SLATE: Are AI-Generated Slides Educationally Effective? A Benchmark for Language Teaching Quality and Learner Knowledge Acquisition

Anonymous ACL submission

Abstract

LLMs have achieved remarkable capabilities in generating language teaching slides. However, a critical mismatch persists between visual polish and actual instructional effectiveness. To address this gap, we introduce **SLATE** (**S**lide-based **L**earning **A**ssessment for **T**eaching **E**ffectiveness), the first benchmark that evaluates AI-generated language teaching slides through instructional effectiveness and learner knowledge acquisition. **SLATE** transforms linguistics olympiad puzzles from low-resource languages with negligible web presence into 90 standardized instructional units comprising 1,133 assessable items, paired with a structured course outline and matched near- and far-transfer test sets. This pretest–posttest design eliminates pretrained knowledge leakage, ensuring gains reflect learning rather than prior recall. Using VLMs as scalable learner proxies and directionally supported by a three-system human pilot, our results show that content validity exhibits a weak association with learning gain, while pedagogical design exhibits a robust positive association. Moreover, most systems show a significant gap between near- and far-transfer accuracy, and even frontier models can produce negative learning gains. **SLATE** reveals a dissociation between artifact quality and instructional effectiveness, calling for a paradigm shift in how generative teaching systems are built, evaluated, and deployed.

1 Introduction

Large language models have revolutionized AI slide generation. Frontier systems such as PPTAgent (Zheng et al., 2025a), AutoPresent (Ge et al., 2025), and SlideBot (Xie et al., 2025) produce visually polished, professionally structured teaching slides. AI-generated slides have become a core application in intelligent education (Liang et al., 2025), and are widely adopted by language teachers for lesson preparation and classroom instruction. Corresponding evaluation benchmarks such as PPTEval (Zheng et al., 2025b), SlidesGen-Bench (Yang et al., 2026), and PresentBench (Chen et al., 2026) assess artifact-level properties including content coherence, visual design, and structural completeness (Maurya et al., 2025).

Benchmark	Scope		Evaluation			
	Slide gen.	Lang. teaching	Artifact quality	Pedagog. grounding	Learner outcomes	Transfer eval.
PPTEval (Zheng et al., 2025b)	✓	✗	✓	✗	✗	✗
SlidesGen-Bench (Yang et al., 2026)	✓	✗	✓	✗	✗	✗
PresentBench (Chen et al., 2026)	✓	✗	✓	✗	✗	✗
TutorBench (Srinivasa et al., 2025)	✗	✓	✗	✓	✗	✗
MathTutorBench (Macina et al., 2025)	✗	✓	✗	✓	✗	✗
SLATE (Ours)	✓	✓	✓	✓	✓	✓

Table 1: Comparison of **SLATE** with existing slide generation and educational benchmarks.

However, many AI-generated language-teaching slides that perform well on these artifact metrics show highly variable instructional outcomes in practice (Do et al., 2024). Decks generated directly from frontier systems often fail to match learners' cognitive pace (Han and Choi, 2025) or align with established patterns of second language acquisition (Yue et al., 2025). More critically, existing evaluation paradigms suffer from a fundamental limitation: they evaluate slides solely on surface-level properties, without ever measuring learner knowledge acquisition, conceptual understanding, or knowledge transfer (Shi et al., 2025). This creates a paradox illustrated in Figure 1: a visually polished and structurally complete teaching deck may produce no effective learning gain, or even result in negative learning outcomes.

To close this gap, we introduce **SLATE** (**S**lide-based **L**earning **A**ssessment for **T**eaching **E**ffectiveness), the first benchmark dedicated to evaluating AI-generated language-teaching slides through instructional effectiveness and learner transfer. As shown in Table 1, **SLATE** is the first to unify learning gain measurement, near/far-transfer evaluation, pretraining leakage control, and human validation within a single assessment framework. Grounded in cognitive load theory, multimedia learning theory, and transfer taxonomy,

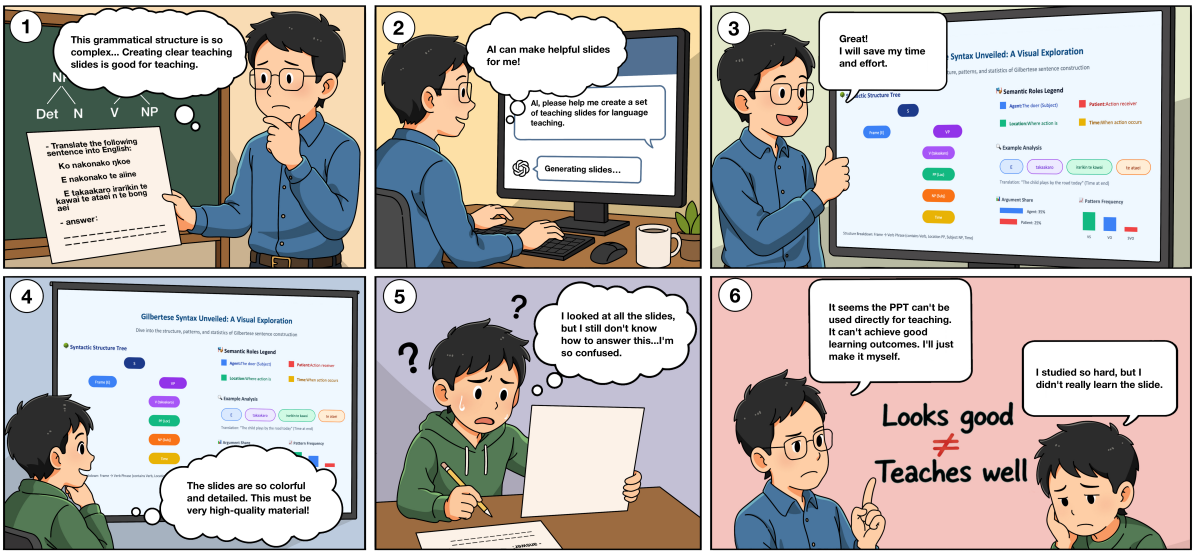

Figure 1: AI-generated slides can appear visually polished and pedagogically plausible, yet still fail to help learners acquire target concepts or solve novel problems after study.

SLATE evaluates decks along four complementary, theory-driven dimensions: linguistic knowledge validity (D1), cognitive-pedagogical design (D2), visual communication quality (D3), and learning and transfer outcomes (D4). **SLATE** transforms linguistics olympiad puzzles from low-resource languages (Bean et al., 2024) into 90 standardized instructional units comprising 1,133 assessable items, and structured course outlines and matched near- and far-transfer test sets.

Evaluating 10 frontier systems on **SLATE** reveals a systematic dissociation between artifact quality and instructional effectiveness: content validity correlates weakly with learning gain, while pedagogical design correlates strongly. All systems show a near/far transfer gap, supporting local pattern reuse but failing to scaffold generalization. Critically, several systems produce *negative* learning gains, a failure mode largely undetectable by artifact-based rubrics. These findings are supported by a 30-participant three-system human pilot (§5.4) and a reliability analysis (§5.1) showing high agreement between our multi-agent framework and expert judgments. Quantitative and qualitative analyses (§4.4) unpack these patterns and mechanisms.

Our core contributions are threefold: ❶ First, we introduce **SLATE**, the first benchmark to evaluate AI-generated language-teaching slides via instructional effectiveness and learner transfer, featuring a theory-grounded four-dimensional framework (D1–D4) and 90 standardized instructional units. ❷ We establish a controlled evaluation protocol grounded in low-resource language tasks, mitigating pretraining knowledge leakage via a pretest–study–posttest–transfer pipeline validated against human learners with 85.7% directional agreement and perfect cross-system ranking consistency (n=30). ❸ We empirically reveal a systematic dissociation between artifact-level quality and learning outcomes across 10 frontier systems: pedagogical design (D2) rather than content accuracy (D1) dominates instructional effectiveness, with several systems producing negative gains invisible to existing rubrics.

2 Related Work

Slide Generation and Evaluation. Driven by rapid advances in multimodal generative AI (Zhang et al., 2026), slide generation has advanced from template-based summarization (Fu et al., 2022) to sophisticated agentic pipelines. PPTAgent (Zheng et al., 2025a) infers functional schemas from reference decks; AutoPresent (Ge et al., 2025) generates Python-pptx code directly; SlidesGen (Liang et al., 2025) decomposes creation into coordinated subtasks; SlideBot (Xie et al., 2025) incorporates principles from learning science into slide planning; and SlideTailor (Zeng et al., 2025) personalizes scientific-deck generation. Existing evaluation frameworks—PPTEval (Zheng et al., 2025b), SlidesGen-Bench (Yang et al., 2026), and PresentBench (Chen et al., 2026)—score content, design, and coherence as artifact properties (Do et al., 2024). All existing work treats slides as *artifacts*: none evaluates whether studying a generated deck

improves learner outcomes.

Educational Benchmarks and Learning Transfer. Most educational benchmarks position LLMs as *students*: LINGOLY (Bean et al., 2024), ModeLing (Chi et al., 2024), PuzzLing (Şahin et al., 2020), AGIEval (Zhong et al., 2023), and MMLU (Hendrycks et al., 2021) test reasoning and knowledge mastery but do not evaluate whether models can *generate* effective teaching materials. TutorBench (Srinivasa et al., 2025) benchmarks interactive tutoring quality, yet still targets model behavior rather than artifact effectiveness (Gao et al., 2025). SLATE shifts this paradigm by evaluating generated artifacts through a pretest–study–posttest–transfer protocol grounded in transfer taxonomy (Barnett and Ceci, 2002), which predicts that far transfer requires explicit abstraction support beyond local pattern reuse (Perkins and Salomon, 1999). A comprehensive review of related work is provided in Appendix A.

3 The SLATE Benchmark

3.1 Task Definition

SLATE formulates language-teaching slide generation as a conditional instructional artifact generation task. Given a puzzle-derived instructional unit U and a structured course outline O, a generation model $\mathcal{G}$ produces a slide deck $S = \mathcal{G}(U, O)$ whose goal is to teach the target linguistic concept. A successful deck must preserve linguistic correctness, organize the concept into a learnable sequence, provide visual support for rule induction, and improve learner performance on source and transfer items.

This formulation differs from standard text-to-slide generation in two ways. First, the source is an instructional unit distilled from a rule-learning problem, not a document. Second, the output is evaluated through learner-side outcomes in addition to artifact quality.

3.2 Benchmark Overview

SLATE covers 7 linguistic task types: paradigm completion (37), translation (31), morphophonology (10), numeral systems (6), syntax/word order (4), error detection (1), and language identification (1). We select linguistics olympiad puzzles for three methodological reasons that jointly address key validity threats in educational AI evaluation. First, all puzzles are drawn from low-resource, endangered, or constructed languages with negligible internet presence, ensuring that the macro pretest baseline (23.78%) reflects genuine zero-knowledge performance rather than prior exposure; any post-study gain can therefore be attributed to the slide deck rather than recalled knowledge. Second, the puzzle-answer pairs are rarely indexed in public corpora, substantially reducing the risk that VLM student models retrieve solutions from pretraining memory rather than learning from instruction. Third, the rule-induction structure of each puzzle—inferring generalizations from a finite example set and applying them to novel inputs—natively instantiates the near- and far-transfer distinction central to our evaluation protocol (Barnett and Ceci, 2002), allowing transfer distance to be operationalized precisely rather than approximated. These categories span diverse instructional demands, from local contrastive explanation to compositional abstraction. Each unit is scored on five factors (rule complexity, exception density, representation density, transfer abstraction, distractor strength) and assigned to one of three difficulty bands. Table 2 summarizes the benchmark statistics.

