Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Scaffolded large language model feedback builds durable translation competence whereas unrestricted use fosters dependency

Xixi Zhang

zhangxixi0628@163.com

School of Foreign Languages, Wuhan College of Arts \& Sciences

Guangyin Jin

School of Computer Science and Engineering, Nanjing University of Science and Technology

Article

Keywords:

Posted Date: November 16th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11512296/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Scaffolded large language model feedback builds durable translation competence whereas unrestricted use fosters dependency

Xixi Zhang[1,*] **and Guangyin Jin**[2,*]

[1]School of Foreign Languages, Wuhan College of Arts & Sciences, Wuhan 430345, China
[2]School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China
[*]Co-corresponding author: zhangxixi0628@163.com; jinguangyin@njust.edu.cn

ABSTRACT

Large language models (LLMs) are rapidly entering translator education, yet it remains unclear whether they build students' own translation competence or merely improve the products students submit. We combined a validation study on public benchmark data with a 16-week quasi-experiment to test a scaffolded human–LLM pedagogy, the Draft–Diagnose–Revise–Reflect (D2R) cycle, for English–Chinese translation. In Study 1, a rubric-guided error-annotation prompt increased the agreement between LLM feedback and expert Multidimensional Quality Metrics judgements on 2,500 WMT segments for all four models tested; with this prompt GPT-4o reached a segment-level Kendall's τ of 0.30, on a par with the strongest neural metric, and the highest error-detection accuracy. In Study 2, 179 English majors received conventional instruction, unrestricted LLM access or the guardrailed D2R cycle. Unrestricted access maximised LLM-assisted performance but yielded no advantage in unassisted translation over conventional instruction at a delayed test 12 weeks after the course (Hedges' $g = -0.08$) and produced the largest AI-dependency gap. Scaffolded use produced durable gains in unassisted translation ($g = 0.70$), a 13% reduction in error rate and higher metacognitive regulation. Process logs indicated that the benefit was transmitted mainly through lower verbatim adoption of LLM output. Pedagogical design, not LLM access per se, determines whether LLMs support or supplant translation learning.

Introduction

Large language models (LLMs) have changed what machine translation can do and, consequently, what translation students can delegate to it. Instruction-tuned models trained at scale[1,2] now produce translations that rival or surpass dedicated neural machine translation (NMT) systems for high-resource language pairs such as English–Chinese[3,4], and they can additionally explain terminology, propose alternative renderings and comment on style on demand. For translator education, this is simultaneously an opportunity and a threat. On the one hand, an always-available conversational partner that can diagnose errors and justify choices could provide the kind of individualised, timely feedback that large translation classes rarely receive[5,6]. On the other hand, a tool that can produce a fluent target text in seconds invites students to outsource precisely the cognitive work—source-text analysis, problem identification, strategic decision-making and self-revision—that constitutes translation competence[7–9].

Research on machine translation in language and translation education has long documented this tension. Studies of NMT in second-language writing and translation classes report gains in lexical accuracy and fluency of the submitted product, together with concerns about over-reliance, superficial post-editing and uncritical trust in machine output[10,11]. Translation pedagogues have consequently argued that technology should be integrated as an object of critical reflection rather than as a black-box shortcut[12–14], a position consistent with social-constructivist models of translator education[15]. Generative LLMs sharpen this debate because, unlike NMT engines, they also act as tutors: they can be prompted to withhold answers, to ask questions or to point to errors without correcting them. Whether these affordances translate into learning, however, depends on how they are orchestrated.

Evidence from other domains suggests that the answer hinges on pedagogical design. A recent meta-analysis of experimental studies reported a large positive effect of ChatGPT on learning performance[16], but most of the included studies measured performance with the tool available, over short periods, and rarely tested retention; moreover, because they compared ChatGPT with no-ChatGPT conditions, the technology was confounded with the instructional method in which it was embedded, so that the observed gains cannot be attributed to either[17]. A large field experiment in high-school mathematics showed that

unrestricted access to GPT-4 improved practice performance yet reduced subsequent unassisted exam performance, whereas a tutor with pedagogical guardrails largely eliminated this harm[18]. Conversely, a randomised trial reported in this journal found that an AI tutor built on research-based instructional principles outperformed in-class active learning[19]. Process-oriented studies point to the mechanism: generative AI can induce "metacognitive laziness"[20] and cognitive offloading[21,22], with reduced neural and behavioural engagement during AI-assisted writing[23]. Taken together, these findings imply a dissociation between *performance with AI* and *learning from AI*, the classic distinction between short-term performance and durable learning that motivates the literature on desirable difficulties[24].

Translation education has not yet been examined through this lens. Existing studies of LLMs in translation classrooms are mostly short, rely on self-reports or on the quality of AI-assisted products, and seldom include a delayed, unassisted measure of competence[6,25]. A recent experiment reported in this journal, for example, found that adding an LLM to collaborative translation groups changed the negotiation process and raised perceived efficiency but did not improve the quality of the translations produced[26]; whether LLM use changes what students can do once the model is removed was not examined. A second, largely ignored problem is the validity of the feedback itself. LLMs hallucinate[27], and a feedback-centred pedagogy is only as good as the accuracy of the error diagnoses it relies on. The machine translation community has developed rigorous tools for exactly this purpose: expert Multidimensional Quality Metrics (MQM) annotations[28,29] released through the WMT metrics shared tasks[30,31] provide public gold standards against which automatic metrics[32–36] and LLM-based evaluators[37–39] can be benchmarked. These resources have not been used to qualify LLM feedback engines before they are deployed with learners.

Here we address both gaps. Grounded in scaffolding theory[40,41], feedback research[5] and the self-explanation literature[42], we designed the Draft–Diagnose–Revise–Reflect (D2R) cycle (Fig. 1), in which a guardrailed LLM tutor diagnoses MQM-typed errors in a student's own draft but does not supply full translations, so that the generative and evaluative work stays with the learner. In Study 1 we used public WMT MQM data, complemented by teacher-annotated learner translations, to compare four open and proprietary LLMs under three prompting strategies and to select a feedback engine. In Study 2 we conducted a 16-week quasi-experiment with three intact classes of English majors who received conventional instruction (CI), unrestricted LLM access (UL) or the D2R cycle with the validated engine (scaffolded LLM, SL), and we assessed unassisted translation competence at post-test and after a 12-week delay spanning the summer break, LLM-assisted performance, error profiles, interaction logs and learner perceptions (Fig. 2). We asked (RQ1) how accurately LLMs can diagnose translation errors relative to expert MQM judgements; (RQ2) whether scaffolded and unrestricted LLM use differ in their effects on durable, unassisted translation competence; (RQ3) whether LLM use creates a gap between assisted and unassisted performance; and (RQ4) which learning processes explain the differences.

The study departs from previous work in three respects. First, both LLM conditions used the same model and interface and differed only in pedagogical orchestration, so that the comparison of SL with UL isolates the effect of pedagogy from that of the technology, the confound that limits media-comparison studies[17]. Second, we introduce a validate-then-deploy procedure in which the feedback engine is qualified against public expert MQM annotations before it reaches learners, turning a resource built for machine translation research into a quality-assurance step for AI-supported teaching. Third, we operationalise the AI-dependency gap—the difference between a student's assisted and unassisted performance on parallel tasks—as an outcome in its own right and relate it to logged interaction behaviour, which allows dependency to be measured rather than inferred from self-reports.

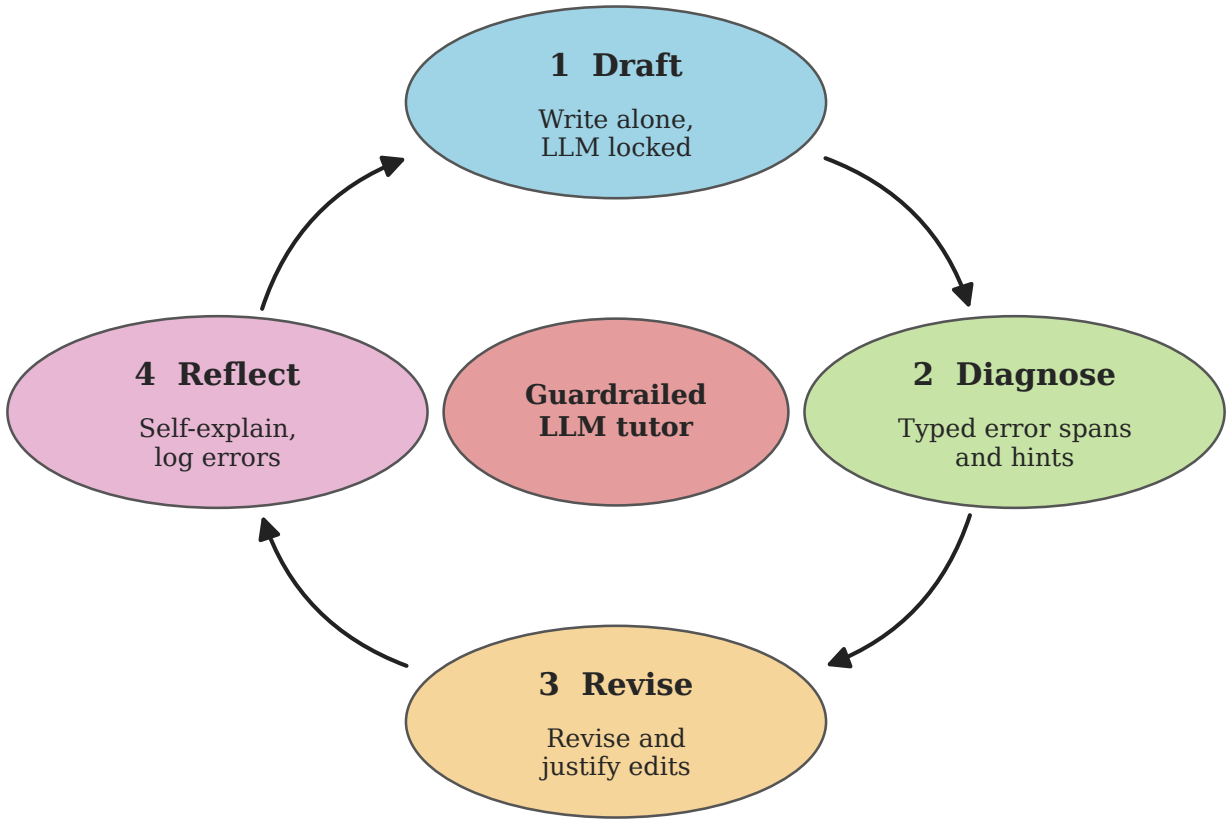


Figure 1. The Draft–Diagnose–Revise–Reflect (D2R) cycle. Students first draft independently while the LLM is locked; the guardrailed GPT-4o tutor, driven by the D2R-MQM prompt, then returns MQM-typed error spans and graded hints rather than corrected text; students revise and justify each edit and may request alternative renderings only for segments they have already attempted; finally, they write a self-explanation and update a personal error log, while teachers audit a sample of LLM feedback each week.

