***	1.22 [1.08, 1.37]***	0.30 [0.16, 0.45]***
Perceived comprehension	0.53 [0.37, 0.69]***	0.17 [0.01, 0.33]	–0.12 [–0.28, 0.05]
Imagery vividness	0.45 [0.22, 0.68]**	1.29 [1.06, 1.52]***	0.43 [0.20, 0.66]**
Emotional engagement	0.39 [0.20, 0.58]***	0.45 [0.26, 0.64]***	0.11 [–0.08, 0.30]
Mental effort	–0.61 [–0.81, –0.42]***	–0.42 [–0.61, –0.23]***	0.13 [–0.06, 0.33]
Distraction	0.11 [–0.13, 0.34]	–0.01 [–0.25, 0.22]	1.40 [1.17, 1.64]***

Note. Effects are covariate-adjusted odds ratios (accuracy) or scale-point differences (ratings), with pointwise 95% Wald CIs. CIs are not multiplicity-adjusted. Stars use Holm correction across six primary or 15 experience tests: * $p < .05$; ** $p < .01$; *** $p < .001$.

In the nested S01 subset (32 participants, 128 tasks, 384 objective items), PT/PTI experience patterns and PTV vividness/distraction increases persisted,

while accuracy contrasts remained inconclusive. The video-constraint estimate was sensitive to this restriction: PTV–PT was −0.010 [−0.169, 0.148] for the three-item composite and −0.063 [−0.161, 0.036] for the two-item version (within-outcome Holm p = .897 and .213).

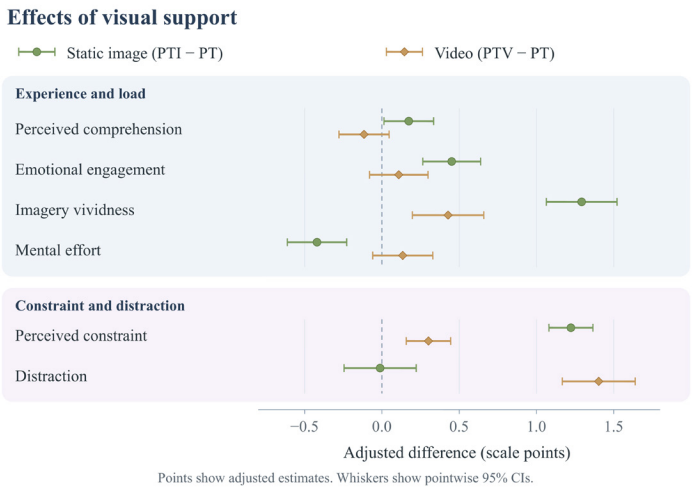


Figure 3: Adjusted contrasts relative to paraphrase alone, on 1–7 scales, with pointwise 95% confidence intervals. Family-adjusted tests are reported in Table 3.

Video Vividness and Distraction

Relative to PT, PTV increased vividness by 0.43 [0.197, 0.659] (experience-family Holm p = .002) and distraction by 1.403 [1.167, 1.639] (p < .001; Figure 3). Its perceived-comprehension, engagement and effort contrasts remained inconclusive after correction. Distraction exceeded the observed minimum of 2, on a scale starting at 1, for 39/64 participants under PTV, none under PT/PTI and one under P. PTV–PT differences were positive for 39 participants and zero for 25 (median = 1; paired Wilcoxon V = 780, two-sided $p = 3.77 \times 10^{-8}$). This unadjusted test corroborated the distraction pattern; ordinal and binary mixed-model checks were numerically unstable.

Open Interpretations and Exploratory Checks

Mean response length ranged from 88.1 Chinese characters under P to 103.3 under PTV. AI-coder count agreement was ICC(A,1) = .909, 95% CI [.874, .935], with 84.8% exact agreement. Neither coder detected focal differences in meaning-unit counts (all within-outcome Holm p = 1.000; all log-count intervals included zero).