Property	Value
Instructional units	90
Assessment sub-items	1,133
Task types	7
Difficulty bands	3
Basic / Intermediate / Challenge	2 / 70 / 18
Avg. teachable rules per unit	5.8
Near-transfer items per unit	2
Far-transfer items per unit	2
Generation systems evaluated	10
Evaluation dimensions	4 (D1–D4)

Table 2: SLATE benchmark statistics.

3.3 Construction Pipeline

The benchmark is constructed through a four-stage pipeline, illustrated in Figure 2.

Stage 1: Source Puzzle Parsing. Raw linguistics olympiad puzzles are parsed into structured representations containing language name, task type, context, sub-items, and answers. This produces 1,133 assessable sub-items across 90 puzzles.

Stage 2: Instructional-Unit Construction. Each puzzle is converted from a problem-solving format into a concept-centered teaching unit by trained annotators. Annotators extract 2–6 teachable linguistic rules per puzzle (morphological alternations, agreement patterns, structural templates, etc.), assess instructional difficulty on a five-dimensional rubric, and generate matched near-

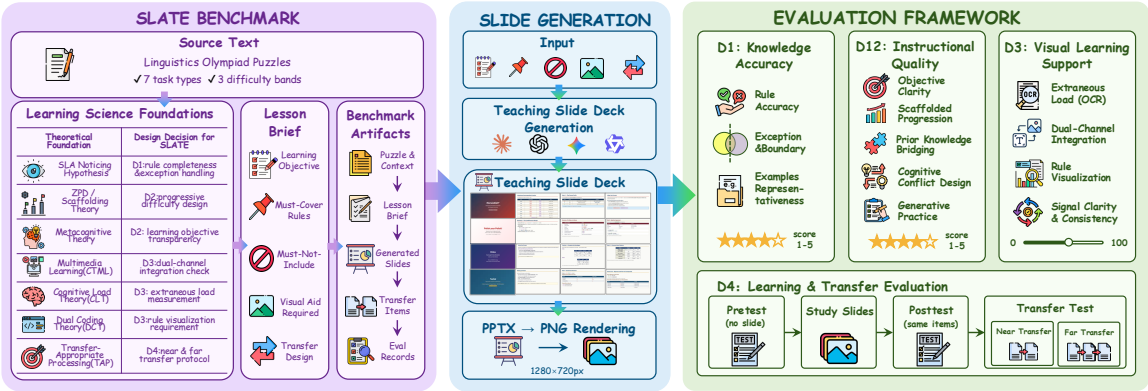

Figure 2: Overview of **SLATE**. Source linguistics olympiad puzzles are transformed into instructional units; generation models produce teaching decks; D1–D4 evaluate content, pedagogy, visual support, and learner transfer.

transfer and far-transfer items. Each unit stores the target rules, key terms, exceptions, representative examples, and transfer items. Annotator qualifications and inter-annotator agreement are reported in Appendix F.

Stage 3: Course-Outline Generation. For each unit, annotators compile a structured outline as the generation specification. The outline contains five mandatory fields: **Learning Objective**, **Must-Cover Rules**, **Must-Not-Include** (prohibiting answer exposure), **Visual Aid Required**, and **Transfer Design**. Transfer answers are never exposed to the generator, preventing evaluation leakage.

Stage 4: Slide Generation. Each generation system receives the same course outline and produces a self-contained HTML teaching deck, which is rendered to PNG pages via Playwright for downstream evaluation. We standardize on HTML to ensure all systems produce output in a uniform, programmatically renderable format, enabling consistent visual evaluation across systems. Slide count targets are difficulty-dependent (Basic: 8–12, Intermediate: 10–14, Challenge: 12–16 pages).

3.4 Evaluation Framework

SLATE evaluates generated decks along four theory-grounded dimensions.

D1: Linguistic Knowledge Validity. D1 assesses whether a deck teaches the intended content correctly, covering rule coverage, data fidelity, metalinguistic accuracy, answer non-leakage, and scope appropriateness.

D2: Cognitive-Pedagogical Design. D2 evaluates instructional organization: cognitive load management, worked examples, scaffolded progression, dual-channel integration, and practice opportunities. Both D1 and D2 are grounded in Cognitive Load Theory (Sweller, 1988) and the Cognitive Theory of Multimedia Learning (Mayer, 2005).

D3: Visual Communication Quality. D3 provides lightweight visual-support diagnostics computed automatically from rendered deck properties, measuring slide count, text density, table usage, color diversity, visual aid completeness, and exercise inclusion. Unlike D1 and D2, D3 does not assess perceptual or cognitive dimensions of visual design; it serves as a coarse filter for identifying decks that lack basic visual scaffolding.

D4: Learning and Transfer. D4 measures learner-side outcomes through a four-stage protocol. The student model first answers source items without instruction (pretest), studies the slide deck as PNG pages, answers the same items again (posttest), and then answers matched near-transfer and far-transfer items. Near-transfer items require applying taught rules to new inputs within the same structural pattern; far-transfer items require generalization across knowledge domain or functional context (Barnett and Ceci, 2002). Reported metrics include posttest gain, near-transfer accuracy, and far-transfer accuracy.

4 Experiments

4.1 Experimental Setup

Generation systems. We evaluate 10 frontier LLM systems under a unified setup: Claude Opus 4.7, Claude Sonnet 4.6, GPT-5.4, Gemini-3.1-Pro, Qwen3.6-Max, Qwen3.6-Plus, DeepSeek-V4-Pro, MiniMax-M2.7, Kimi-K2.6, and GLM-5.1. Each system receives the same course outline for each of the 90 instructional units and generates a self-

contained HTML teaching deck. All systems complete all 90 units.

Student model and D4 protocol. D4 uses Qwen3-VL-8B-Instruct as the student model, selected based on the ablation study (Section§ 4.3), which evaluates four VL models of varying capacity (see Appendix L for full model names) and confirms that the 8B model balances sensitivity to instructional input with sufficient baseline capability. For each unit, the student answers source items without instruction (pretest), studies the rendered slide deck as PNG pages, answers the same items again (posttest), and finally answers near-transfer and far-transfer items. We report puzzle-level macro averages. The pretest baseline is 23.78%. The cross-system ranking validity of this VLM proxy is validated via a three-system human pilot (§5.4, Appendix M, Appendix N).

D1/D2 multi-agent evaluation. D1 and D2 are scored by a multi-agent panel of three heterogeneous LLMs (Claude Opus 4.7, Gemini-3.1-Pro, and GPT-5.4), each independently rating all 10 sub-dimensions (D1.1–D1.5, D2.1–D2.5) on a 1–5 scale. Final scores are averaged across the three judges to reduce single-model bias. Each reported score represents the mean of 5 independent runs per judge.

To validate this setup, three linguistics experts (all with doctoral training; two with low-resource language experience; one with >3 years of language-teaching and slide-design experience) independently rated a stratified sample of 29 decks after rubric calibration. We report inter-rater reliability and human–model agreement in Section§ 5.

D3 evaluation. D3 is computed automatically from rendered HTML properties: slide count, words per slide (WPS), table count, color diversity, visual aid completeness (VA%), and exercise inclusion rate.

4.2 Main Results

Table 3 presents the full results across all four dimensions. On D4, Claude Opus 4.7 and Qwen3.6-Max achieve the largest positive gains (↑5.46% and ↑5.14%), followed by MiniMax-M2.7 (↑3.27%) and GPT-5.4 (↑2.61%). Four systems produce negative gains, confirming that SLATE discriminates systems on instructional effectiveness rather than artifact plausibility alone.

The most striking pattern in Table 3 is the dissociation between artifact quality and learning outcomes. GPT-5.4 scores highest on D1 (4.20) yet ranks only fourth on gain; MiniMax-M2.7 scores below average on D1 (3.77) yet ranks third. This dissociation is not explained by visual design alone, though visual completeness appears necessary: all systems with VA%<80% produce negative gains. Across all systems, near-transfer scores (31–42%) consistently exceed far-transfer scores (11–16%), indicating that current decks support pattern reuse more effectively than conceptual generalization.

4.3 Ablation Study

To isolate the contribution of slides from other factors, we conduct a controlled ablation with four student VLMs of increasing capacity across six conditions (Table 4).

Three findings stand out. First, slides produce positive learning gain for 3 of 4 student models; only the smallest model (Qwen2.5-VL-3B) shows negative slide effect (↓1.4%). This is consistent with the expertise reversal effect in Cognitive Load Theory (Sweller et al., 2011): instructional scaffolding that benefits intermediate learners can impose extraneous cognitive load on learners whose processing capacity is insufficient to integrate multi-page visual input. The monotonic scaling of slide effects across the four models (↓1.4, ↑4.3, ↑4.0, ↑5.0%) rules out the possibility that the 8B result is an isolated artifact: the trend is systematic and capacity-dependent. Second, puzzle context is the strongest single factor (↑5.6 to ↑12.5%), but slides alone can partially compensate for missing context (↑4.1 to ↑10.0% gain over zero-information baseline). Third, full-frame presentation outperforms 3-frame sampling by ∼3%, confirming that intermediate slides contain valuable instructional content. Figure 3 visualizes these effect decompositions across four student models.

4.4 Analysis

High near-transfer, weak far-transfer. Across systems, near-transfer accuracy (∼35–42%) substantially exceeds far-transfer (∼11–16%). This 20+% gap persists even for the strongest systems, suggesting that current decks help learners recognize and reuse local patterns but do not reliably support rule abstraction for novel contexts. Figure 4 visualizes this pattern: systems cluster in the lower-right region, with strong near-transfer but uniformly weak far-transfer regardless of gain.

System	Artifact Quality		D3	D4: Learning & Transfer			
	D1	D2	VA%	Gain (%)	Post	Near	Far
Claude Opus 4.7 (Anthropic, 2026a)	**4.19**	**4.42**	**99**	↑**5.46**	**29.25**	**41.67**	12.22
Qwen3.6-Max (Qwen Team, 2026)	3.33	3.95	**99**	↑5.14	28.92	39.89	12.36
MiniMax-M2.7 (MiniMax, 2026)	3.77	3.91	**99**	↑3.27	27.06	38.33	15.00
GPT-5.4 (OpenAI, 2026)	4.20	4.24	**99**	↑2.61	26.40	35.00	13.33
Claude Sonnet 4.6 (Anthropic, 2026b)	4.17	3.72	98	↑1.85	25.63	**41.67**	12.22
Qwen3.6-Plus (Qwen Team, 2026)	4.01	4.23	**99**	↑0.38	24.17	41.11	13.33
Kimi-K2.6 (Moonshot AI, 2026)	4.01	4.13	91	↓0.89	22.90	36.11	15.56
GLM-5.1 (GLM-5-Team et al., 2026)	4.05	3.60	80	↓1.24	22.54	31.67	13.89
DeepSeek-V4-Pro (DeepSeek-AI, 2026)	3.11	2.02	61	↓1.68	22.10	35.39	11.24
Gemini-3.1-Pro (Google DeepMind, 2026)	3.60	2.88	79	↓2.51	21.28	32.02	**16.29**

Table 3: **Main results on SLATE** (sorted by D4 gain). D1/D2 are multi-agent judge averages (1–5 scale; 5 runs × 3 judges). VA%: visual aid completeness (full D3 breakdown in Appendix). D4 gain is measured against the pretest baseline of 23.78%. **Bold**: best; underline: second best.