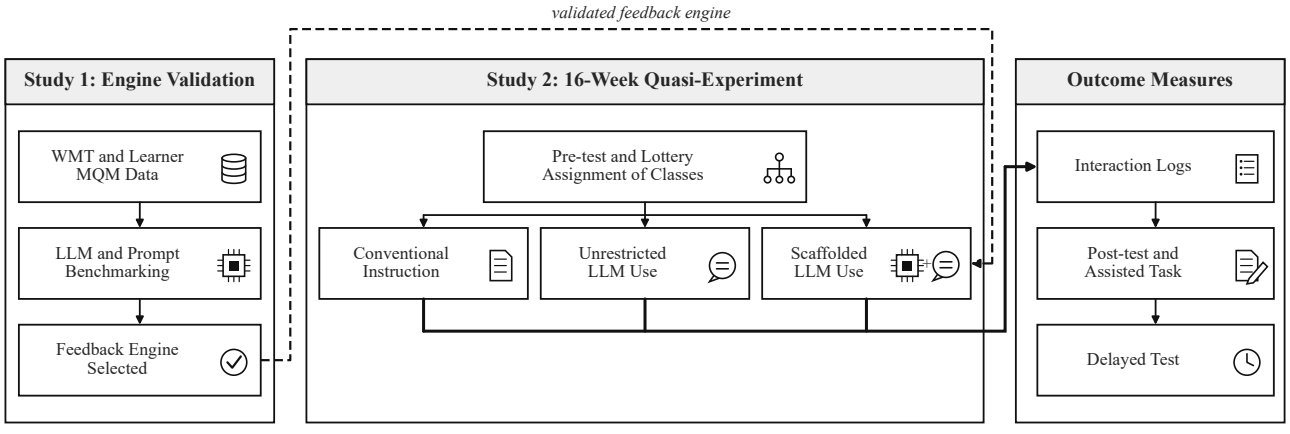


Figure 2. Research design and data-collection procedure. Study 1 benchmarked LLM feedback against expert MQM annotations on public WMT data and on learner translations to select the feedback engine for the scaffolded condition. Study 2 compared conventional instruction (CI), unrestricted LLM access (UL) and scaffolded LLM use (SL) over 16 weeks, with pre-, post- and 12-week delayed unassisted tests and an LLM-assisted task at week 16.

Methods

Study 1: validation of LLM feedback

Data

Public data comprised the Chinese-to-English test sets of the WMT22 and WMT23 general machine translation tasks together with the expert MQM annotations released through the WMT metrics shared tasks[30,31] and the MT Metrics Eval repository[43]. We randomly sampled 1,300 (WMT22) and 1,200 (WMT23) segment–system pairs, stratified by system, from the 15 systems per year with complete annotations. Because MQM annotations for English-to-Chinese learner translations are not publicly available, two experienced translation teachers (more than ten years of teaching each) annotated 600 segments sampled from pilot translations produced by 60 students of the 2024 cohort, who did not take part in Study 2, using the typology described below (inter-annotator agreement on segment scores: Kendall's $\tau = 0.61$; disagreements were resolved by discussion).

Error typology and scoring

We used a seven-category teaching adaptation of the MQM framework[28,29]: mistranslation, omission/addition, terminology, grammar, collocation/fluency, style/register and cohesion, each with minor, major or critical severity. The original WMT categories were mapped onto this typology (Accuracy/Mistranslation and Untranslated → mistranslation; Accuracy/Omission and Addition → omission/addition; Terminology → terminology; Fluency/Grammar, Spelling and Punctuation → grammar; Style/Awkward → collocation/fluency; Fluency/Register and Locale convention → style/register; Fluency/Inconsistency → cohesion), and non-translation errors were treated as critical. Segment scores were computed as negative weighted error counts (minor = 1, major = 5, critical = 10, with minor punctuation errors weighted 0.1 as in the WMT protocol), so that higher scores indicate better quality. Because WMT segments are annotated sentence by sentence, cohesion errors are rare in the public data, and category-level results for cohesion rest mainly on the learner set.

Scorers and prompts

We evaluated four instruction-tuned LLMs: Llama-3.1-70B-Instruct[44], Qwen2.5-72B-Instruct[45], DeepSeek-V3[46] and GPT-4o (version 2024-08-06)[47]. Each model was queried with temperature 0 using three prompts: (i) GEMBA-DA, which requests a 0–100 direct-assessment score[37]; (ii) GEMBA-MQM, which requests a list of error spans with category and severity[38]; and (iii) D2R-MQM, our prompt, which extends GEMBA-MQM with definitions of the seven categories, severity anchors illustrated by six bilingual examples and a short rationale step in which the model first summarises the meaning of the source before listing errors[48]. Error lists were converted to MQM scores with the weights above. Baselines were sentence-level BLEU[32], chrF[33], COMET-22[35] and xCOMET-XL[36], all computed with references.

Evaluation

Segment-level agreement was measured with Kendall's τ_b between automatic and human scores, with 95% CIs from 300 bootstrap resamples of segments. System-level agreement was measured with Pearson's r between system-mean scores across the 30 WMT systems. Category-level error detection was evaluated as binary presence of each category per segment, yielding precision, recall and F1 relative to the human annotations.

Study 2: quasi-experiment

Participants and design

Participants were third-year undergraduate English majors enrolled in a compulsory English–Chinese translation course at a public university in China during the spring semester of 2025 (February–June). All had taken the Test for English Majors Band 4 (TEM-4) in the preceding year, and its total score (0–100) served as the measure of general English proficiency. Three intact classes taught by the same instructor were assigned to the three conditions by lottery. Of 186 enrolled students, 179 completed all assessments and were analysed (CI 58 of 61, UL 61 of 63, SL 60 of 62; attrition did not differ between groups). An a priori power analysis (G*Power, ANCOVA, three groups, two covariates) indicated that 158 students were required to detect a medium group effect ($f = 0.25$) with power 0.80 at $\alpha = 0.05$; the analysed sample exceeded this number. The LLM-feedback study (Study 1) was completed in January 2025, before the semester began. The study was approved by the institutional ethics committee of the university and was conducted in accordance with the Declaration of Helsinki and relevant guidelines and regulations. All participants provided written informed consent; participation or non-participation did not affect course grades, and the CI and UL classes were offered the D2R platform after the study.

Instructional conditions

All groups followed the same syllabus (16 weeks, 2 × 90 min per week), source texts, weekly translation assignments (approximately 300-word texts alternating between English-to-Chinese and Chinese-to-English) and teacher lectures on translation strategies. The conditions differed only in how technology was used during assignments and in-class practice (Table 1). In CI, students completed assignments individually, could consult dictionaries, corpora and conventional NMT

systems, and received written teacher feedback on a rotating subsample of assignments together with whole-class debriefing. In UL, students used a GPT-4o chat interface on the course platform without restrictions and received the same teacher debriefing. In SL, students used the same interface configured as the D2R tutor with the D2R-MQM feedback engine selected in Study 1 and the guardrails listed in Table 1. The platform logged all interactions in both LLM groups.

Table 1. Instructional conditions. All conditions shared the syllabus, teacher, source texts and assessment schedule.

Feature	Conventional instruction (CI)	Unrestricted LLM (UL)	Scaffolded LLM (SL, D2R)
Drafting	Independent; dictionaries, corpora, NMT allowed	Free use of GPT-4o at any stage	Independent first draft; LLM locked until draft submitted
Feedback source	Teacher (rotating subsample) and peers	GPT-4o on demand; teacher debriefing	D2R-MQM error spans with category, severity and graded hints; teacher audit of 10% of feedback per week
Full translations from LLM	–	Allowed	Blocked; alternative renderings only for segments already drafted, after one revision attempt
Revision	Optional	Optional	Mandatory; each edit tagged with error category and short justification
Reflection	End-of-unit class discussion	End-of-unit class discussion	Weekly self-explanation and personal error log reviewed by the LLM and teacher
Low-confidence flags	–	–	Style/register and cohesion feedback marked “check with your teacher”

Translation tests

Unassisted competence was measured at week 1 (pre-test), week 16 (post-test) and in the first week of the following autumn semester (delayed test, 12 weeks after the post-test; week 28 of the study calendar). No translation instruction took place between the post-test and the delayed test. Each test comprised two passages totalling about 500 source words (497–512 across forms): one English-to-Chinese passage of consecutive sentences from a single FLORES-200 devtest article[49,50] and one Chinese-to-English passage from the WMT21 news test set, which does not overlap with the Study 1 data; passages were matched on length, readability and domain. Three parallel forms were counterbalanced across time points in a Latin-square rotation. Tests were taken on lab computers with internet access disabled and only a printed monolingual dictionary allowed (90 min). At week 16, students additionally completed a parallel assisted task (a fourth form, 503 words) in which all groups had unrestricted access to the same GPT-4o interface. Two trained raters blind to condition and time point annotated all translations with the MQM typology above; each rater’s penalty per 100 source words was converted to a 0–100 score (100 − penalty), and the mean of the two raters was analysed. The error profiles reported below are the averaged annotations of the two raters. Reference-based COMET-22 and xCOMET-XL scores were computed as robustness outcomes.

Process measures

Interaction logs were aggregated per student and week. Each request was classified as full-translation request, error diagnosis, explanation/terminology query, request for alternative renderings or reflection/self-explanation. Given the volume of the logs, classification was performed by a codebook-guided GPT-4o classifier; two human coders independently coded a stratified random sample of 1,500 requests (agreement between coders: Cohen’s $\kappa = 0.84$; agreement between the classifier and the coders’ consensus labels: $\kappa = 0.81$), and misclassifications in this sample were corrected. Verbatim adoption was the proportion of characters in each final translation that belonged to a longest common subsequence shared with any LLM output in the same session, averaged over assignments. Substantive self-revisions were edits between draft and final version that altered meaning, terminology or structure, normalised per 100 source words.

Questionnaires

At weeks 1 and 16, students completed a 10-item translation self-efficacy scale constructed following Bandura’s guidelines[51] (Cronbach’s $\alpha = 0.87$ and 0.90 at weeks 1 and 16) and the 8-item regulation subscale adapted from the Metacognitive Awareness Inventory[52] ($\alpha = 0.84$ and 0.87), both on 5-point scales; scale scores are item means. At week 16, students rated cognitive load with the 10-item instrument of Leppink et al.[53] (0–10; $\alpha = 0.80$, 0.83 and 0.88 for intrinsic, extraneous and germane load), and students in the LLM groups rated eight perception items on 5-point Likert scales.

Statistical analysis

Unassisted scores were analysed with a linear mixed-effects model including Group, Time, their interaction and standardised TEM-4 score as fixed effects and random intercepts for students, fitted by restricted maximum likelihood; the omnibus interaction was tested with a likelihood-ratio test on maximum-likelihood fits. Post-test, delayed and assisted scores were additionally analysed with ANCOVAs adjusting for pretest score and TEM-4, with pairwise contrasts corrected using the Holm procedure[54]. Effect sizes are reported as partial η^2 and as Hedges' g[55] for gain scores, with 95% percentile bootstrap CIs (1,500 resamples). Error rates were log-transformed and analysed with a Group × Time mixed model. Moderation by prior proficiency was tested by adding Group × Pretest interactions to a regression of retained gain, with Johnson–Neyman intervals[56]. Mediation was tested in the two LLM groups with a parallel multiple-mediator model and 2,000 bootstrap resamples[57]. Perception items were compared with Mann–Whitney U tests. Inter-rater reliability was assessed with a two-way random-effects, absolute-agreement, single-rater ICC[58]. All tests were two-sided with $\alpha = 0.05$. Analyses were conducted in Python 3 with statsmodels[59] and SciPy[60].