Reader stratum moderated perceived comprehension, vividness, engagement and effort ($\chi^2(3)$ = 59.43, 32.56, 19.96 and 39.33; all seven-outcome Holm p < .001). No interaction was detected for constraint, distraction or accuracy (p ≥ .210), and no condition-by-familiarity interaction survived

correction ($p \geq .071$). These exploratory moderators concern recruitment strata and self-reported familiarity, not measured literary competence. Task position was associated with distraction ($\chi^2(3) = 23.56$, Holm $p = .0002$); no other position effect survived correction. Fixed-standard-error sensitivity thresholds were approximately 0.20 points for constraint, 0.23–0.34 for single-item ratings and ORs of 6.20–6.64 for accuracy, describing local precision rather than prospective power.

DISCUSSION

Support and the Sense of Interpretive Freedom

AI assistance improved reported experience while increasing the sense that interpretation was directed. Paraphrases improved perceived comprehension, vividness and engagement and reduced effort; static imagery further increased vividness and engagement and reduced effort, while producing the highest constraint. The constraint increases—0.75 points for PT–P and 1.22 for PTI–PT—show why experience ratings alone are insufficient. A poem can become easier to imagine while its meaning feels more predetermined.

Perceived constraint differs from the range of interpretations readers produce. Neither AI coder detected focal differences in meaning-unit counts, which can include several elaborations of one reading. Coder agreement supports annotation reproducibility, not validity as a measure of interpretive breadth. Human coding that distinguishes alternative readings from elaborations is needed to assess whether support narrows textually defensible interpretations.

Video and Attention

Video added vividness together with a clear distraction cost: 0.43 and 1.40 points relative to PT. Increased distraction persisted in S01 and paired responses; the video-constraint estimate was unstable. Static images covered all four lines and allowed free inspection, whereas videos showed one continuous scene without pause or scrubbing. Coverage and pacing could contribute to these differences. As animation research emphasizes (Tversky et al., 2002), comparisons should isolate informational content and playback control.

Design Implications

Figure 4 translates these findings into three design priorities. First, start with the poem, reveal paraphrases or visuals on request, and allow an easy return to the text. Second, label a visual as one possible reading and distinguish text-supported details from inferred ones. Third, provide pause, scrubbing and skip controls so readers can decide when video competes with the poem.

These proposed features require evaluation. Comparisons should match visual content while varying motion and playback control, assessing experience, perceived constraint and distraction separately, alongside more discriminating comprehension tasks and human assessment of alternative readings.

Scope and Limitations

The findings concern 64 readers, eight poems and the tested packages. Online sessions were unsupervised; devices and viewing conditions were unrecorded, and compliance with search restrictions was unverified. Three recognition items per task had limited resolution at near-ceiling accuracy, and objective scoring covered fewer tasks. The constraint composite requires construct validation. Analyses were not preregistered; moderation and robustness checks were exploratory. Recruitment strata and self-reported familiarity do not measure literary competence.

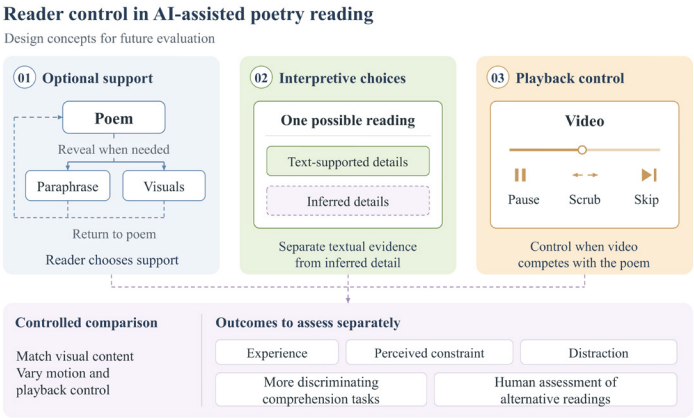


Figure 4: Proposed design concepts and evaluation plan for reader-controlled AI support.

CONCLUSION

AI paraphrases and static imagery improved reported reading experience while increasing perceived interpretive constraint; video added vividness and a robust distraction cost. These findings suggest that literary AI can shift interpretive authority toward the representations it supplies. By foregrounding particular relations, details and emotional tones, support can make a poem easier to imagine while making its meaning feel more predetermined. An appealing presentation is therefore an incomplete criterion for literary support.