Cond.	Setting	Student Model			
		Q25-3B	Q3-4B	Q25-7B	Q3-8B
A	Ctx only	14.6	19.9	16.9	22.9
B	Ctx + slides	13.1	24.2	20.9	28.0
V1	No input	9.0	9.2	8.0	10.4
V2	Slides only	13.1	17.7	16.1	20.4
V3	Ctx + 3-frame	13.3	20.8	17.7	24.4
V4	3-frame only	12.0	12.6	12.1	13.8
Effect decomposition (%)					
B−A	Slide effect	↓1.4	↑4.3	↑4.0	↑5.0
A−V1	Context effect	↑5.6	↑10.7	↑8.8	↑12.5
V2−V1	Slide compens.	↑4.1	↑8.5	↑8.1	↑10.0
B−V3	Full vs. 3-fr.	↓0.2	↑3.4	↑3.1	↑3.5

Table 4: **Ablation study** across 4 student VLMs and 6 conditions using Claude Opus 4.7 slides. Upper block: accuracy (%); lower block: effect sizes (%).

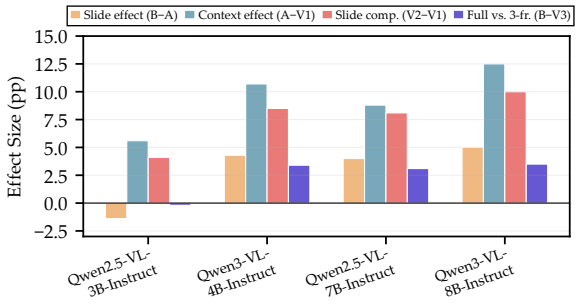

Figure 3: Ablation effect decomposition across 4 student VLMs. Context is the strongest factor; slide effects scale with model capacity; full-frame consistently outperforms 3-frame sampling.

Task-type stratified gains. We group the 90 units into four macro categories: Paradigm Completion (37), Translation (31), Morphophonology (10), and Other (12; numeral systems, syntax, error detection, language ID). We report mean gain per category (Table 5). Paradigm completion is the most teachable category (↑3.5% mean gain; 47.2% near-transfer), likely because paradigm tasks present explicit input-output mappings that naturally lend themselves to tabular visualization and pattern-based instruction. Translation tasks show the weakest gains (↓1.1%), consistent with their reliance on contextual inference and compositional semantics that are harder to scaffold within static visual materials. The near/far transfer gap is largest for paradigm tasks (30.7%) and smallest for the Other category (9.6%), suggesting that pattern-based tasks benefit most from slides but also show the sharpest drop-off when generalization is required. Note that Morphophonology ($n = 10$) and Other ($n = 12$, merging four subtypes) have small sample sizes; conclusions for these categories are exploratory.

Artifact quality vs. learning outcomes. To quantify the relationship between artifact quality and instructional effectiveness, we compute Spearman rank correlations between D1/D2 scores and D4 gain across 10 systems (Figure 5). Within the evaluated systems, D2 (pedagogical design) shows a stronger positive association with gain ($\rho = 0.72$, $p = 0.019$) than D1 (knowledge validity; $\rho = 0.38$, $p = 0.28$). However, with only $N = 10$ systems, confidence intervals are wide and individual outliers can substantially shift these estimates; these correlations should be interpreted as suggestive rather than conclusive. The qualitative pattern is nonetheless informative: GPT-5.4 scores highest on D1 (4.20) yet ranks only fourth on gain, while MiniMax-M2.7 scores below average on D1 (3.77) yet ranks third, illustrating that content validity alone is not sufficient for effective instruction.

Case study. Per-puzzle analysis reveals extreme variation: the best single-puzzle gain is ↑68% (Wik-Mungkan compound nouns, Claude Opus 4.7), while the worst is ↓50% (Jahai smell verb semantics, Gemini-3.1-Pro). In the success case, a

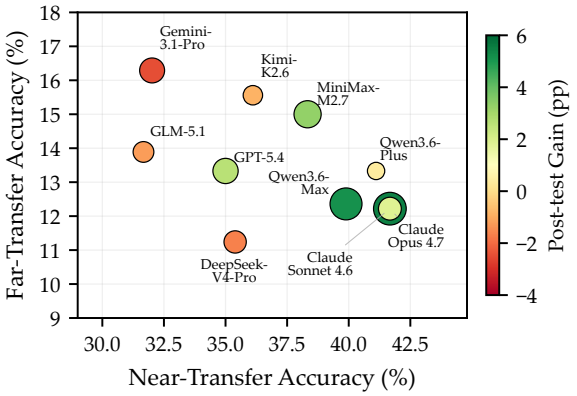

Figure 4: Near- vs. far-transfer accuracy for each system (color = posttest gain, size = |gain|). Systems cluster in a high-near / low-far region, indicating a systematic gap between pattern reuse and conceptual generalization.

System	Parad. *(n=37)*	Trans. *(n=31)*	Morph. *(n=10)*	Other *(n=12)*
Claude Opus 4.7	↑8.5	↑0.8	↑2.2	↑11.1
Qwen3.6-Max	↑9.4	↑0.5	↑2.1	↑6.2
MiniMax-M2.7	↑7.9	↓0.9	↓2.2	↑4.4
GPT-5.4	↑4.5	↓0.4	↑2.3	↑4.9
Claude Sonnet 4.6	↑4.7	↓0.8	↓2.3	↑3.3
Qwen3.6-Plus	↑3.0	↓1.4	↓5.6	↑1.8
Kimi-K2.6	↓0.0	↓1.4	↓2.1	↓1.2
GLM-5.1	↑0.8	↓3.6	↓2.1	↓0.9
DeepSeek-V4-Pro	↓1.2	↓2.5	↓2.4	↓1.9
Gemini-3.1-Pro	↓2.9	↓1.1	↓5.9	↓2.8
Mean	↑3.5	↓1.1	↓1.6	↑2.5

Table 5: **D4 gain (%) by task type.** Paradigm completion is the most teachable; translation tasks show weakest gains.

low-pretest puzzle (4%) reaches 72% after studying a deck with clear tabular rule presentation. In the failure case, a high-pretest puzzle (72%) drops to 22%, suggesting the deck disrupts useful prior knowledge. We also identify an artifact-quality gap case: GPT-5.4 achieves perfect near-transfer (100%) on a Dutch spelling puzzle yet causes a 30% posttest decline, demonstrating that visual polish and local pattern learning do not guarantee effective instruction. Detailed analysis with representative slide excerpts is provided in Appendix G.

5 Discussion

5.1 Evaluation Reliability

Table 6 reports agreement statistics for D1 and D2. Human–human ICC reaches 0.80 for D1 and 0.72 for D2, indicating excellent and good reliability respectively. Human–model Pearson r reaches 0.80 for D1 and 0.69 for D2. Agreement is stronger on objective content-validity sub-dimensions (e.g., rule coverage) and more moderate on subjective pedagogical sub-dimensions (e.g., scaffolding quality). We observe that model judges tend to overscore D2 when decks exhibit surface structural completeness without genuine pedagogical depth, motivating the use of multi-agent panels.

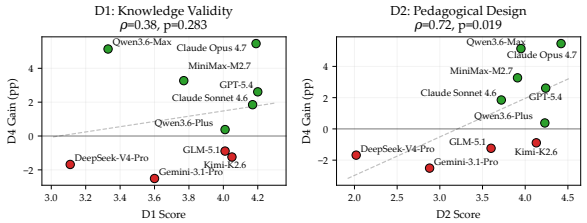

Figure 5: D1/D2 scores vs. D4 gain. D2 (pedagogical design) shows a stronger positive association with learning gain ($\rho = 0.72$) than D1 (knowledge validity; $\rho = 0.38$), though $N = 10$ limits statistical power. Green = positive gain; red = negative gain.

Metric	H–H ICC	H–M Pearson r
D1.1 Rule coverage	0.82	0.81
D1.2 Data fidelity	0.78	0.79
D1.3 Metalinguistic acc.	0.71	0.75
D1.4 Answer non-leakage	0.85	0.83
D1.5 Scope approp.	0.76	0.77
D1 average	0.80	0.80
D2.1 Cognitive load	0.73	0.72
D2.2 Worked examples	0.68	0.68
D2.3 Scaffolding	0.65	0.65
D2.4 Dual-channel integ.	0.70	0.70
D2.5 Practice quality	0.58	0.51
D2 average	0.72	0.69

Table 6: D1/D2 reliability. H–H: human–human ICC; H–M: human–model Pearson r.

5.2 Construct Validity: Error Injection

The error injection study (Appendix E) provides construct-oriented validation by selectively corrupting strong reference decks along four failure families. Each perturbation degrades its target dimension most severely: knowledge corruption drops D1 by 4.5 points (on a 10-point scale) while leaving D2/D3 nearly intact; pedagogical corruption drops D2 by 4.4 points; and visual corruption drops D3 by 5.3 points. Assessment-integrity corruption is diagnostically distinct: it inflates posttest scores by 0.5 points while destroying far-transfer performance (−5.0 points), confirming that answer leakage creates artificial gains without genuine learning. This selective sensitivity supports the claim that **SLATE** dimensions measure interpretable instructional properties rather than surface fluency.

5.3 Failure Modes and Hypothesized Mechanisms

Three recurring failure modes emerge. First, *content-faithfulness drift*: decks appear fluent and pedagogically structured, yet examples or rule formulations contain errors (reflected in weak D1.2 data fidelity scores). This is supported by D1 sub-scores: rule coverage and scope score high across systems (surface-level completeness), but data fidelity is consistently weaker, as plausibility-optimizing models fail to catch constructed-example errors.

Second, *pedagogical shallowness*: systems produce the form of a lesson (headings, worked examples, practice sections) without instructional coherence. D2 correlates far more strongly with D4 gain ($\rho = 0.72$) than D1 ($\rho = 0.38$), showing instructional organization matters more than content correctness alone, consistent with Cognitive Load Theory's emphasis on sequencing and scaffolding.

Third, *near-transfer bias*: all systems support local pattern reuse far better than deeper abstraction, as shown by the persistent 20+% near/far transfer gap. Following Barnett and Ceci (2002), our far-transfer items require cross-domain generalization, which fails without explicit abstraction support. Current decks rarely provide the contrastive examples or analogical bridges needed for novel context application.

5.4 VLM Student Validity

Using a VLM as a learner proxy is scalable but raises validity concerns. The pretest baseline of 23.78% likely reflects pretrained language exposure rather than domain-specific memorization, since the items require structured rule application. The ablation study validates that slide effects scale with model capacity (negative for 3B, positive for 8B+), consistent with research on the expertise reversal effect (Sweller, 1988).