Results

Study 1: LLMs can diagnose translation errors with expert-like validity when guided by an MQM rubric

Table 2 summarises the agreement of the 16 scorers (four LLMs × three prompts and four reference-based metrics) with expert MQM judgements on the 2,500 WMT segments and on the 600 learner translations described in Methods.

On the WMT data, the D2R-MQM prompt yielded the highest segment-level agreement with human MQM scores for every model (Fig. 3a), with gains over GEMBA-MQM of $\Delta\tau = 0.044$ to 0.065 and over GEMBA-DA of 0.048 to 0.058. GEMBA-MQM, in contrast, did not consistently outperform the simpler direct-assessment prompt (it was better for Qwen2.5-72B and worse for the other three models), indicating that asking for error spans helps only when categories and severities are clearly anchored. GPT-4o with D2R-MQM achieved the highest agreement among the LLM configurations ($\tau = 0.298$, 95% bootstrap confidence interval (CI) [0.269, 0.329]), on a par with xCOMET-XL ($\tau = 0.295$ [0.268, 0.322]) and above COMET-22 ($\tau = 0.269$), chrF ($\tau = 0.129$) and BLEU ($\tau = 0.119$). System-level correlations over the 30 WMT systems were high for the neural metrics and the D2R-MQM configurations ($r = 0.79$–0.88; GPT-4o 0.82, xCOMET-XL 0.88) and lower for BLEU and chrF ($r = 0.64$ and 0.53; Table 2), in line with the known weakness of lexical overlap metrics for strong systems[30].

Because feedback must point students to specific problems rather than merely grade them, we also evaluated category-level error detection. Error detection was far from perfect for all models: GPT-4o with D2R-MQM detected annotated error categories with a precision of 0.41 and a recall of 0.43 (F1 = 0.42) on WMT data and an F1 of 0.52 on learner translations, the best values among the LLMs (Table 2). Detection was most reliable for mistranslation (F1 = 0.57), collocation/fluency (0.45) and grammar (0.44) and weakest for style/register (0.34) and cohesion (0.30), which require discourse-level judgement (Fig. 3b). Segment-level agreement was higher on learner translations than on professional MT output for nearly all scorers, presumably because learner translations contain more and more conspicuous errors; on this set Qwen2.5-72B with D2R-MQM reached the highest value ($\tau = 0.408$ [0.355, 0.462]), with GPT-4o close behind (0.376 [0.329, 0.426]). We selected GPT-4o with the D2R-MQM prompt as the feedback engine for Study 2 because it combined the highest agreement on the public WMT data with the highest error-detection F1 on both sets, and we flagged the two weakest categories in the tutor's output as "check with your teacher".

Study 2: participants and measurement quality

A total of 179 third-year English majors in three intact classes completed all assessments (CI, $n = 58$; UL, $n = 61$; SL, $n = 60$). The groups did not differ in sex distribution, age, TEM-4 score, prior generative-AI use, pretest translation score, metacognitive regulation or translation self-efficacy (all $p > 0.44$; Table 3). Inter-rater reliability of the MQM-based translation scores was excellent (ICC(2,1) = 0.97)[58], and human scores correlated strongly with reference-based COMET-22 ($r = 0.82$) and xCOMET-XL ($r = 0.75$) scores computed on the same texts, supporting the validity of the outcome measure.

Scaffolded, but not unrestricted, LLM use improved durable unassisted translation competence

Unassisted translation scores evolved differently across groups (Fig. 4). A linear mixed-effects model with random intercepts for students and TEM-4 as a covariate revealed a significant Group × Time interaction (likelihood-ratio $\chi^2(4) = 25.29$, $p < 0.001$; Table 4). Relative to CI, the SL group improved by an additional 5.03 points from pre- to post-test ($b = 5.03$, SE = 1.45, $p < 0.001$) and by 5.34 points from pre-test to the delayed test ($b = 5.34$, SE = 1.45, $p < 0.001$). The corresponding interactions for UL were small and non-significant (post-test: $b = -0.17$, $p = 0.908$; delayed: $b = -0.75$, $p = 0.602$).

Analyses of covariance (ANCOVAs) controlling for pretest score and TEM-4 confirmed these findings. At post-test, the group effect was significant ($F(2, 174) = 9.38$, $p < 0.001$, $\eta_p^2 = 0.10$); the adjusted means were 69.7 (CI), 69.9 (UL) and 74.4 (SL), with SL exceeding both CI (4.73 points, Holm-adjusted $p < 0.001$) and UL (4.50 points, $p < 0.001$), whereas UL and CI did not differ ($p = 0.855$). The advantage of scaffolding persisted at the delayed test 12 weeks later ($F(2, 174) = 10.03$, $p <$

Table 2. Agreement between automatic scorers and expert MQM annotations (Study 1). Segment-level Kendall's τ on 2,500 WMT22/23 zh–en segments and 600 learner en–zh segments; system-level Pearson r on 30 WMT systems; micro-averaged category-level error detection on WMT data (F1 on learner data after the slash). The configuration selected for Study 2 is shown in bold.

		Segment-level τ		System r	Error detection (WMT / learner F1)		
Scorer	Prompt	WMT zh–en	Learner en–zh	WMT	Precision	Recall	F1
BLEU	–	0.119	0.097	0.639	–	–	–
chrF	–	0.129	0.156	0.529	–	–	–
COMET-22	–	0.269	0.375	0.837	–	–	–
xCOMET-XL	–	0.295	0.364	0.879	–	–	–
Llama-3.1-70B	GEMBA-DA	0.184	0.239	0.669	–	–	–
Llama-3.1-70B	GEMBA-MQM	0.167	0.312	0.717	0.28	0.39	0.33 / 0.47
Llama-3.1-70B	D2R-MQM	0.232	0.344	0.794	0.33	0.39	0.36 / 0.51
Qwen2.5-72B	GEMBA-DA	0.201	0.281	0.680	–	–	–
Qwen2.5-72B	GEMBA-MQM	0.215	0.303	0.729	0.30	0.41	0.35 / 0.45
Qwen2.5-72B	D2R-MQM	0.259	0.408	0.848	0.36	0.39	0.38 / 0.50
DeepSeek-V3	GEMBA-DA	0.241	0.285	0.827	–	–	–
DeepSeek-V3	GEMBA-MQM	0.231	0.369	0.775	0.33	0.41	0.36 / 0.49
DeepSeek-V3	D2R-MQM	0.290	0.361	0.882	0.39	0.41	0.40 / 0.52
GPT-4o	GEMBA-DA	0.248	0.319	0.739	–	–	–
GPT-4o	GEMBA-MQM	0.234	0.338	0.701	0.33	0.40	0.36 / 0.49
GPT-4o	D2R-MQM	**0.298**	**0.376**	0.818	0.41	0.43	0.42 / 0.52

Table 3. Baseline characteristics of the three instructional groups. Values are means (standard deviations) unless stated otherwise. CI, conventional instruction; UL, unrestricted LLM; SL, scaffolded LLM (D2R).

Characteristic	CI (n = 58)	UL (n = 61)	SL (n = 60)	Test
Female, %	87.9	82.0	85.0	$\chi^2(2) = 0.83$, $p = 0.662$
Age (years)	21.17 (0.65)	21.10 (0.72)	21.25 (0.57)	$F(2, 176) = 0.82$, $p = 0.443$
TEM-4 score	69.45 (6.18)	68.31 (6.45)	68.82 (6.54)	$F(2, 176) = 0.47$, $p = 0.625$
Prior GenAI use (1–5)	3.02 (1.13)	2.89 (1.05)	2.93 (1.12)	$F(2, 176) = 0.22$, $p = 0.804$
Pretest translation score	66.81 (9.55)	66.22 (9.89)	65.55 (9.44)	$F(2, 176) = 0.25$, $p = 0.776$
Metacognitive regulation (1–5)	3.35 (0.58)	3.25 (0.64)	3.34 (0.67)	$F(2, 176) = 0.49$, $p = 0.611$
Translation self-efficacy (1–5)	3.07 (0.56)	3.03 (0.69)	3.05 (0.62)	$F(2, 176) = 0.08$, $p = 0.926$

Table 4. Linear mixed-effects model of unassisted translation scores. Reference categories are the conventional-instruction group and the pre-test. Model fitted by restricted maximum likelihood with random intercepts for students; the likelihood-ratio test compares models with and without the Group × Time interaction (maximum likelihood).

Fixed effect	b	SE	95% CI	p
Intercept	66.26	1.09	[64.12, 68.39]	< 0.001
Group: UL	0.46	1.52	[-2.52, 3.45]	0.762
Group: SL	-0.68	1.53	[-3.67, 2.31]	0.656
Time: post-test	3.49	1.03	[1.47, 5.52]	< 0.001
Time: delayed	1.58	1.03	[-0.45, 3.60]	0.127
TEM-4 (z)	5.86	0.52	[4.84, 6.88]	< 0.001
UL × post-test	-0.17	1.44	[-3.00, 2.66]	0.908
SL × post-test	5.03	1.45	[2.19, 7.88]	< 0.001
UL × delayed	-0.75	1.44	[-3.58, 2.08]	0.602
SL × delayed	5.34	1.45	[2.50, 8.18]	< 0.001
Random intercept variance = 37.66; residual variance = 31.00; ICC = 0.55				
Group × Time LRT: $\chi^2(4) = 25.29$, $p < 0.001$; $N_{\text{obs}} = 537$, $N_{\text{students}} = 179$				