Design should preserve readers' agency through optional, progressive support, visible interpretive choices, clear distinctions between textual evidence and model inference, and reader-controlled playback. Evaluation should assess experience, perceived constraint, distraction and alternative readings separately. Easier reading does not establish that interpretive openness has been preserved. The goal is to support understanding while sustaining the reader's role in making meaning.

DATA AND MATERIALS AVAILABILITY

Analysis scripts and stimulus materials are available from the corresponding author on reasonable request. Any sharing of participant-level responses

is subject to the scope of participant consent. Source poems derive from chinese-poetry; poem–paraphrase pairs were prepared for this study.

REFERENCES

Bates, D., Mächler, M., Bolker, B. and Walker, S. (2015). Fitting linear mixed-effects models using lme4. Journal of Statistical Software, 67(1), pp. 1–48. https://doi.org/10.18637/jss.v067.i01

Chen, C. and Jamiat, N. (2023). A quantitative study on the effects of an interactive multimodal application to promote students' learning motivation and comprehension in studying Tang poetry. Frontiers in Psychology, 14, article 1189864. https://doi.org/10.3389/fpsyg.2023.1189864

Chinese Poetry Contributors (n.d.). chinese-poetry: The most comprehensive database of Chinese poetry. GitHub. https://github.com/chinese-poetry/chinese-poetry

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L. and Chen, W. (2022). LoRA: Low-rank adaptation of large language models. International Conference on Learning Representations. https://openreview.net/forum?id=nZeVKeeFYf9

Kalyuga, S., Ayres, P., Chandler, P. and Sweller, J. (2003). The expertise reversal effect. Educational Psychologist, 38(1), pp. 23–31. https://doi.org/10.1207/S15326985EP3801_4

Liang, Q., Ma, X., Hopkins, T. and Wang, Y. (2025). LivePoem: Improving the learning experience of classical Chinese poetry with AI-generated musical storyboards. Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, pp. 10307–10315. https://doi.org/10.24963/ijcai.2025/1145

Mayer, R. E. and Moreno, R. (2003). Nine ways to reduce cognitive load in multimedia learning. Educational Psychologist, 38(1), pp. 43–52. https://doi.org/10.1207/S15326985EP3801_6

Mayer, R. E., Hegarty, M., Mayer, S. and Campbell, J. (2005). When static media promote active learning: Annotated illustrations versus narrated animations in multimedia instruction. Journal of Experimental Psychology: Applied, 11(4), pp. 256–265. https://doi.org/10.1037/1076-898X.11.4.256

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. Cognitive Science, 12(2), pp. 257–285. https://doi.org/10.1207/s15516709cog1202_4

Tversky, B., Morrison, J. B. and Betrancourt, M. (2002). Animation: Can it facilitate? International Journal of Human-Computer Studies, 57(4), pp. 247–262. https://doi.org/10.1006/ijhc.2002.1017

APPENDIX A: MODEL ADAPTATION AND EVALUATION

Pairs were randomly split into 14,942 training and 1,660 validation examples (seed 42). LoRA training used two epochs, batch size 16, learning rate 2×10^{-4} with linear decay and .05 warmup, maximum sequence length 4,096, weight decay .01, rank 8, $\alpha = 32$ and dropout .10. Epoch 2 had the lowest validation loss.

Adapted Flash, untuned Flash and untuned V4 Pro were compared on the same 1,660 validation prompts using the rubric in Method. Analyses used available-case dimension means and complete-prompt means; paired 95% CIs used 5,000 prompt-level bootstrap resamples. Complete-prompt means were 4.483, 4.336 and 4.368, respectively. Paired gains were 0.146 against untuned Flash (n = 1,558; 95% CI [0.117, 0.174]) and 0.117 against Pro

(n = 1,548; [0.088, 0.145]). Avoidance of unsupported additions showed the largest gain over untuned Flash (0.420 [0.358, 0.484]); completeness and imagery/emotion intervals included zero.

These numerical results are automated scores; expert review is described in Method. Evaluation reused the checkpoint-selection validation set, with no independent test set. The reader experiment compared presentation packages, not paraphrase sources.