To distinguish genuine learning from contextual priming, we conduct a content perturbation control: preserving slide structure and visual design but replacing teaching content with material from unrelated languages. Under this condition, posttest gains fall below the zero-information baseline, confirming that observed gains are content-driven. We further validate the VLM proxy through a three-system human pilot with 30 participants, covering high-gain Claude Opus 4.7 (Anthropic, 2026a), intermediate-gain Claude Sonnet 4.6 (Anthropic, 2026b), and negative-gain Gemini-3.1-Pro (Google DeepMind, 2026). Human learners show a significant overall posttest gain ($t(29) = 6.78$, $p < 0.001$), with 85.7% per-puzzle directional agreement and perfect cross-system ranking consistency (Spearman $\rho = 1.0$) between VLM and human results. Full experimental details, group-level results and statistical analyses are provided in Appendix M and Appendix N.

5.5 Broader Implications

The dissociation between artifact quality and instructional effectiveness has implications beyond slide generation. Any "generation + evaluation" benchmark that relies solely on artifact-side rubrics risks the same validity gap: materials that appear well-structured may not function as intended when deployed. This parallels findings in the LLM-as-judge literature (Shankar et al., 2024), where judge scores correlate imperfectly with downstream outcomes. We note that **SLATE**'s current scope is limited to explicit rule-based instruction with discrete, teachable rules and clear input-output mappings. Real language teaching also involves context-dependent pragmatics and communicative interaction; extending the evaluation framework to these less rule-governed domains is an important direction for future work.

6 Conclusion

The central finding of this work is that artifact quality and instructional effectiveness are dissociated: AI-generated slides that appear pedagogically well-structured under rubric evaluation do not reliably produce learning gains or support knowledge transfer. Within the evaluated systems, pedagogical design (D2) shows a stronger association with learning outcomes than content validity (D1), suggesting that instructional organization may matter more than factual coverage alone. The persistent near/far transfer gap across all systems indicates that current generation approaches support pattern reuse but not the deeper abstraction needed for genuine conceptual generalization. A content perturbation control and a three-system human pilot (n=30) support the validity of the VLM-based evaluation protocol.

The most important next steps are scaling human validation and incorporating transfer-oriented objectives into generation pipelines. Extending **SLATE** to interactive, multi-session formats is a promising direction for closing the near/far transfer gap.

Limitations

VLM vs. human learners. D4 relies on a VLM student proxy whose cross-system ranking validity has been confirmed via a three-system human pilot (n=30) spanning the full range of instructional effectiveness observed in our experiments. While VLM learners differ from human learners in several aspects—including episodic memory, interactive question-asking ability, and static image processing—their perfect ranking consistency with human learners across tested systems supports their use for scalable cross-system comparison. Extending validation to additional model architectures and larger participant cohorts remains a valuable direction for future work.

Static, single-session evaluation. SLATE evaluates one-shot, non-interactive study. Real instruction involves iterative feedback, learner questions, and spaced repetition, none of which are captured in the current protocol. Extending the benchmark to multi-turn and interactive formats is a natural next step.

Ethics Statement

SLATE is an evaluation benchmark, not evidence that AI-generated teaching slides are ready for unsupervised educational deployment. Incorrect explanations or misleading visualizations may negatively affect learners. Human oversight remains necessary in real instructional settings.

References

Anthropic. 2026a. Introducing claude opus 4.7.

Anthropic. 2026b. System card: Claude sonnet 4.6.

Stefano Bannò, Kate Knill, and Mark J. F. Gales. 2024. Grammatical error feedback: An implicit evaluation approach. *Preprint*, arXiv:2408.09565.

Susan M. Barnett and Stephen J. Ceci. 2002. When and where do we apply what we learn? a taxonomy for far transfer. *Psychological bulletin*, 128 4:612–37.

Andrew M. Bean, Simi Hellsten, Harry Mayne, Jabez Magomere, Ethan A. Chi, Ryan Chi, Scott A. Hale, and Hannah Rose Kirk. 2024. Lingoly: A benchmark of olympiad-level linguistic reasoning puzzles in low-resource and extinct languages. *Preprint*, arXiv:2406.06196.

Xin-Sheng Chen, Jiayu Zhu, Pei lin Li, Hanzheng Wang, Shuojin Yang, and Meng-Hao Guo. 2026. Presentbench: A fine-grained rubric-based benchmark for slide generation. *Preprint*, arXiv:2603.07244.

Nathan A. Chi, Teodor Malchev, Riley Kong, Ryan A. Chi, Lucas Huang, Ethan A. Chi, R. Thomas McCoy, and Dragomir Radev. 2024. modeling: A novel dataset for testing linguistic reasoning in language models. *Preprint*, arXiv:2406.17038.

Chad Clark and Royce Kimmons. 2023. Cognitive load theory. *EdTechnica*.

DeepSeek-AI. 2026. DeepSeek-V4: Towards highly efficient million-token context intelligence.

Tiffany D. Do, Usama Bin Shafqat, Elsie Ling, and Nikhil Sarda. 2024. Paige: Examining learning outcomes and experiences with personalized ai-generated educational podcasts. *Preprint*, arXiv:2409.04645.

Tsu-Jui Fu, William Yang Wang, Daniel McDuff, and Yale Song. 2022. Doc2ppt: Automatic presentation slides generation from scientific documents. *Preprint*, arXiv:2101.11796.

Weibo Gao, Qi Liu, Linan Yue, Fangzhou Yao, Rui Lv, Zheng Zhang, Hao Wang, and Zhenya Huang. 2025. Agent4edu: Generating learner response data by generative agents for intelligent education systems. *Preprint*, arXiv:2501.10332.

Jiaxin Ge, Zora Zhiruo Wang, Xuhui Zhou, Yi-Hao Peng, Sanjay Subramanian, Qinyue Tan, Maarten Sap, Alane Suhr, Daniel Fried, Graham Neubig, and Trevor Darrell. 2025. Autopresent: Designing structured visuals from scratch. *Preprint*, arXiv:2501.00912.

Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. 2022. Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text. *Preprint*, arXiv:2202.06935.

GLM-5-Team, :, Aohan Zeng, Xin Lv, Zhenyu Hou, Zhengxiao Du, Qinkai Zheng, Bin Chen, Da Yin, Chendi Ge, Chenghua Huang, Chengxing Xie, Chenzheng Zhu, Congfeng Yin, Cunxiang Wang, Gengzheng Pan, Hao Zeng, Haoke Zhang, Haoran Wang, and 168 others. 2026. Glm-5: from vibe coding to agentic engineering. *Preprint*, arXiv:2602.15763.

Google DeepMind. 2026. Gemini 3.1 Pro: Best for complex tasks and bringing creative concepts to life.

Richard Hake. 1998. Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *American Journal of Physics - AMER J PHYS*, 66.

Junzhi Han and Jinho D. Choi. 2025. Beyond linear digital reading: An LLM-powered concept mapping approach for reducing cognitive load. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 805–817, Vienna, Austria. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Preprint*, arXiv:2009.03300.

Enkelejda Kasneci, Kathrin Sessler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, Stephan Krusche, Gitta Kutyniok, Tilman Michaeli, Claudia Nerdel, Jürgen Pfeffer, Oleksandra Poquet, Michael Sailer, Albrecht Schmidt, Tina Seidel, and 4 others. 2023. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103:102274.

Xin Liang, Xiang Zhang, Yiwei Xu, Siqi Sun, and Chenyu You. 2025. Slidegen: Collaborative multimodal agents for scientific slide generation. *Preprint*, arXiv:2512.04529.

Xinyi Lu and Xu Wang. 2024. Generative students: Using llm-simulated student profiles to support question item evaluation. In *Proceedings of the Eleventh ACM Conference on Learning @ Scale*, page 16–27. ACM.

Jakub Macina, Nico Daheim, Ido Hakimi, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2025. Mathtutorbench: A benchmark for measuring open-ended pedagogical capabilities of llm tutors. *Preprint*, arXiv:2502.18940.

Kaushal Kumar Maurya, Kv Aditya Srivatsa, Kseniia Petukhova, and Ekaterina Kochmar. 2025. Unifying AI tutor evaluation: An evaluation taxonomy for pedagogical ability assessment of LLM-powered AI tutors. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1234–1251, Albuquerque, New Mexico. Association for Computational Linguistics.

Richard E. Mayer. 2002. Multimedia learning. volume 41 of *Psychology of Learning and Motivation*, pages 85–139. Academic Press.

Richard E. Mayer. 2005. Cognitive theory of multimedia learning. In Richard E. Mayer, editor, *The Cambridge Handbook of Multimedia Learning*, pages 31–48. Cambridge University Press, Cambridge, UK.

MiniMax. 2026. MiniMax-M2.7.

Moonshot AI. 2026. Kimi K2.6: Advancing open-source coding.

C. Donald Morris, John D. Bransford, and Jeffery J. Franks. 1977. Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior*, 16(5):519–533.

John Norris and Lourdes Ortega. 2002. Effectiveness of l2 instruction: A research synthesis and quantitative meta-analysis. *Language Learning*, 50:417 – 528.

OpenAI. 2026. Introducing gpt-5.4.

David Perkins and Gavriel Salomon. 1999. Transfer of learning. 11.

Qwen Team. 2026. Qwen3.6: Large language model series.

RICHARD W. SCHMIDT. 1990. The role of consciousness in second language learning1. *Applied Linguistics*, 11(2):129–158.

Shreya Shankar, J. D. Zamfirescu-Pereira, Björn Hartmann, Aditya G. Parameswaran, and Ian Arawjo. 2024. Who validates the validators? aligning llm-assisted evaluation of llm outputs with human preferences. *Preprint*, arXiv:2404.12272.

Yao Shi, Rongkeng Liang, and Yong Xu. 2025. Educationq: Evaluating llms' teaching capabilities through multi-agent dialogue framework. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, page 32799–32828. Association for Computational Linguistics.

Rakshith S Srinivasa, Zora Che, Chen Bo Calvin Zhang, Diego Mares, Ernesto Hernandez, Jayeon Park, Dean Lee, Guillermo Mangialardi, Charmaine Ng, Ed-Yeremai Hernandez Cardona, Anisha Gunjal, Yunzhong He, Bing Liu, and Chen Xing. 2025. Tutorbench: A benchmark to assess tutoring capabilities of large language models. *Preprint*, arXiv:2510.02663.

John Sweller. 1988. Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12(2):257–285.

John Sweller, Paul Ayres, Slava Kalyuga, and Paul Chandler. 2011. The expertise reversal effect. *Faculty of Education - Papers*, 38.

Eric Xie, Danielle Waterfield, Michael Kennedy, and Aidong Zhang. 2025. Slidebot: A multi-agent framework for generating informative, reliable, multimodal presentations. *Preprint*, arXiv:2511.09804.

Songlin Xu, Hao-Ning Wen, Hongyi Pan, Dallas Dominguez, Dongyin Hu, and Xinyu Zhang. 2025. Classroom simulacra: Building contextual student generative agents in online education for learning behavioral simulation. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, page 1–26. ACM.

Yunqiao Yang, Wenbo Li, Houxing Ren, Zimu Lu, Ke Wang, Zhiyuan Huang, Zhuofan Zong, Mingjie Zhan, and Hongsheng Li. 2026. SlidesGenBench: Evaluating slides generation via computational and quantitative metrics. *arXiv preprint arXiv:2601.09487*.

Zhihao Yuan, Yunze Xiao, Ming Li, Weihao Xuan, Richard Tong, Mona Diab, and Tom Mitchell. 2026. Towards valid student simulation with large language models. *Preprint*, arXiv:2601.05473.