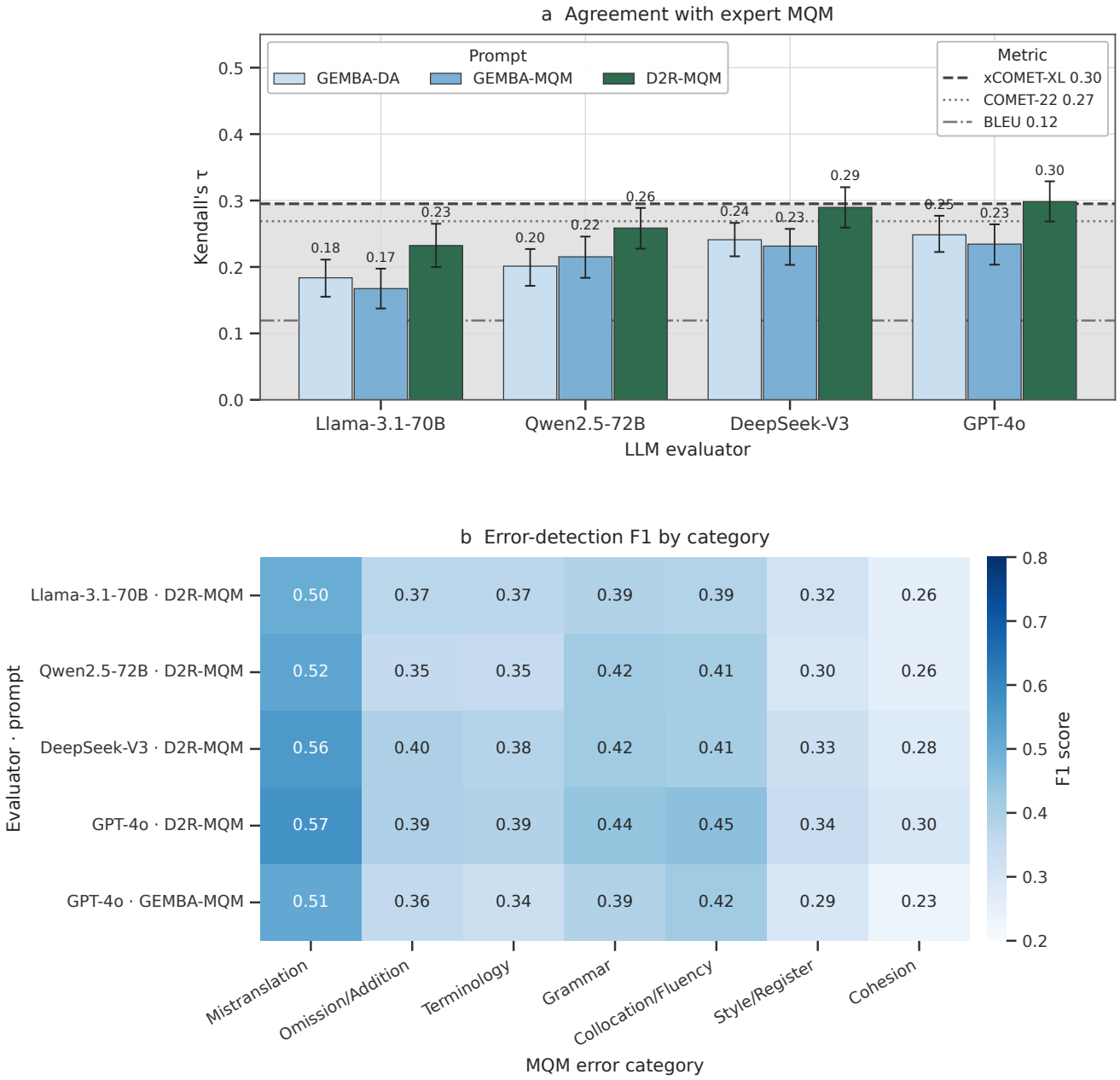


Figure 3. Validity of LLM-generated translation feedback against expert MQM annotations. (**a**) Segment-level Kendall's τ between automatic scores and human MQM scores on WMT22/23 zh–en data for four LLMs under three prompting strategies; error bars are 95% bootstrap CIs (300 resamples). Horizontal lines mark reference-based metrics; the shaded area lies below xCOMET-XL. (**b**) Category-level F1 of error detection for the D2R-MQM prompt across models and for GPT-4o with GEMBA-MQM.

0.001, $\eta_p^2 = 0.10$; adjusted means 67.7, 67.3 and 72.9; SL vs CI 5.15 points, SL vs UL 5.52 points, both $p < 0.001$). Over the summer, scores declined in all groups, but the decline was smallest and not significant in SL (delayed minus post-test: −1.61 points, $t(59) = -1.61$, $p = 0.113$), whereas it was significant in CI (−1.92 points, $p = 0.020$) and UL (−2.50 points, $p = 0.005$). In standardised terms, the retained gain of SL students exceeded that of CI by $g = 0.70$ (95% CI [0.35, 1.08]) and that of UL by $g = 0.69$ [0.35, 1.05], whereas UL did not differ from CI ($g = -0.08$ [−0.45, 0.28]; Figs. 5 and 6b). The same pattern held when xCOMET-XL replaced human scores as the outcome (SL vs CI: $b = 0.030$, $p = 0.007$; UL vs CI: $b = -0.016$, $p = 0.146$).

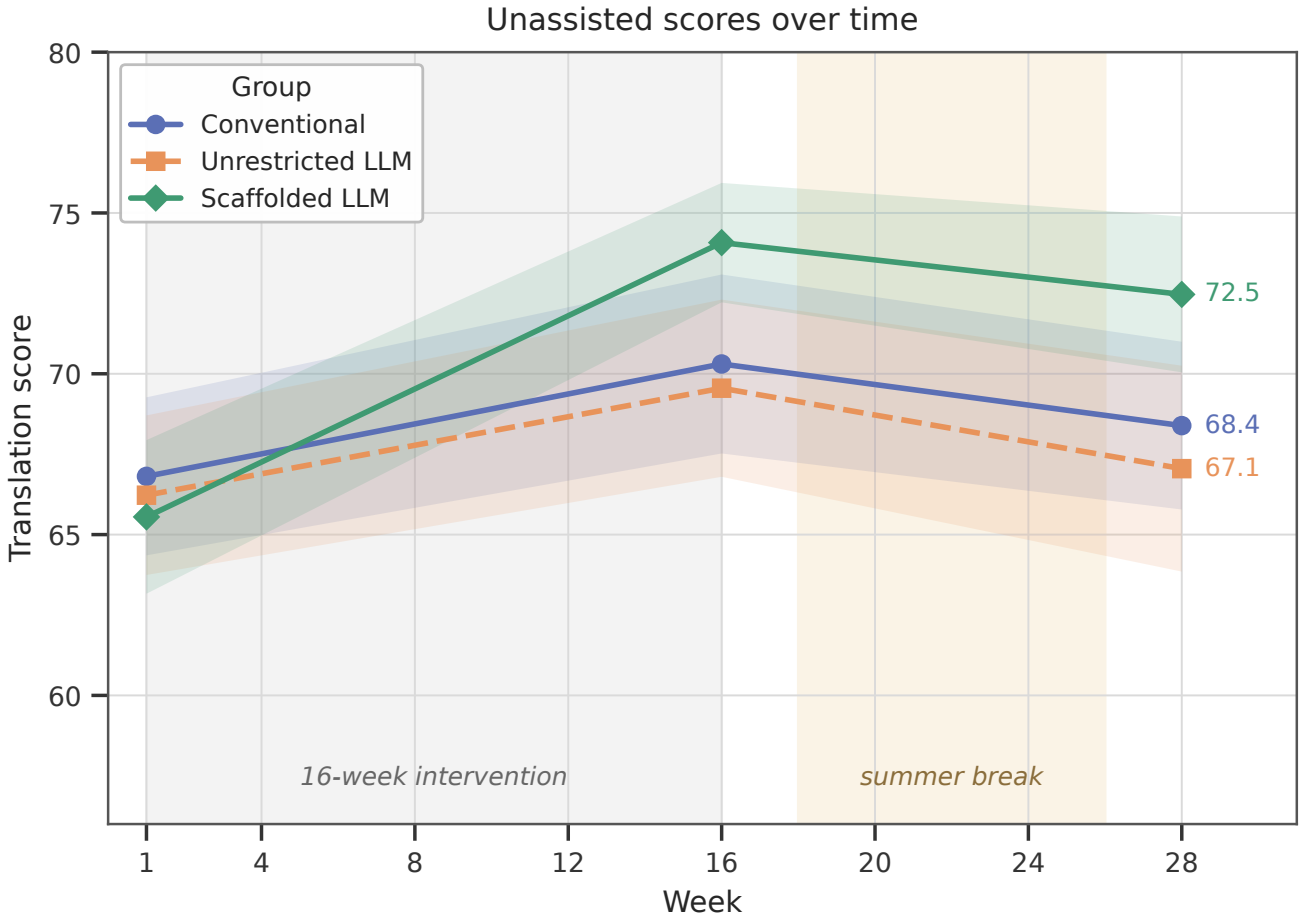


Figure 4. Unassisted translation scores over time. Group means at pre-test (week 1), post-test (week 16) and the delayed test 12 weeks later (week 28, start of the autumn semester); shaded bands are 95% CIs. The grey area marks the intervention period and the beige area the summer break. CI, conventional instruction; UL, unrestricted LLM; SL, scaffolded LLM.

Unrestricted use dissociated assisted performance from unassisted competence

At week 16, all students also completed a parallel translation task with full access to the same LLM interface. Here the ranking changed: UL students produced the best LLM-assisted translations (adjusted mean 83.9), followed by SL (82.2) and CI (77.0; $F(2, 174) = 29.27$, $p < 0.001$, $\eta_p^2 = 0.25$), with the difference between UL and SL not reaching significance after Holm correction ($p = 0.073$). Thus, unrestricted practice made students efficient users of the tool without making them better translators. We quantified this dissociation as the AI-dependency gap, i.e., the difference between each student's assisted and unassisted post-test scores (Fig. 6a). The gap was about twice as large in UL (14.1 ± 7.8 points) as in CI (7.2 ± 8.6) and SL (7.9 ± 6.0; $F(2, 176) = 15.32$, $p < 0.001$); the UL–CI difference was large ($g = 0.84$ [0.45, 1.26]), whereas SL did not differ from CI ($g = 0.09$ [−0.27, 0.49]). Scaffolded students therefore achieved assisted performance close to that of UL students while keeping the competence that remains when the tool is removed.

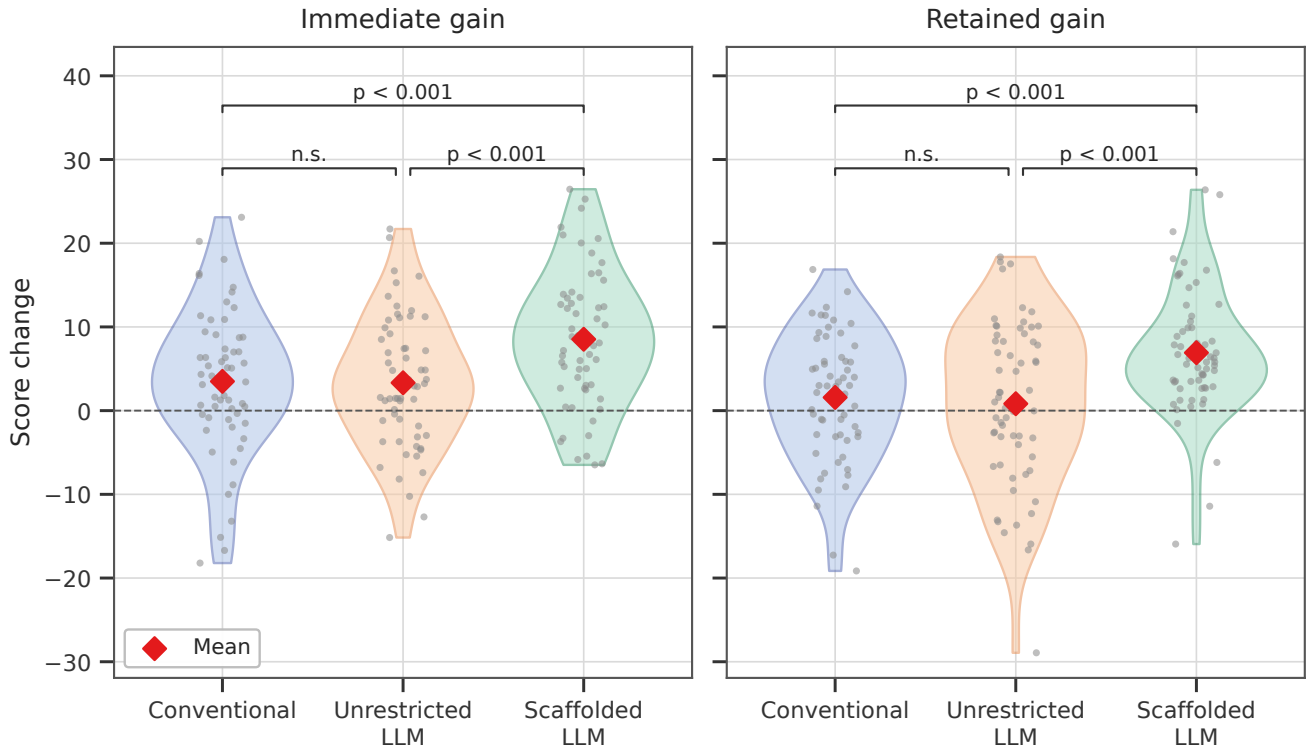


Figure 5. Distribution of individual learning gains. Immediate gain is post-test minus pre-test; retained gain is delayed test minus pre-test. Violins show kernel densities, grey dots are students and red diamonds group means. Brackets show Holm-adjusted ANCOVA contrasts controlling for pretest score and TEM-4; n.s., $p > 0.05$.