Murong Yue, Wenhan Lyu, Jennifer Suh, Yixuan Zhang, and Ziyu Yao. 2025. Mathvc: An llm-simulated multi-character virtual classroom for mathematics education. *Preprint*, arXiv:2404.06711.

Wenzheng Zeng, Mingyu Ouyang, Langyuan Cui, and Hwee Tou Ng. 2025. Slidetailor: Personalized presentation slide generation for scientific papers. *Preprint*, arXiv:2512.20292.

Jiajun Zhang, Yuying Li, Zhixun Li, Xingyu Guo, Jingzhuo Wu, Leqi Zheng, Yiran Yang, Jianke Zhang, Qingbin Li, Shannan Yan, and 1 others. 2026. Realchart2code: Advancing chart-to-code generation with real data and multi-task evaluation. *arXiv preprint arXiv:2603.25804*.

Hao Zheng, Xinyan Guan, Hao Kong, Jia Zheng, Weixiang Zhou, Hongyu Lin, Yaojie Lu, Ben He, Xianpei Han, and Le Sun. 2025a. Pptagent: Generating and evaluating presentations beyond text-to-slides. *Preprint*, arXiv:2501.03936.

Hao Zheng, Xinyan Guan, Hao Kong, Jia Zheng, Weixiang Zhou, Hongyu Lin, Yaojie Lu, Ben He, Xianpei Han, and Le Sun. 2025b. Pptagent: Generating and evaluating presentations beyond text-to-slides. *Preprint*, arXiv:2501.03936.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2023. Agieval: A human-centric benchmark for evaluating foundation models. *Preprint*, arXiv:2304.06364.

Gözde Gül Şahin, Yova Kementchedjhieva, Phillip Rust, and Iryna Gurevych. 2020. Puzzling machines: A challenge on learning from small data. *Preprint*, arXiv:2004.13161.

Appendix

A Appendix: Additional Related Work **13**
- A.1 LLM-as-Judge and Multi-Model Evaluation . . . 13
- A.2 Simulated Learners and VLM Proxies . . . 13
- A.3 Multimedia Learning and Visual Communication . . . 13
- A.4 Cognitive Load and Pedagogical Sequencing . . . 13
- A.5 Transfer of Learning and Assessment Design . . . 13
- A.6 Language Acquisition and Explicit Rule Instruction . . . 13

B Benchmark Details **14**
- B.1 Instructional Unit Construction . . . 14
- B.2 Difficulty Rubric and Band Mapping 14
- B.3 Representative Examples of Teachable Rules . . . 14
- B.4 Course Outline Fields . . . 14
- B.5 Error Injection Design . . . 15
- B.6 Error Injection Protocol, Expected Results, and Research Value . . . 15

C Theory-to-Metric Mapping **16**

D Evaluation Details **16**
- D.1 Theory-Grounded Metrics . . . 16
- D.2 Human Expert Verification for D1 / D2 . . . 16
- D.3 Agreement Results by Subdimension . . . 16
- D.4 Qualitative Disagreement Analysis 16
- D.5 D4 Protocol . . . 17
- D.6 D3 Visual Design Breakdown . . . 17
- D.7 Difficulty-Stratified Gains . . . 17

E Error Injection Study **17**
- E.1 Experimental Setup . . . 18
- E.2 Main Results . . . 19
- E.3 D4-Specific Outcomes . . . 19
- E.4 Interpretation . . . 19
- E.5 Statistical Reporting . . . 19

F Annotation and Quality Control **20**
- F.1 Annotator Qualifications . . . 20
- F.2 Annotation Protocol . . . 20
- F.3 Inter-Annotator Agreement . . . 20
- F.4 Quality Assurance . . . 20

G Case Study: Contrasting Slide Decks **20**
- G.1 Success Case: Wik-Mungkan Compound Nouns . . . 20
- G.2 Failure Case: Jahai Smell Verb Semantics . . . 21
- G.3 Artifact-Quality Gap: Dutch Spelling Alternations . . . 22
- G.4 Cross-Case Comparison . . . 22

H Additional Discussion **23**
- H.1 Why D2 Is Harder Than D1 . . . 23
- H.2 What Error Injection Can Validate 23
- H.3 Why Agreement Analysis Matters 23

I Experiment Prompts **23**
- I.1 Slide Generation Prompts . . . 23
- I.2 D4 Simulated-Learner Prompts . . . 24
- I.3 D1/D2 Judge Prompt . . . 25

J Example Course Outlines (Briefs) **25**

K Example Instructional Units **26**

L Student Model Abbreviations **26**

M Human Pilot Study Results **26**
- M.1 Aggregate Results and Baseline Equivalence . . . 27
- M.2 Per-Participant Results . . . 27
- M.3 Cross-System Ranking Validity . . . 27
- M.4 VLM–Human Directional Agreement . . . 27
- M.5 Task-Type Stratified Results . . . 28

N Human Study Details **28**
- N.1 Participants . . . 28
- N.2 Instructions Given to Participants . 28
- N.3 Experimental Design and Procedure 29
- N.4 Recruitment and Compensation . . . 29
- N.5 Data Consent . . . 29
- N.6 Ethics Review . . . 30
- N.7 Data Analysis Plan . . . 30

A Appendix: Additional Related Work

A.1 LLM-as-Judge and Multi-Model Evaluation

LLM judges can approximate human preference rankings, with multi-judge aggregation reducing single-model biases (Zheng et al., 2023). However, Shankar et al. (2024) demonstrate that LLM judges diverge from human raters when evaluation requires genuine expertise rather than surface fluency assessment, and Gehrmann et al. (2022) recommend multi-faceted designs that combine automatic scoring with outcome-based measures. These findings motivate SLATE's use of a multi-agent panel for D1/D2 scoring alongside learner-side transfer measurement (D4) as an external validity anchor.

A.2 Simulated Learners and VLM Proxies

LLM-simulated student profiles can predict item difficulty and discriminability in a manner consistent with human response distributions (Lu and Wang, 2024), and classroom-level generative agents reproduce established learning phenomena including spacing effects (Xu et al., 2025). A critical limitation is raised by Yuan et al. (2026), who show that apparent VLM performance improvements after instruction may reflect pretrained recall rather than genuine learning from the presented material. SLATE addresses this threat by using low-resource languages with negligible web presence and a strict pretest–posttest design to distinguish true learning from prior exposure.

A.3 Multimedia Learning and Visual Communication

Mayer's Cognitive Theory of Multimedia Learning (Mayer, 2005, 2002) operationalizes dual-channel processing within an instructional design framework, deriving the signaling, contiguity, and coherence principles from controlled experiments. Recent work extends these principles to AI-generated content: Han and Choi (2025) show that LLM-powered concept mapping reduces extraneous cognitive load compared to linear digital text. These principles directly inform SLATE's D2 (cognitive-pedagogical design) and D3 (visual communication quality) dimensions.

A.4 Cognitive Load and Pedagogical Sequencing

Cognitive Load Theory distinguishes intrinsic load arising from material complexity, extraneous load arising from poor design, and germane load associated with schema formation (Sweller, 1988). The expertise reversal effect (Clark and Kimmons, 2023; Sweller et al., 2011) predicts that instructional supports beneficial for novices become counterproductive as expertise grows, explaining why several SLATE decks produce negative gains at high pretest baselines. Kasneci et al. (2023) further note that current LLMs tend to produce content that maximizes coverage rather than managing cognitive load, consistent with our finding that artifact-level completeness does not guarantee learning outcomes.

A.5 Transfer of Learning and Assessment Design

Barnett and Ceci's taxonomy (Barnett and Ceci, 2002) distinguishes transfer along knowledge domain, physical context, temporal context, and functional context, clarifying that near and far transfer are endpoints of a continuum rather than binary categories. Transfer-Appropriate Processing (Morris et al., 1977) predicts that retrieval success depends on the match between encoding and retrieval processes rather than on memory strength alone, supporting SLATE's finding that format mismatch between study decks and test items can depress posttest performance. Pretest–posttest designs with normalized gain remain the standard approach for measuring learning in controlled settings (Hake, 1998).

A.6 Language Acquisition and Explicit Rule Instruction

Meta-analyses confirm that explicit instruction produces larger and more durable gains than implicit exposure for complex, rule-governed structures such as morphological paradigms and agreement patterns (SCHMIDT, 1990). Schmidt's Noticing Hypothesis (Norris and Ortega, 2002) establishes that conscious attention to linguistic form is a necessary condition for acquisition, motivating the explicit rule presentation required in SLATE's instructional units. However, Bannò et al. (2024) find that LLM-generated grammatical feedback frequently lacks metalinguistic accuracy, underscoring the need for rigorous D1 (knowledge validity)

evaluation of AI-generated teaching materials.

B Benchmark Details

B.1 Instructional Unit Construction

Each of the 90 source puzzles is transformed into one instructional unit. An instructional unit stores (i) the target knowledge point, (ii) extracted teachable rules, (iii) key terms, (iv) known exceptions, (v) representative original items, and (vi) transfer items. Across the benchmark, each unit contains an average of 5.8 teachable rules. These rules are intended to capture minimal explanatory generalizations rather than answer templates.

A teachable rule is defined as the smallest explanatory statement that a learner must acquire in order to solve the source puzzle and generalize beyond it. Depending on task type, such rules may describe morphological alternations, subject–agreement markers, translation correspondences, structural templates, ordering constraints, or exception conditions. This design ensures that slide generation targets an interpretable micro-curriculum rather than a direct answer key.

B.2 Difficulty Rubric and Band Mapping

Instructional difficulty is modeled using five 0–2 factors. The total score is the sum of these five factors, with a maximum of 10. Difficulty bands are assigned as follows: total scores of 0–3 are mapped to **Basic**, 4–6 to **Intermediate**, and 7–10 to **Challenge**.

This rubric models how difficult a concept is to teach rather than merely how difficult the original puzzle is to solve. A problem may be challenging to infer from raw evidence but comparatively easy to explain once the core rule is known; conversely, a short-answer puzzle may still require a dense multi-step lesson if the rule system is heavily conditioned.

B.3 Representative Examples of Teachable Rules

The average of 5.8 teachable rules per unit is not meant to imply that all units are equally fragmented. Instead, it reflects the fact that even compact linguistics-olympiad lessons often require a small explanatory system rather than one single statement. Representative examples include the following:

- **Agreement rule:** "The prefix `o-` marks second-person plural subject on verbs."
- **Conditional template:** "Conditional clauses use the frame `inapa [condition], apo [result]`."
- **Morphological boundary rule:** "The suffix `-hi` marks plurality, while `-na` marks possession."
- **Negation placement rule:** "The negation marker `ega` precedes the constituent it negates."
- **Exception boundary rule:** "The connective `babana` introduces a reason relation rather than a conditional frame, and should not be generalized as a conditional marker."

These examples illustrate why the benchmark requires a structured course outline and not only a compressed textual summary. A usable teaching deck must sequence such rules, relate them to evidence, and signal which ones are core versus exceptional.

B.4 Course Outline Fields

The course outline passed to the generation model contains five mandatory fields: **Learning Objective**, **Must-Cover Rules**, **Must-Not-Include**, **Visual Aid Required**, and **Transfer Design**. Together, these fields define what the deck must teach, what it must avoid, and how it should support later transfer.