Scaffolding reduced meaning-level and discourse-level errors

Error profiles derived from the MQM annotations of the unassisted tests showed where the gains came from (Fig. 7). From pre-test to the delayed test, the total error rate decreased by 13.4% in SL (11.28 to 9.77 errors per 100 source words), compared with 3.5% in CI (10.86 to 10.48), whereas it rose slightly in UL (10.99 to 11.13, +1.3%); a mixed model of log error rates confirmed a significant SL × Time interaction ($b = -0.12$, $p = 0.018$), equivalent to an additional 12% reduction relative to CI, but no UL × Time interaction ($b = 0.03$, $p = 0.520$). In SL, the largest reductions concerned cohesion (−29%), collocation/fluency (−17%), mistranslation (−14%) and grammar (−11%). UL students improved in grammar (−9%) and terminology (−9%), the categories most easily "looked up", whereas omission/addition (+13%), style/register (+9%) and cohesion errors (+6%) increased, and their post-test improvement in total errors (10.99 to 10.54) had disappeared by the delayed test. The share of major and critical errors also decreased most in SL (from 47.4% to 43.1%, vs 47.4% to 47.5% in CI and 47.8% to 47.0% in UL; Fig. 7b).

How students used the LLM: interaction logs

The platform logged 24,245 LLM requests from the two LLM groups during weeks 2–15, distributed over 1,640 active student-weeks (3.2% of the 1,694 possible student-weeks showed no platform use), with visible dips in the onboarding week and the mid-term examination week (week 8). UL students most often asked the model to translate the source text outright (54.0% of requests), and this share rose over the semester (0.54 percentage points per week, $p = 0.001$; Fig. 8a). In SL, where full translations were blocked until a draft had been submitted, only 3.7% of requests sought full renderings (of segments already attempted), and 75.4% were metalinguistic—error diagnosis, explanations and reflection prompts—compared with 30.7% in UL (Fig. 8b). SL students also spent more time on task (115 vs 81 min per week), made about twice as many substantive self-revisions (8.26 ± 2.15 vs 4.34 ± 1.90 per 100 words) and adopted far less LLM text verbatim in their final versions (21.6% ± 8.0% vs 57.5% ± 15.9%; Fig. 8c). Across both LLM groups, more self-revision was associated with larger retained gains ($r = 0.31$, $p = 0.001$) and more verbatim adoption with smaller retained gains ($r = -0.37$, $p < 0.001$; Fig. 8d,e) and with a larger AI-dependency gap ($r = 0.47$, $p < 0.001$).

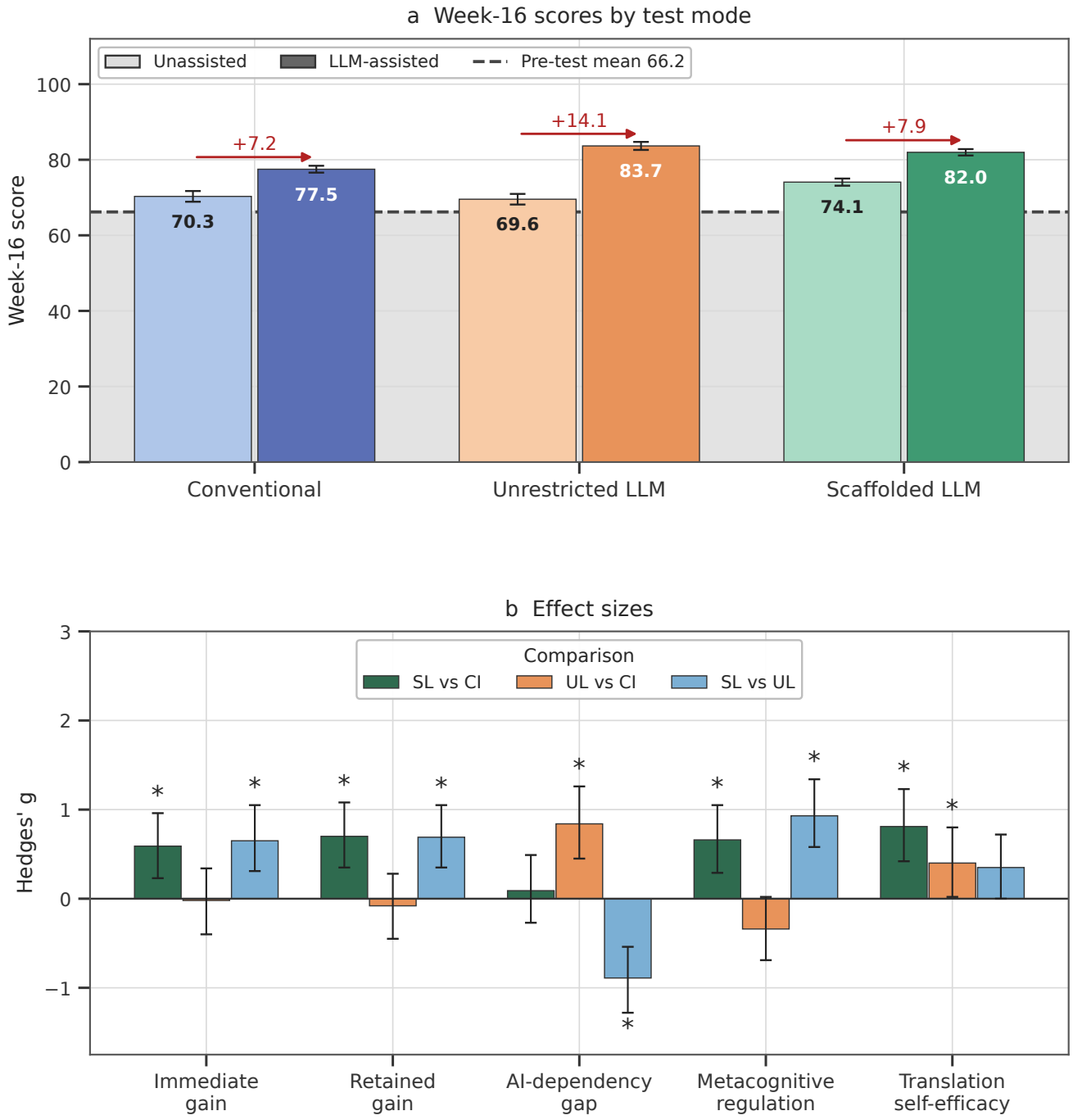


Figure 6. AI dependency and standardised effects. (**a**) Week-16 scores on the unassisted test (light bars) and on a parallel LLM-assisted task (dark bars), mean ± SE; red arrows give the AI-dependency gap and the shaded area lies below the pooled pre-test mean. (**b**) Between-group differences (Hedges' g) with 95% bootstrap CIs (1,500 resamples) for learning, dependency and self-regulation outcomes; *CI excludes zero.

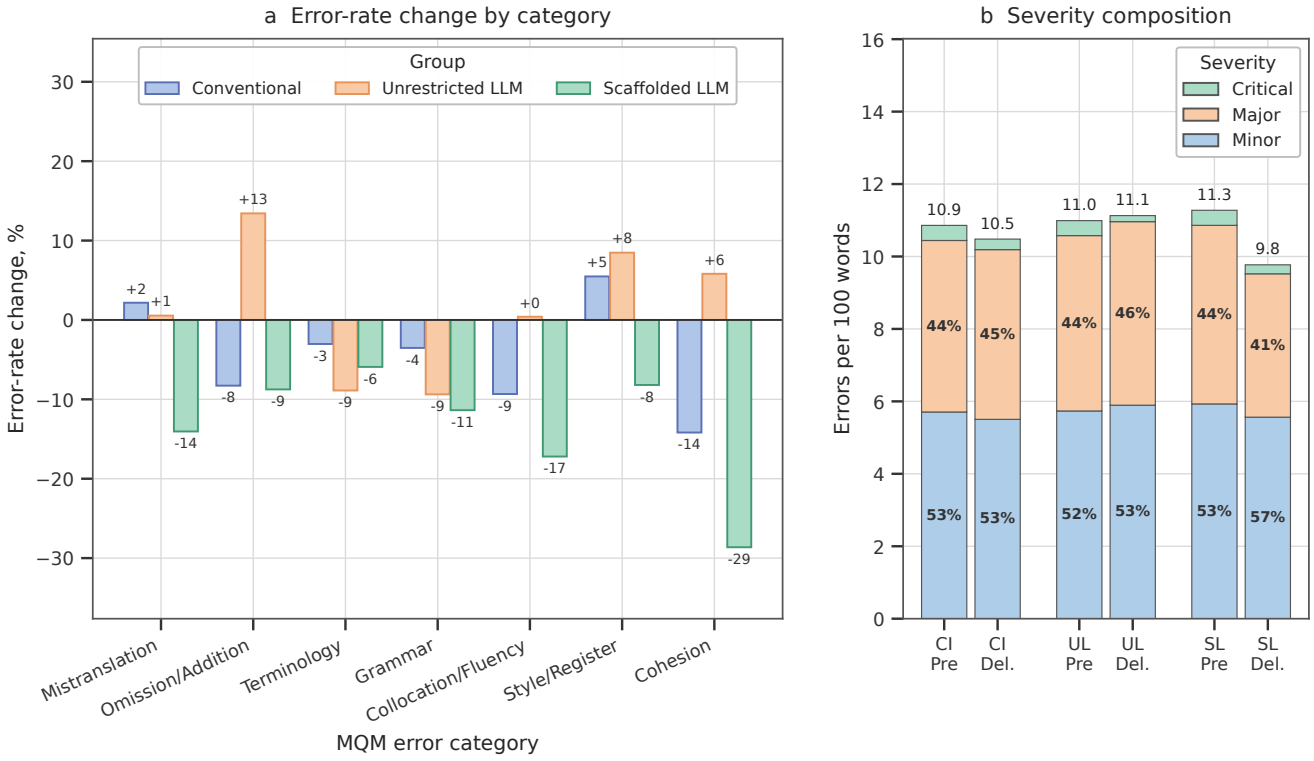


Figure 7. Error profiles in unassisted translations. (**a**) Relative change in error rate per MQM category from pre-test to the 12-week delayed test. (**b**) Errors per 100 source words at pre-test and delayed test (Del.), decomposed by severity; segment labels give the share of each severity level and numbers above bars the total rate.