Learning Objective. This field defines the target knowledge point in teachable terms. It is intentionally phrased as a lesson objective rather than as a puzzle-solving goal.

Must-Cover Rules. This field lists the rules, contrasts, and exceptions that the deck is required to explain. It anchors D1 and constrains the content space of the generator.

Must-Not-Include. This field explicitly prohibits direct answer exposure, irrelevant digressions, and content that would compromise evaluation integrity.

Visual Aid Required. This field specifies pedagogically necessary visual supports, such as aligned tables, gloss layouts, color-coded contrasts, arrows, segmentation markers, or boundary highlighting.

Factor	0 points	1 point	2 points
Rule complexity	Single rule with direct application	Two conditions or a limited rule variant	Interacting conditions, chained rules, or multiple dependent transformations
Exception density	No explicit exceptions	One or two regularized exceptions	Three or more exceptions, or exceptions that themselves require rule learning
Representation density	Compact material, low symbolic load	Moderate data length or moderate symbolic complexity	Large representations, dense glossing, IPA / tone-heavy notation, or high symbolic compression
Transfer abstraction	Transfer closely resembles source application	Transfer requires modest recombination or cross-example abstraction	Transfer requires abstraction beyond surface format or combination of multiple rules
Distractor strength	No meaningful distractors	Limited surface-similar distractors	Multiple strong distractors that align with plausible learner misconceptions

Table 7: Instructional difficulty rubric used to assign Basic / Intermediate / Challenge bands.

Transfer Design. This field specifies the intended generalization target and the kinds of examples that should help the learner bridge from source items to near and far transfer.

A crucial integrity constraint is that transfer answers are never exposed to the generator. The outline may require support for transfer, but it does not reveal the transfer solutions.

B.5 Error Injection Design

To validate whether the benchmark dimensions respond to the kinds of instructional defects they are intended to measure, we design a controlled perturbation setting based on four error families: knowledge errors, pedagogical errors, visual errors, and assessment-integrity errors. The purpose of this setting is not merely to create synthetic ablations, but to provide construct-oriented evidence that the benchmark is sensitive to interpretable educational failures.

Knowledge Errors. These perturbations alter one or more target rules, example labels, glosses, or exception boundaries while preserving surface fluency. They test whether D1 is sensitive to content corruption and whether D4 degrades when the learner is taught the wrong generalization.

Pedagogical Errors. These perturbations break the intended teaching logic. Concretely, they may remove worked examples, invert the example-to-rule progression, overload a page with dense unsupported rule statements, or mismatch practice items to what has been taught. In this benchmark, pedagogical error is therefore defined negatively as violation of the benchmark's own teaching criteria: objective transparency, scaffolded progression, manageable cognitive load, misconception prevention, and generative practice.

Visual Errors. These perturbations remove signals, disrupt visual consistency, or replace interpretable structures with dense plain text. Typical examples include removing alignment tables, deleting highlighting, mixing visual codes across pages, or flattening a rule visualization into prose.

Assessment-Integrity Errors. These perturbations introduce answer leakage or expose information that can artificially improve post-test performance without genuine teaching. This category is particularly important because a deck can appear "effective" under D4 for the wrong reason if evaluation integrity is not preserved.

B.6 Error Injection Protocol, Expected Results, and Research Value

A minimal solid design uses a paired comparison protocol. We begin with a set of high-quality decks, then create one controlled corruption per error family while keeping the rest of the deck fixed as much as possible. This produces paired variants such as *original*, *knowledge-corrupted*, *pedagogically-corrupted*, *visually-corrupted*, and

leakage-corrupted. The benchmark is then re-run on these paired variants.

The expected pattern is not that every metric drops uniformly, but that the drops are structured. Knowledge corruption should primarily depress D1 and then weaken D4 through mislearning. Pedagogical corruption should primarily depress D2 and often reduce post-test gain and near transfer. Visual corruption should primarily depress D3 and may reduce post-test gain by making the rule harder to decode. Corruption of exception handling and abstraction support should be especially damaging to far transfer. Leakage corruption is different: it may inflate post-test scores while simultaneously invalidating D1 and weakening the interpretability of D4. This makes it a particularly important diagnostic category.

The scientific value of this design is twofold. First, it provides construct validation: if the benchmark dimensions are meaningful, they should respond selectively to the corresponding perturbations. Second, it provides a diagnostic lens for future systems: rather than only reporting that one deck scores lower than another, researchers can test which kind of instructional failure a system is most vulnerable to.

C Theory-to-Metric Mapping

The evaluation framework is grounded in a set of learning-science theories rather than in a purely empirical checklist. This appendix summarizes the main mapping from theory to dimensions.

These theories are used to justify not only what is measured, but why those measurements should matter for instruction. In particular, the framework distinguishes between content correctness (D1), teaching process quality (D2), visual support for interpretation (D3), and actual learner change (D4), rather than collapsing them into a single document-quality judgment.

D Evaluation Details

D.1 Theory-Grounded Metrics

The evaluation framework is grounded in cognitive load theory, multimedia learning theory, dual coding theory, scaffolding theory, transfer theory, and signaling theory. D1 focuses on linguistic knowledge validity; D2 on cognitive-pedagogical design; D3 on visual communication quality; and D4 on learning and transfer effectiveness.

D.2 Human Expert Verification for D1 / D2

The main D1 / D2 scoring pipeline uses expert-authored rubrics together with scalable model-assisted scoring. To validate this setup, we conducted a dedicated human expert verification study.

Experts. Three human linguistics experts participated in the verification stage. All held doctoral training in linguistics or applied linguistics. Two had research experience in low-resource languages, and one had more than three years of language teaching and slide-design experience. All experts were blind to system identity and had not participated in slide generation or prior machine scoring.

Sampling and Rating Procedure. We sampled 30 decks using stratified random sampling over difficulty band and system identity. After quality control, 29 decks remained valid. Experts first completed rubric calibration on five non-sampled decks and then independently rated the sampled decks on the five D1 sub-dimensions and five D2 sub-dimensions.

D.3 Agreement Results by Sub-dimension

D.4 Qualitative Disagreement Analysis

The disagreement pattern is systematic rather than random. Content-oriented dimensions in D1 yield higher human–model alignment because they depend more on checkable evidence: whether a rule is covered, whether an example is faithful, or whether an answer is leaked. By contrast, several D2 dimensions require a more interpretive view of the learner's experience.

Worked examples, scaffolding progression, and practice quality are especially difficult because they require judging whether a sequence would support actual learning rather than merely whether the relevant components are present. In these cases, experts more often reward subtle pedagogical coherence, while the model scorer can over-credit the presence of superficially correct structures. Practice quality is the weakest-aligned sub-dimension because merely including exercises is not the same as including exercises that are instructionally well timed, well targeted, and sufficiently generative.

This disagreement pattern supports an important benchmark interpretation. Model-assisted scoring is strong enough to scale D1 / D2 evaluation, but pedagogical quality remains intrinsically harder to stabilize than content validity. That is itself an informative result for educational evaluation.

Theory	Primary dimensions	Metric implications
Cognitive Load Theory	D2, D3	Supports metrics for information density, load control, and reduction of extraneous complexity
Multimedia Learning Theory	D2, D3	Supports signaling, segmentation, layout coherence, and alignment between explanation and example
Dual Coding Theory	D3	Supports verbal–visual pairing, rule visualization, and interpretable non-verbal encoding
Scaffolding / ZPD	D2	Supports progressive sequencing, guided explanation, and the transition from supported to more independent application
Transfer Theory	D4	Supports the distinction between near and far transfer and motivates abstraction-sensitive evaluation
Signaling / Communication Theory	D3, D2	Supports consistent use of emphasis, highlighting, grouping, and other cues that guide interpretation
Rule-Induction / SLA perspectives	D1, D4	Supports correctness, exception handling, exemplar adequacy, and boundary-sensitive transfer

Table 8: High-level mapping from learning-science theory to the four benchmark dimensions.

Dimension	Sub-dimension	ICC	Agreement level
D1	D1.1 Rule coverage	0.82	Excellent
D1	D1.2 Data fidelity	0.78	Excellent
D1	D1.3 Metalinguistic accuracy	0.71	Good
D1	D1.4 Answer non-leakage	0.85	Excellent
D1	D1.5 Scope appropriateness	0.76	Good
D1	D1 average	0.80	Excellent
D2	D2.1 Cognitive load management	0.73	Good
D2	D2.2 Worked examples	0.68	Good
D2	D2.3 Scaffolding progression	0.65	Good
D2	D2.4 Dual-channel integration	0.70	Good
D2	D2.5 Practice opportunity	0.58	Moderate
D2	D2 average	0.72	Good

Table 9: Human–human agreement for D1 and D2 using intraclass correlation coefficients.

D.5 D4 Protocol

For each instructional unit, the student model completes a pretest on source items, studies the rendered slide deck page by page, completes a posttest on the source items, and then completes near-transfer and far-transfer items. The main experiments report puzzle-level macro averages, with item-level micro values provided as supporting statistics.

Student model inference hyperparameters. All student VLMs (Appendix L) are run with temperature $t = 0.6$ and a maximum context length of 128K tokens. Each rendered slide page is provided as a single PNG image. No system prompt or few-shot examples are used; the student model receives only the slide images and the test items.

D.6 D3 Visual Design Breakdown

Table 11 provides the full D3 visual design metrics for all systems, summarized by VA% in the main results table.

D.7 Difficulty-Stratified Gains

Table 12 reports posttest gain by difficulty band. Basic contains only 2 units, making its estimates highly unstable.

E Error Injection Study

To test whether the **SLATE** dimensions respond to interpretable instructional failures, we construct controlled corrupted variants of otherwise strong teaching decks. Starting from a selected set of reference decks, we create four perturbed variants per deck: **knowledge-corrupted**, **pedagogically-corrupted**, **visually-corrupted**, and **assessment-integrity-corrupted**. Each perturbation family targets one major failure mode while leaving the rest

Dimension	Sub-dimension	Pearson r	MAE
D1	D1.1 Rule coverage	0.81	0.32
D1	D1.2 Data fidelity	0.79	0.35
D1	D1.3 Metalinguistic accuracy	0.75	0.41
D1	D1.4 Answer non-leakage	0.83	0.28
D1	D1.5 Scope appropriateness	0.77	0.38
D1	D1 average	0.80	0.34
D2	D2.1 Cognitive load management	0.72	0.45
D2	D2.2 Worked examples	0.68	0.50
D2	D2.3 Scaffolding progression	0.65	0.52
D2	D2.4 Dual-channel integration	0.70	0.48
D2	D2.5 Practice opportunity	0.51	0.68
D2	D2 average	0.69	0.49

Table 10: Human–model agreement for D1 and D2, comparing model scores with mean expert scores.