Who benefits and why: moderation and mediation

We next asked whether the benefits of scaffolding depended on prior proficiency (Fig. 9a). Neither Group × Pretest interaction on retained gain was significant (SL: $b = -0.01$, $p = 0.965$; UL: $b = 0.08$, $p = 0.605$). Johnson–Neyman analysis showed that the advantage of SL over CI was significant for pretest scores between 51.2 and 80.5, which covers the large majority of students (observed range 40.7–86.2); at the extremes of the distribution, where data were sparse, the confidence band included zero. The effect of UL was not significant at any level. The benefit of scaffolding was therefore not confined to a particular proficiency band.

To identify mechanisms, we estimated a parallel mediation model among the LLM groups with condition (SL vs UL) as predictor, change in metacognitive regulation and verbatim adoption as mediators and retained gain as outcome, controlling for pretest score[56,57] (Table 5 and Fig. 9b). Scaffolding increased metacognitive regulation ($a_1 = 0.41$ [0.26, 0.57]) and reduced verbatim adoption by about 36 percentage points ($a_2 = -3.60$ [−4.02, −3.17] in units of 10 percentage points). Verbatim adoption in turn predicted smaller retained gains ($b_2 = -1.42$ [−2.69, −0.23]), and the indirect effect through verbatim adoption was significant (5.12 [0.80, 9.79]). The path from metacognitive change to retained gain was positive but imprecise ($b_1 = 2.12$ [−2.01, 6.01]), so the indirect effect through metacognition did not reach significance (0.86 [−0.84, 2.50]). With both mediators in the model, the direct effect of condition was close to zero ($c' = -0.04$ [−5.75, 5.30]; total effect $c = 5.94$). Because condition and verbatim adoption were strongly related, the individual path estimates are imprecise, as the wide intervals show, but the pattern indicates that the advantage of scaffolding arises largely because it keeps the generative work with the student.

Learner experience, cognitive load and self-beliefs

Students' perceptions mirrored the behavioural data (Fig. 10a). Both LLM groups rated the feedback as similarly accurate (SL 3.88 vs UL 3.92, $p = 0.697$) and were equally willing to use their approach again (4.08 vs 4.00, $p = 0.583$). UL students found the LLM's help easier to act on (4.11 vs 3.60, $p = 0.004$), but SL students more often reported understanding why their translations were wrong (4.02 vs 2.95, $p < 0.001$), believing that the LLM had improved their own ability (3.93 vs 3.54, $p = 0.014$) and being able to translate well without it (3.63 vs 2.77, $p < 0.001$). UL students more often reported relying on the LLM more than they wanted (3.74 vs 2.52) and trusting its output without checking (3.46 vs 2.30; both $p < 0.001$).

Cognitive-load ratings[53] showed that scaffolding did not simply make the task harder (Fig. 10b). Intrinsic load was lowest

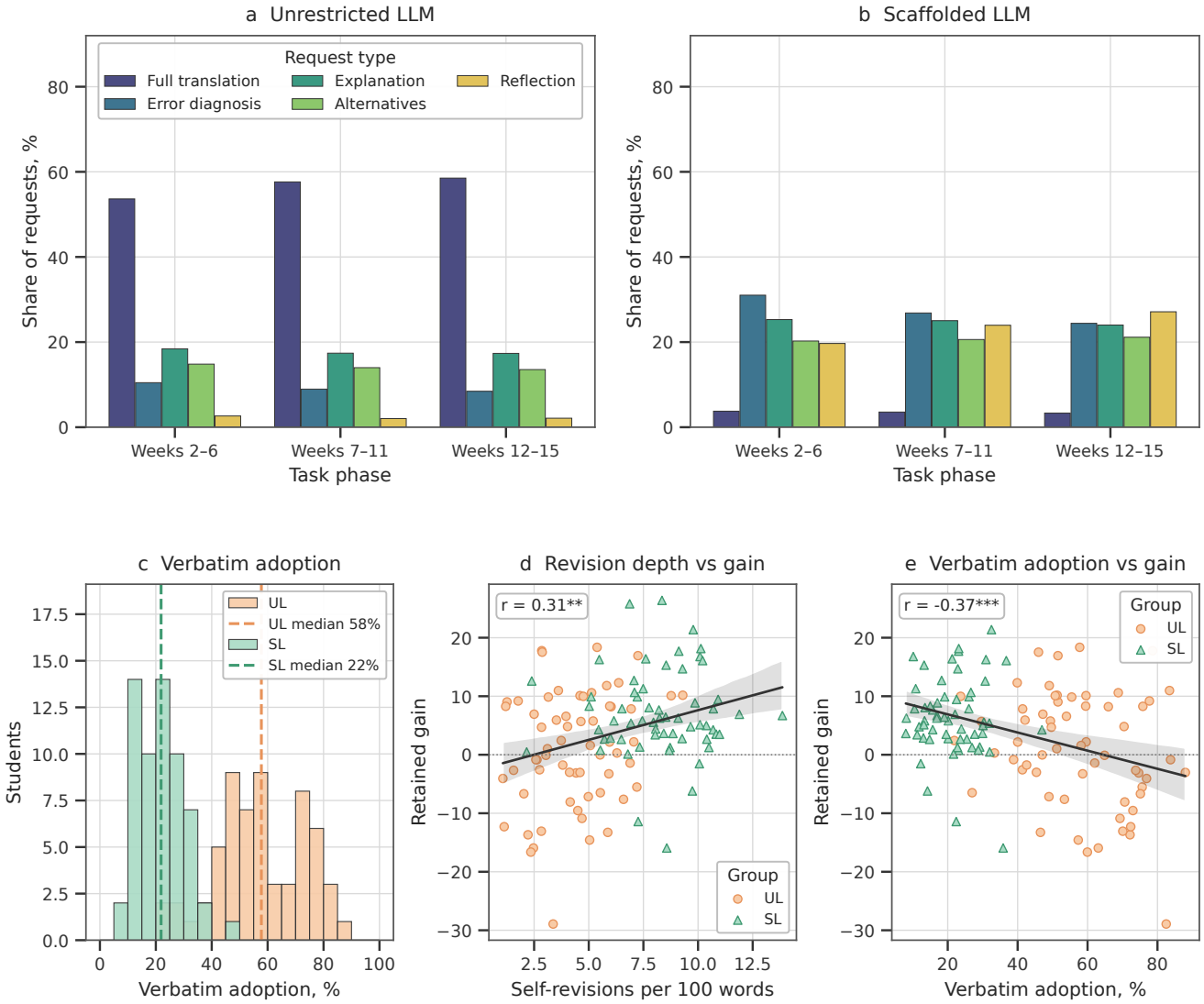


Figure 8. Interaction processes in the two LLM groups. (**a, b**) Composition of coded LLM requests in the initial (weeks 2–6), middle (weeks 7–11) and final (weeks 12–15) phases of the semester in the unrestricted (a) and scaffolded (b) groups. (**c**) Distribution of verbatim adoption, i.e., the proportion of each student's final translations copied from LLM output (longest-common-subsequence overlap); dashed lines mark group medians. (**d, e**) Retained gain as a function of substantive self-revisions (d) and verbatim adoption (e); lines show pooled least-squares fits across the LLM groups with 95% confidence bands; ***Pearson $p < 0.001$.

Table 5. Parallel mediation of the scaffolding effect on retained gain (LLM groups, n = 121). Condition coded SL = 1, UL = 0; all paths adjusted for pretest score. Verbatim adoption is expressed in units of 10 percentage points. CIs are percentile intervals from 2,000 bootstrap resamples.

Path	Estimate	95% bootstrap CI	Interpretation
a_1: SL → Δ metacognitive regulation	0.41	[0.26, 0.57]	
a_2: SL → verbatim adoption (per 10%)	-3.60	[-4.02, -3.17]	
b_1: Δ metacognition → retained gain	2.12	[-2.01, 6.01]	
b_2: verbatim adoption → retained gain	-1.42	[-2.69, -0.23]	
Indirect via metacognition ($a_1 b_1$)	0.86	[-0.84, 2.50]	n.s.
Indirect via verbatim adoption ($a_2 b_2$)	5.12	[0.80, 9.79]	significant
Direct effect (c')	-0.04	[-5.75, 5.30]	n.s.
Total effect (c)	5.94		direct effect ≈ 0

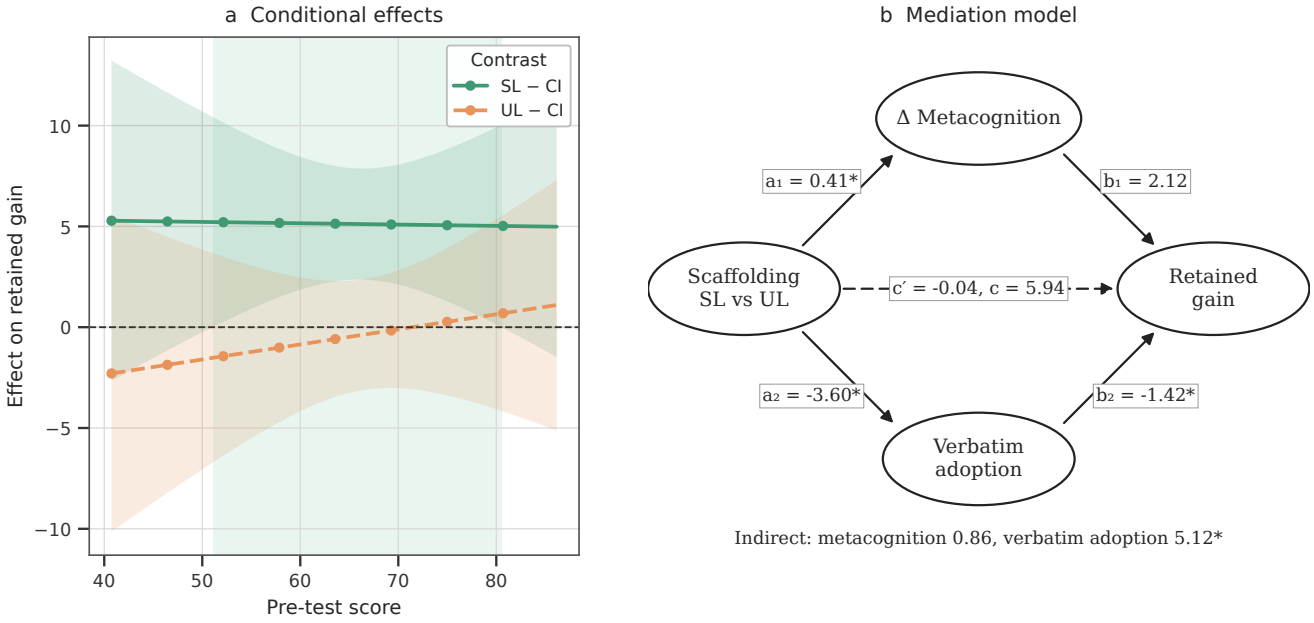


Figure 9. Moderation by prior proficiency and mediation of the scaffolding effect. (**a**) Johnson–Neyman plot of the conditional effects of SL and UL relative to CI on retained gain across pretest scores; shaded bands are 95% CIs and the light green region marks pretest scores at which the SL effect is significant. (**b**) Parallel mediation model comparing SL with UL; *95% bootstrap CI excludes zero; the dashed arrow denotes a non-significant direct effect. Bootstrap CIs are given in Table 5.

in UL (CI 5.51, UL 4.14, SL 5.15; $F(2, 176) = 17.85$, $p < 0.001$), consistent with offloading of the core task, and extraneous load was also lower in the LLM groups (3.84, 2.81 and 3.32; $F = 6.86$, $p = 0.001$), whereas germane load—the effort invested in schema construction—was markedly higher in SL (5.41, 4.99 and 6.76; $F = 25.05$, $p < 0.001$). Finally, ANCOVAs on post-semester self-beliefs showed group effects on translation self-efficacy ($F(2, 175) = 8.74$, $p < 0.001$, $\eta_p^2 = 0.09$) and metacognitive regulation ($F(2, 175) = 17.20$, $p < 0.001$, $\eta_p^2 = 0.16$; Fig. 10c). SL exceeded both other groups on both measures. UL students, by contrast, reported slightly higher self-efficacy than CI students ($b = 0.17$, $p = 0.035$) but lower metacognitive regulation ($b = -0.15$, $p = 0.039$), a pattern suggestive of an illusion of competence.

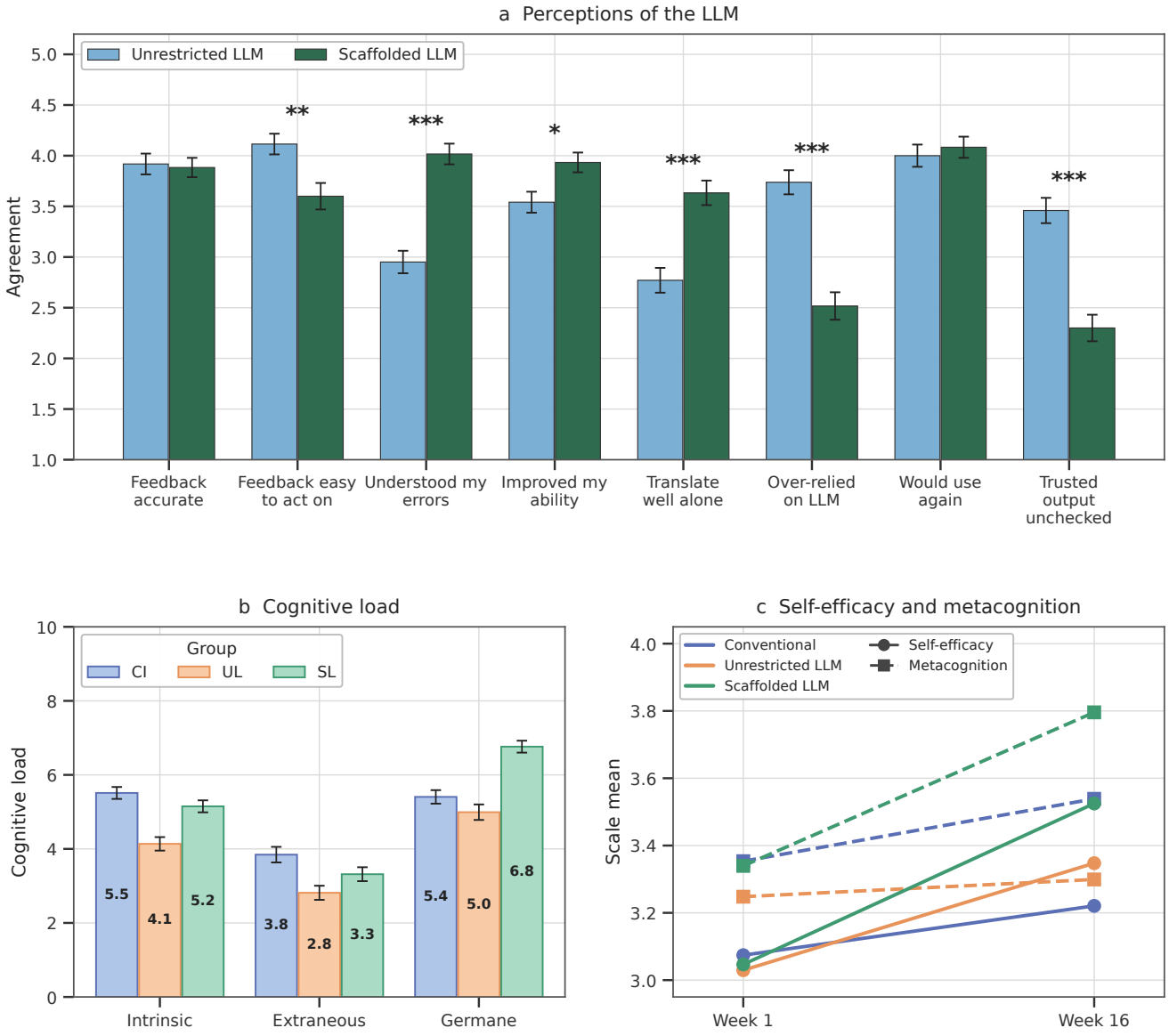


Figure 10. Learner perceptions, cognitive load and self-beliefs. (**a**) Mean agreement (± SE) with end-of-semester perception items on a 5-point scale in the unrestricted and scaffolded groups; ***Mann–Whitney $p < 0.001$. (**b**) Intrinsic, extraneous and germane cognitive load on a 0–10 scale, mean ± SE. (**c**) Group means of translation self-efficacy (solid) and metacognitive regulation (dashed) at weeks 1 and 16.

Discussion

This study set out to determine whether LLMs help translation students learn or merely help them perform. The answer depends on design. When students had unrestricted access to a state-of-the-art LLM for a full semester, they became highly proficient at producing good translations *with* the model, but their ability to translate *without* it was no better than that of students taught conventionally, and the gap between the two kinds of performance was about twice as large as in either other group. When the same model was embedded in a scaffolded cycle that preserved independent drafting, delivered validated diagnostic feedback instead of answers and required justification and reflection, students achieved substantial gains in unassisted translation that survived a summer without instruction, fewer meaning-level and discourse-level errors and stronger metacognitive regulation, while losing little in assisted performance. These findings extend to translation education the dissociation between AI-assisted performance and learning previously reported in mathematics[18] and the positive effects of research-informed AI tutoring[19], and they are consistent with the observation that adding an LLM to translation groups changes the process without improving the product[26]. Because UL and SL used the same model, our data also show that the effect of LLM use is not a property of the technology: the same tool produced either no learning advantage or a large one depending on how it was orchestrated, which helps to reconcile the optimistic meta-analytic literature[16]—which largely measures performance with the tool and rarely separates technology from method[17]—with process studies warning of metacognitive laziness and cognitive offloading[20,22,23].

The process data clarify why the conditions diverged. Unrestricted access turned the LLM predominantly into a translation engine, and more than half of the final text of UL students was adopted verbatim. Such use offloads source-text analysis and problem solving, the sub-competences that models of translation competence place at the centre of expertise[7,8], and it lowered intrinsic load at the cost of germane processing[53]. The D2R guardrails reversed this division of labour. Because the tutor flagged error spans and categories without supplying replacements, students had to generate solutions themselves, which is a desirable difficulty that slows performance but strengthens learning[24], and the requirement to justify each revision is a form of prompted self-explanation[42]. The mediation analysis is consistent with this account: the effect of scaffolding on retained competence was carried mainly by lower verbatim adoption, with no direct effect remaining. Scaffolding also clearly strengthened metacognitive regulation, but the link between this self-reported change and retained gain was too imprecise to establish it as a second pathway, possibly because change scores on an 8-item scale are noisy measures of what is ultimately a process variable. In terms of feedback theory[5], the D2R tutor moved feedback from the task level ("here is a better translation") to the process and self-regulation levels ("this span contains a major accuracy error; what in the source did you misread?"), the levels most strongly associated with transfer.

Our results also have methodological implications for AI-supported education. Feedback-centred designs presuppose that the feedback is correct, yet the accuracy of LLM feedback is rarely verified before classroom deployment. By benchmarking candidate engines against public expert MQM annotations[29–31], we could show that rubric-guided LLM error annotation reaches segment-level agreement with experts comparable to that of the strongest learned metric[35,36], although not its system-level accuracy, and that agreement is, if anything, higher on learner translations. Prompt design mattered for every model, whereas merely asking for error spans did not, which suggests that the instructional value of an LLM tutor depends as much on the evaluative scaffolding embedded in its prompt as on the underlying model[38,39]. At the same time, even the best configuration flagged error categories with a precision of only about 0.4 to 0.6, so a sizeable share of its diagnoses are false alarms, and detection of style and cohesion problems was weakest, although these are precisely the discourse-level competences that distinguish advanced translators. A design in which students must evaluate each diagnosis before acting on it turns this imperfection into a learning opportunity rather than a risk. Teacher audits and the explicit flagging of low-confidence categories, as implemented here, are therefore a necessary complement rather than an optional extra, and they position the teacher as the arbiter of AI feedback rather than its competitor[12,15].

For practice, three implications follow. First, banning LLMs and granting unrestricted access are both suboptimal: CI students retained less than SL students and were weaker LLM users, whereas UL students became dependent. Second, guardrails should be designed around the competence one wishes to develop; in translation this means protecting independent drafting and source-text analysis, delaying access to full renderings and rewarding justified revision. Third, assessment must include unassisted and delayed measures, because assisted performance alone would have ranked the unrestricted condition first and concealed its cost. More broadly, translator education is likely to require explicit training in evaluating and post-editing LLM output[9,13,14]; our data suggest that such skills are built most effectively on top of, rather than instead of, students' own translation competence.

Several limitations should be noted. First, the study was a quasi-experiment with intact classes taught at one university; although the groups were equivalent at baseline and analyses controlled for prior proficiency, unmeasured class or teacher effects cannot be excluded, and a cluster-randomised trial across institutions would strengthen causal inference. Second, the conditions differed in more than one respect (guardrails, validated feedback engine, reflection tasks), so the present design cannot isolate the contribution of each D2R component; component-wise ablations are an important next step. Third, the work focused on English–Chinese translation by advanced undergraduates; effects may differ for lower-resource language pairs,

where LLM feedback is less accurate, and for professional trainees. Fourth, the LLM-feedback validation relied partly on news-domain WMT data, and specialised genres such as legal or literary texts may pose different challenges. Fifth, the delayed test took place 12 weeks after the intervention, across a summer break during which students' private use of LLMs could not be controlled, and longer follow-ups are needed to determine whether the dependency observed in the unrestricted group persists or is reversible. Finally, verbatim adoption and request types are behavioural proxies for cognitive engagement, and the request types were assigned by a validated automatic classifier rather than coded entirely by hand; combining them with keystroke logging, eye tracking or think-aloud protocols would provide a more fine-grained picture of how students process LLM feedback.