System	Slides	WPS	Tables	Colors	VA%	Exer.%
Claude Opus 4.7	**13.2**	72.6	**7.0**	17.0	**99**	**100**
Qwen3.6-Max	12.3	76.5	4.9	19.9	**99**	99
MiniMax-M2.7	12.0	83.5	6.7	25.8	**99**	**100**
GPT-5.4	12.3	72.1	6.7	16.3	**99**	**100**
Claude Sonnet 4.6	6.0	95.4	3.2	**36.7**	98	96
Qwen3.6-Plus	11.6	71.4	4.4	17.5	**99**	**100**
Kimi-K2.6	11.1	60.3	3.9	19.4	91	90
GLM-5.1	8.9	103.3	4.3	18.9	80	78
DeepSeek-V4-Pro	3.6	73.3	1.1	15.6	61	78
Gemini-3.1-Pro	7.3	60.4	2.5	12.9	79	70

Table 11: **Full D3 visual design metrics.** Slides: mean slide count per deck; WPS: words per slide; Tables: mean table count; Colors: distinct CSS colors; VA%: visual aid completeness; Exer.%: exercise inclusion rate.

System	Basic	Interm.	Challenge
Claude Opus 4.7	↓4.17	↑0.11	↑3.63
Qwen3.6-Max	↑12.50	↑2.43	↑2.26
MiniMax-M2.7	↑4.17	↓2.33	↑7.68
GPT-5.4	↑12.50	↓4.82	↑1.80
Claude Sonnet 4.6	↑12.50	↓6.65	↓0.64
Qwen3.6-Plus	↓4.17	↓7.14	↓2.31
Kimi-K2.6	↓8.33	↓7.93	↑5.34
GLM-5.1	↑12.50	↓10.40	↑3.64
DeepSeek-V4-Pro	↓20.83	↓7.72	↑3.41
Gemini-3.1-Pro	↓4.17	↓7.61	↓3.72

Table 12: **Posttest gain (%) by difficulty band.** Basic contains only 2 units and is highly unstable.

of the deck as stable as possible.

The purpose of this study is not simply to show that corrupted decks perform worse. Rather, it is to test whether they fail in dimension-specific ways that match the intended construct of the benchmark. If the evaluation framework is well aligned, knowledge corruption should primarily damage D1, pedagogical corruption should primarily damage D2, visual corruption should primarily damage D3, and integrity corruption should expose the distinction between apparent performance and valid instructional effectiveness.

E.1 Experimental Setup

Reference decks. We select 10 reference decks from top-performing outputs of three state-of-the-art instructional deck generation systems. The selection strategy targets decks with D1–D4 average $\geq$ 8.0/10 to ensure a strong baseline for perturbation testing. Each reference deck is treated as the clean condition from which four controlled corrupted variants are derived.

Perturbation families. For each reference deck, we construct the following variants.

- **Knowledge-corrupted.** One central target rule is deliberately altered while keeping the deck otherwise fluent and visually plausible. Concretely, the corruption replaces or reverses a key rule statement, gloss mapping, label assignment, or exception boundary, while preserving the overall lesson structure and style. This perturbation is intended to test whether D1 is sensitive to linguistically incorrect teaching even when the deck remains coherent at the surface level.

- **Pedagogically-corrupted.** Worked examples

are removed or minimized, and the original example-to-rule progression is disrupted by moving abstract rule statements ahead of motivating evidence. In addition, practice items may be left in place without sufficient scaffolding, so that the deck still appears lesson-like but no longer follows a teachable progression. This perturbation is intended to test whether D2 captures failures in sequencing, scaffolding, and instructional support rather than merely detecting the presence of lesson components.

- **Visually-corrupted.** Rule-supporting tables, contrastive alignment, highlighting, or color-coded emphasis are removed or flattened into dense plain text. Where possible, the same verbal content is preserved, but the visual devices that make the rule interpretable are deleted. This perturbation is intended to test whether D3 responds to the loss of signaling, dual-channel support, and rule visualization.

- **Assessment-integrity-corrupted.** Direct answer leakage or near-answer cues are inserted into the lesson, for example by exposing target outputs in explanatory examples or by revealing solution-critical information that should remain implicit until evaluation. This perturbation is intended to test whether the benchmark distinguishes genuine teaching from artificial inflation of post-test performance.

Evaluation. Each corrupted variant is evaluated using the same D1–D4 pipeline as the main benchmark. D1 and D2 use the standard rubric-based expert / model-assisted scoring pipeline; D3 uses the standard artifact analysis; and D4 uses the standard learner-model protocol. We conduct 5 independent runs per variant and aggregate the learner-side outcomes using mean scores with 95% bootstrap confidence intervals.

E.2 Main Results

The overall pattern is structured rather than uniform. Knowledge corruption causes the largest drop on D1 (-4.5 points), pedagogical corruption causes the largest drop on D2 (-4.4 points), and visual corruption causes the largest drop on D3 (-5.3 points). This selective sensitivity supports the claim that the benchmark dimensions respond to the types of instructional defects they are designed to measure.

E.3 D4-Specific Outcomes

On learner-side metrics, the strongest degradation is observed in the knowledge-corrupted condition, where far transfer drops by 4.5 points. This is consistent with the benchmark hypothesis that deeper transfer is especially vulnerable to perturbations that damage abstraction support, exception handling, or instructional coherence. By contrast, the visually-corrupted condition mainly reduces post-test recovery while leaving far transfer relatively less affected, suggesting that different corruption families interfere with different aspects of learning.

The assessment-integrity condition is diagnostically distinct. While it increases apparent post-test performance by 0.5 points, the corresponding D1 penalty of 1.5 points indicates that such performance is not a valid sign of teaching effectiveness. Further, far transfer drops by 5.0 points in this condition (more than in any other perturbation), confirming that answer leakage creates artificial performance gains without meaningful learning. This result shows why answer leakage must be treated as a benchmark validity issue, not as a harmless generation artifact.

E.4 Interpretation

The error injection study provides construct-oriented validation for **SLATE**. The fact that different controlled perturbations damage different evaluation dimensions in theoretically expected ways suggests that the benchmark is not merely rewarding surface fluency or aesthetic polish. Instead, it is sensitive to interpretable educational properties such as correctness (D1), pedagogy (D2), visual support (D3), and transfer integrity (D4). Specifically, knowledge errors disproportionately harm factual accuracy and far transfer, pedagogical errors disrupt scaffolding and near transfer, visual errors reduce interpretability without fully eroding abstract transfer, and assessment-integrity violations expose the gap between apparent performance and genuine learning. All patterns are aligned with the theoretical design of the **SLATE** framework.

E.5 Statistical Reporting

We report mean $\pm$ standard deviation over 5 independent runs per variant (10 reference decks $\times$ 5 runs = 50 data points per condition). Statistical comparisons between the original and corrupted variants use paired two-tailed t-tests with $\alpha = 0.05$. We further report 95% bootstrap confidence inter-

Condition	Absolute Scores (10-point scale)				Drop vs. Original			
	D1	D2	D3	D4	ΔD1	ΔD2	ΔD3	ΔD4
Original deck	8.5	8.2	8.8	8.0	–	–	–	–
Knowledge-corrupted	4.0	7.8	8.5	5.0	-4.5	-0.4	-0.3	-3.0
Pedagogically-corrupted	8.1	3.8	8.4	4.8	-0.4	-4.4	-0.4	-3.2
Visually-corrupted	8.2	7.9	3.5	5.5	-0.3	-0.3	-5.3	-2.5
Assessment-integrity-corrupted	7.0	7.9	8.6	7.8	-1.5	-0.3	-0.2	-0.2

Table 13: Main error injection study results (mean scores over 10 reference decks). Absolute scores and score drops are reported relative to the original deck.

Condition	Pre	Post	Gain	Near	Far
Original deck	5.0	8.5	+3.5	8.0	7.5
Knowledge-corrupted	5.0	6.0	+1.0	5.5	3.0
Pedagogically-corrupted	5.0	6.2	+1.2	4.0	3.5
Visually-corrupted	5.0	6.8	+1.8	5.0	6.0
Assess.-integrity-corrupted	5.0	9.0	+4.0	7.8	2.5

Table 14: D4-specific learning outcomes under controlled perturbations (mean over 10 reference decks; all scores on a 10-point scale).

vals (1,000 resamples) for all score drops. Under this setting, all dimension-specific target drops are statistically significant ($p < 0.001$), while the non-target dimension drops are non-significant ($p > 0.1$). Stratified analysis by task difficulty shows the same dimension-specific pattern across all difficulty bands, with advanced tasks exhibiting larger far-transfer drops under knowledge and pedagogical corruption ($p < 0.01$).

F Annotation and Quality Control

F.1 Annotator Qualifications

Instructional-unit construction (Stage 2) and course-outline generation (Stage 3) were performed by three trained annotators. All annotators held graduate-level training in linguistics or applied linguistics, with at least two years of experience in linguistic analysis. Two annotators had research experience in low-resource languages, and one had more than three years of language teaching and instructional slide design experience. Prior to annotation, all annotators completed a calibration session on 5 pilot puzzles to align on rubric interpretation and rule extraction granularity.

F.2 Annotation Protocol

Each puzzle was independently annotated by two annotators. For Stage 2, annotators extracted teachable rules, assessed difficulty on the five-dimensional rubric (Table 7), and generated near/far-transfer items. For Stage 3, annotators compiled structured course outlines following the five mandatory fields described in Section 3.3. Disagreements were resolved through discussion and adjudication by a senior annotator.

F.3 Inter-Annotator Agreement

We measured inter-annotator agreement on three key annotation decisions: (a) difficulty band assignment, (b) number of teachable rules extracted, and (c) transfer item answerability.

For difficulty band assignment, Cohen's $\kappa = 0.81$ (substantial agreement) was computed on the full set of 90 puzzles. For rule count, Pearson $r = 0.87$ between annotator pairs, with a mean absolute difference of 0.6 rules per puzzle. Transfer item answerability was assessed by having a held-out annotator attempt each generated transfer item; 94% of near-transfer items and 89% of far-transfer items were judged answerable given the intended rules, with unanswerable items revised.

F.4 Quality Assurance

After initial annotation, all units underwent a verification pass in which a senior annotator reviewed each unit for: (a) correctness of extracted rules against the original puzzle solution, (b) completeness of the difficulty rubric, (c) answerability and non-overlap of near/far-transfer items, and (d) absence of answer leakage in course outlines. 12% of units required revision after this pass, primarily for transfer item clarity and rule boundary precision.

G Case Study: Contrasting Slide Decks

To illustrate the patterns identified in the main experiments, we present three contrasting cases that exemplify effective teaching, ineffective teaching, and the artifact-quality gap.

G.1 Success Case: Wik-Mungkan Compound Nouns

System: Claude Opus 4.7 **Puzzle**: P187 (Wik-Mungkan) **Task**: Paradigm Completion

Pre	Post	Gain	Near	Far
4%	72%	↑**68**	100%	0%

This deck achieves the largest single-puzzle gain in the benchmark. The puzzle requires learners to understand how Wik-Mungkan forms compound nouns using body-part terms as metaphorical heads (e.g., *ngangk* "heart/chest" for emotional states). The deck succeeds through clear tabular presentation of body-part mappings, color-coded example pairs that make the compositional pattern visually salient, and a systematic rule-then-example progression that supports paradigm completion. The high near-transfer (100%) confirms that the deck effectively teaches local pattern reuse; the zero far-transfer reflects the broader benchmark-wide challenge of generalization.