In conclusion, LLMs can be powerful allies in translator education, but only when they are positioned as diagnosticians rather than as substitutes for the learner's own work. A scaffolded design that validates the model's feedback, keeps the generative work with the learner and makes reflection obligatory produced durable gains in unassisted translation competence, whereas unrestricted access produced proficient tool users who were no better translators than their conventionally taught peers. As LLMs become ubiquitous, the central pedagogical question is no longer whether students should use them, but how their use should be structured so that the competence remains with the student.

Conflict of interest

The authors declare no competing interests.

Data availability

The WMT22 and WMT23 MQM annotations are publicly available from the MT Metrics Eval repository (https://github.com/google-research/mt-metrics-eval), and FLORES-200 is available from its original release[50]. Anonymised participant-level data, the learner MQM annotations, the coded interaction logs and the analysis code are available from the corresponding author upon reasonable request.

Author contributions statement

X.Z. conceived and designed the study and the translation experiment, collected experimental data, and drafted the manuscript. G.J. developed the LLM intervention framework and assisted with data analysis. All authors reviewed and revised the manuscript and approved the final version.

Funding

This work was supported by the Youth Research Project of Wuhan College of Arts & Sciences (Grant No. 2026xqk17).

Code availability

The D2R-MQM prompts, platform guardrail configuration and analysis scripts are available from the corresponding author upon reasonable request.

References

1. Ouyang, L. *et al.* Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, vol. 35, 27730–27744 (2022).
2. OpenAI. GPT-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).
3. Hendy, A. *et al.* How good are GPT models at machine translation? A comprehensive evaluation. *arXiv preprint arXiv:2302.09210* (2023).
4. Jiao, W. *et al.* Is ChatGPT a good translator? Yes with GPT-4 as the engine. *arXiv preprint arXiv:2301.08745* (2023).
5. Hattie, J. & Timperley, H. The power of feedback. *Rev. Educ. Res.* **77**, 81–112 (2007).
6. Kohnke, L., Moorhouse, B. L. & Zou, D. ChatGPT for language teaching and learning. *RELC J.* **54**, 537–550 (2023).
7. PACTE. Building a translation competence model. In Alves, F. (ed.) *Triangulating Translation: Perspectives in Process Oriented Research*, 43–66 (John Benjamins, Amsterdam, 2003).
8. Göpferich, S. Towards a model of translation competence and its acquisition: the longitudinal study TransComp. In Göpferich, S., Jakobsen, A. L. & Mees, I. M. (eds.) *Behind the Mind: Methods, Models and Results in Translation Process Research*, 11–37 (Samfundslitteratur, Copenhagen, 2009).

9. European Master's in Translation. EMT competence framework 2022. European Commission, Directorate-General for Translation (2022).

10. Lee, S.-M. The impact of using machine translation on EFL students' writing. *Comput. Assist. Lang. Learn.* **33**, 157–175 (2020).

11. Klimova, B., Pikhart, M., Benites, A. D., Lehr, C. & Sanchez-Stockhammer, C. Neural machine translation in foreign language teaching and learning: a systematic review. *Educ. Inf. Technol.* **28**, 663–682, DOI: 10.1007/s10639-022-11194-2 (2023).

12. Massey, G. & Ehrensberger-Dow, M. Machine learning: implications for translator education. *Lebende Sprachen* **62**, 300–312, DOI: 10.1515/les-2017-0021 (2017).

13. O'Brien, S. Translation as human–computer interaction. *Transl. Spaces* **1**, 101–122 (2012).

14. Koponen, M. Is machine translation post-editing worth the effort? A survey of research into post-editing and effort. *The J. Specialised Transl.* **25**, 131–148 (2016).

15. Kiraly, D. *A Social Constructivist Approach to Translator Education: Empowerment from Theory to Practice* (St. Jerome, Manchester, 2000).

16. Deng, R., Jiang, M., Yu, X., Lu, Y. & Liu, S. Does ChatGPT enhance student learning? A systematic review and meta-analysis of experimental studies. *Comput. & Educ.* **227**, 105224, DOI: 10.1016/j.compedu.2024.105224 (2025).

17. Weidlich, J., Gašević, D., Drachsler, H. & Kirschner, P. ChatGPT in education: an effect in search of a cause. *J. Comput. Assist. Learn.* **41**, e70105 (2025).

18. Bastani, H. *et al.* Generative AI without guardrails can harm learning: evidence from high school mathematics. *Proc. Natl. Acad. Sci.* **122**, e2422633122, DOI: 10.1073/pnas.2422633122 (2025).

19. Kestin, G., Miller, K., Klales, A., Milbourne, T. & Ponti, G. AI tutoring outperforms in-class active learning: an RCT introducing a novel research-based design in an authentic educational setting. *Sci. Reports* **15**, 17458, DOI: 10.1038/s41598-025-97652-6 (2025).

20. Fan, Y. *et al.* Beware of metacognitive laziness: effects of generative artificial intelligence on learning motivation, processes, and performance. *Br. J. Educ. Technol.* **56**, 489–530, DOI: 10.1111/bjet.13544 (2025).

21. Risko, E. F. & Gilbert, S. J. Cognitive offloading. *Trends Cogn. Sci.* **20**, 676–688 (2016).

22. Gerlich, M. AI tools in society: impacts on cognitive offloading and the future of critical thinking. *Societies* **15**, 6 (2025).

23. Kosmyna, N. *et al.* Your brain on ChatGPT: accumulation of cognitive debt when using an AI assistant for essay writing task. *arXiv preprint arXiv:2506.08872* (2025).

24. Bjork, R. A. Memory and metamemory considerations in the training of human beings. In Metcalfe, J. & Shimamura, A. (eds.) *Metacognition: Knowing about Knowing*, 185–205 (MIT Press, Cambridge, MA, 1994).

25. Barrot, J. S. Using ChatGPT for second language writing: pitfalls and potentials. *Assess. Writ.* **57**, 100745 (2023).

26. Quan, X. The impact of large language models on collaborative learning processes and outcomes in higher education for English translation. *Sci. Reports* **16**, 18345, DOI: 10.1038/s41598-026-52424-8 (2026).

27. Ji, Z. *et al.* Survey of hallucination in natural language generation. *ACM Comput. Surv.* **55**, 248, DOI: 10.1145/3571730 (2023).

28. Lommel, A., Uszkoreit, H. & Burchardt, A. Multidimensional Quality Metrics (MQM): a framework for declaring and describing translation quality metrics. *Tradum@GRAVEACCENT@atica* **12**, 455–463 (2014).

29. Freitag, M. *et al.* Experts, errors, and context: a large-scale study of human evaluation for machine translation. *Transactions Assoc. for Comput. Linguist.* **9**, 1460–1474 (2021).

30. Freitag, M. *et al.* Results of WMT22 metrics shared task: stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, 46–68 (2022).

31. Freitag, M. *et al.* Results of WMT23 metrics shared task: metrics might be guilty but references are not innocent. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, 578–628 (2023).

32. Papineni, K., Roukos, S., Ward, T. & Zhu, W.-J. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311–318 (2002).

33. Popović, M. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, 392–395 (2015).

34. Rei, R., Stewart, C., Farinha, A. C. & Lavie, A. COMET: a neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2685–2702 (2020).

35. Rei, R. *et al.* COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, 578–585 (2022).

36. Guerreiro, N. M. *et al.* xCOMET: transparent machine translation evaluation through fine-grained error detection. *Transactions Assoc. for Comput. Linguist.* **12**, 979–995 (2024).

37. Kocmi, T. & Federmann, C. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, 193–203 (2023).

38. Kocmi, T. & Federmann, C. GEMBA-MQM: detecting translation quality error spans with GPT-4. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, 768–775 (2023).

39. Fernandes, P. *et al.* The devil is in the errors: leveraging large language models for fine-grained machine translation evaluation. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, 1066–1083 (2023).

40. Vygotsky, L. S. *Mind in Society: The Development of Higher Psychological Processes* (Harvard University Press, Cambridge, MA, 1978).

41. Wood, D., Bruner, J. S. & Ross, G. The role of tutoring in problem solving. *J. Child Psychol. Psychiatry* **17**, 89–100 (1976).

42. Chi, M. T. H., De Leeuw, N., Chiu, M.-H. & LaVancher, C. Eliciting self-explanations improves understanding. *Cogn. Sci.* **18**, 439–477 (1994).

43. Google Research. MT Metrics Eval: data and tools for evaluating machine translation metrics (WMT MQM annotations). *GitHub* https://github.com/google-research/mt-metrics-eval (2023).

44. Dubey, A., Jauhri, A., Pandey, A. *et al.* The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783* (2024).

45. Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115* (2024).

46. DeepSeek-AI. DeepSeek-V3 technical report. *arXiv preprint arXiv:2412.19437* (2024).

47. Hurst, A., Lerer, A., Goucher, A. P. *et al.* GPT-4o system card. *arXiv preprint arXiv:2410.21276* (2024).

48. Wei, J. *et al.* Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, vol. 35, 24824–24837 (2022).

49. Goyal, N. *et al.* The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions Assoc. for Comput. Linguist.* **10**, 522–538 (2022).

50. NLLB Team. Scaling neural machine translation to 200 languages. *Nature* **630**, 841–846, DOI: 10.1038/s41586-024-07335-x (2024).

51. Bandura, A. Guide for constructing self-efficacy scales. In Pajares, F. & Urdan, T. (eds.) *Self-Efficacy Beliefs of Adolescents*, 307–337 (Information Age Publishing, Greenwich, CT, 2006).

52. Schraw, G. & Dennison, R. S. Assessing metacognitive awareness. *Contemp. Educ. Psychol.* **19**, 460–475 (1994).

53. Leppink, J., Paas, F., Van der Vleuten, C. P. M., Van Gog, T. & Van Merriënboer, J. J. G. Development of an instrument for measuring different types of cognitive load. *Behav. Res. Methods* **45**, 1058–1072 (2013).

54. Holm, S. A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* **6**, 65–70 (1979).

55. Hedges, L. V. Distribution theory for Glass's estimator of effect size and related estimators. *J. Educ. Stat.* **6**, 107–128 (1981).

56. Hayes, A. F. *Introduction to Mediation, Moderation, and Conditional Process Analysis: A Regression-Based Approach* (Guilford Press, New York, 2018), 2nd edn.

57. Preacher, K. J. & Hayes, A. F. Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behav. Res. Methods* **40**, 879–891 (2008).

58. Koo, T. K. & Li, M. Y. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J. Chiropr. Medicine* **15**, 155–163 (2016).

59. Seabold, S. & Perktold, J. Statsmodels: econometric and statistical modeling with Python. In *Proceedings of the 9th Python in Science Conference*, 92–96 (2010).

60. Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272, DOI: 10.1038/s41592-019-0686-2 (2020).