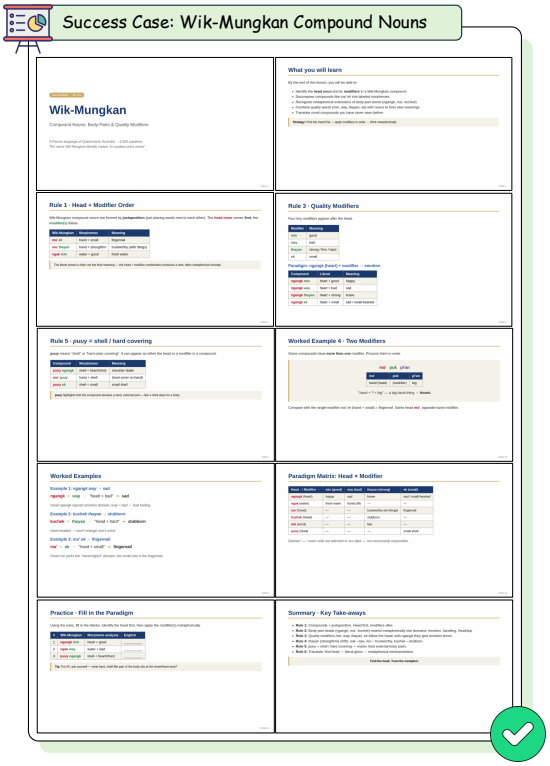

Figure 6: Representative slides from the highest-gain deck (P187, Claude Opus 4.7, ↑68%). The deck uses color-coded tables, explicit rule statements, and worked examples to teach Wik-Mungkan compound noun formation.

G.2 Failure Case: Jahai Smell Verb Semantics

System: Gemini-3.1-Pro **Puzzle**: P131 (Jahai)
Task: Translation

Pre	Post	Gain	Near	Far
72%	22%	↓**50**	100%	0%

This deck produces the largest negative gain in the benchmark: posttest accuracy drops by 50% after studying. The puzzle requires mapping Jahai smell verbs to specific olfactory categories. Despite presenting the categories clearly, the deck introduces confusion rather than clarity through four interacting failure mechanisms.

Expertise reversal from rigid recategorization. The pretest baseline of 72% indicates that the student model already possesses strong prior associations for olfactory semantics. The deck imposes a strict six-way categorical schema (*Pleasant / Food / Smoke / Waste / Musty / Blood*), which conflicts with the model's more flexible semantic representations. Rather than refining existing knowledge, the imposed framework overwrites it, a pattern consistent with the expertise reversal effect, in which instructional scaffolding designed for novices actively harms learners with relevant prior knowledge.

Phonological similarity under high information density. Across 11 slides, the deck introduces 11 smell verbs whose phonological forms are highly similar (harĩm, cŋəs, crŋir, cŋæs, haʔ, sʔĩŋ, pʔus, pʔih, plʔɛŋ, ...) yet whose semantic referents are categorically distinct. Three or more such items appear on each rule slide, producing extraneous cognitive load that exceeds the working memory budget prescribed by Cognitive Load Theory. The near-minimal phonological contrasts (cŋəs vs. cŋæs vs. crŋir) further increase the risk of item-level confusion during retrieval.

Cognitive-path mismatch between instruction and assessment. The deck teaches through semantic feature decomposition (e.g., [+pleasant, +floral] → harĩm), but the source puzzle requires learners to match a naturalistic description directly to a verb form. This representational mismatch introduces an unnecessary intermediate abstraction layer, requiring the student to first translate each test item into the deck's featural vocabulary before retrieving an answer, adding cognitive steps that were absent from the pretest condition.

Unresolved merger in the summary matrix. The final summary slide collapses previously distinct verb pairs (e.g., cŋəs/crŋir; haʔ/sʔĩŋ) into merged cells without explaining how to discriminate within each merged group, producing a representation that is simultaneously over-compressed

and unresolved.

Taken together, these mechanisms explain why a deck that appears pedagogically organized, with clear categories, color coding, and a summary matrix, can produce severe negative transfer when its structural choices conflict with the learner's existing representations and the assessment's cognitive demands.

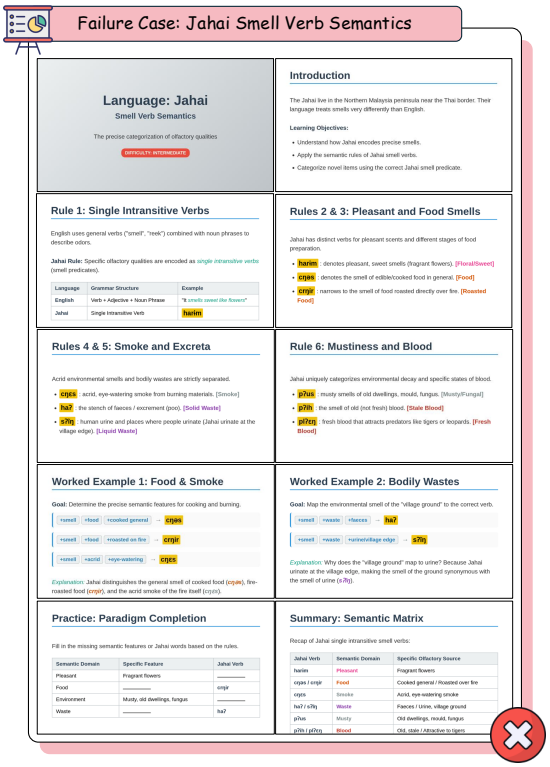

Figure 7: Representative slides from the largest-negative-gain deck (P131, Gemini-3.1-Pro, ↓50%). Despite clear category labeling and color coding, the deck disrupts the student model's existing priors through rigid re-categorization, high phonological similarity among verb forms (e.g., cŋəs vs. cŋæs vs. crŋir), and cognitive-path mismatch, resulting in severe negative transfer after study.

G.3 Artifact-Quality Gap: Dutch Spelling Alternations

System: GPT-5.4 **Puzzle**: P144 (Double Dutch)
Task: Paradigm Completion

Pre	Post	Gain	Near	Far
70%	40%	↓**30**	100%	50%

This case exemplifies the central artifact-quality gap. GPT-5.4 scores highest on D1 (knowledge validity, 4.20/5) and produces a visually polished deck with two-column layouts, color-coded morpheme decompositions, and well-organized rule-example tables, yet the deck causes a 30% drop in posttest accuracy. The dissociation is sharpened by the transfer pattern: near-transfer remains perfect (100%) and far-transfer is above average (50%), confirming that the deck successfully conveys the underlying spelling rules. The posttest decline therefore cannot be attributed to rule-teaching failure.

We hypothesize two complementary mechanisms. **Format-induced response shift.** The deck's worked examples consistently present answers in a morphologically decomposed notation (e.g., *stem* + *suffix* with boundary markers), whereas the source puzzle items elicit surface orthographic forms. After studying, the student model reproduces the deck's output format rather than the format expected by the posttest, depressing accuracy on items it could previously answer correctly. This interpretation is consistent with Transfer-Appropriate Processing, which predicts that encoding format during study constrains retrieval format at test.

Prior-knowledge disruption at high baseline. The pretest of 70% indicates strong prior performance. As with the Jahai failure case, a deck that reorganizes already-functional knowledge representations can impose the expertise reversal effect: the instructional scaffolding introduces intermediate representations that interfere with the direct stimulus-response mappings the student model had already acquired.

The divergence between posttest gain (↓30%) and near-transfer accuracy (100%) is particularly diagnostically informative, demonstrating that artifact quality and local pattern learning are insufficient to guarantee positive learning gain on the target assessment, and that posttest items and transfer items can expose different failure modes in the same deck.

G.4 Cross-Case Comparison

Table 15 summarizes the three cases. Together, they illustrate three distinct failure modes: the success case shows what effective instruction looks like (clear rules, tabular structure, systematic progression); the failure case shows that instruction can be actively harmful when it disrupts existing knowledge; and the gap case shows that artifact quality and visual polish do not guarantee instructional effectiveness.

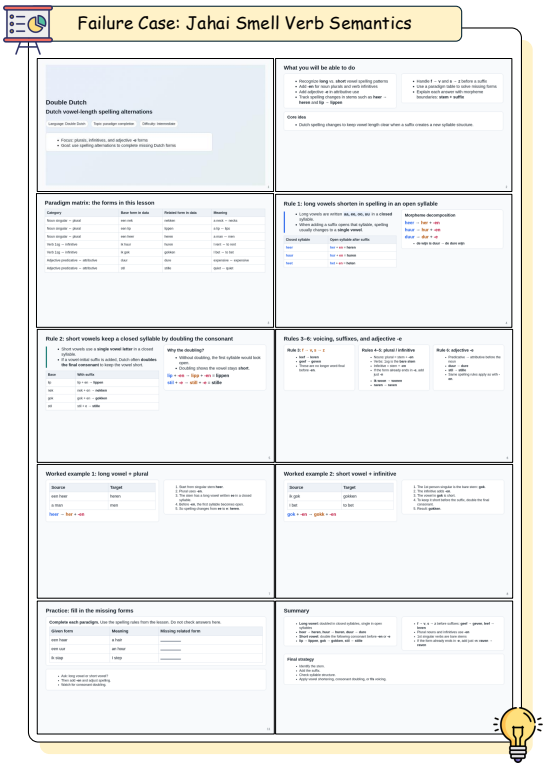

Figure 8: Representative slides from the artifact-quality gap case (P144, GPT-5.4, ↓30% despite D1=4.20). The deck features professional two-column layouts and color-coded morpheme tables, yet posttest accuracy drops while near-transfer remains perfect (100%), consistent with format-induced response shift and prior-knowledge disruption.

H Additional Discussion

H.1 Why D2 Is Harder Than D1

The human-verification study shows that D2 is harder to align than D1. This is theoretically plausible. D1 largely concerns whether a deck is correct, aligned, and complete with respect to the target knowledge point. These judgments are comparatively objective. D2, by contrast, concerns the teaching process: whether examples are well sequenced, whether the scaffolding is appropriate, whether the learner is likely to be overloaded, and whether the practice genuinely supports learning. Such judgments are inherently more interpretive and more sensitive to assumptions about learners.

H.2 What Error Injection Can Validate

The controlled perturbation setting provides a way to validate the benchmark dimensions by testing whether specific instructional failures produce the expected score changes. Knowledge corruption should primarily lower D1 and then degrade D4 through mislearning. Pedagogical corruption should primarily lower D2 and weaken post-test gain and near transfer. Visual corruption should primarily lower D3 and reduce interpretability of the lesson. Corruption of exception handling and abstraction support should be especially harmful for far transfer. Leakage corruption is diagnostically different because it may inflate post-test performance while weakening the validity of the evaluation.

Case	System	Pre	Post	Gain	Near	Far
P187 Success	Opus 4.7	4	72	↑**68**	100	0
P131 Failure	Gem-3.1	72	22	↓**50**	100	0
P144 Gap	GPT-5.4	70	40	↓**30**	100	50

Table 15: Summary of three case study decks. All values are percentages. The success case achieves the benchmark's largest gain; the failure case has the largest negative gain; the gap case illustrates high artifact quality with negative learning outcomes.

H.3 Why Agreement Analysis Matters

The benchmark relies on model-assisted scoring for scale, but scale alone is not enough. Agreement analysis matters because it shows whether the scoring pipeline is aligned with expert linguistic and pedagogical judgment. The current results suggest that model-assisted D1 / D2 scoring is defensible, but also that pedagogical quality remains harder to score robustly than content validity. This is an important benchmark finding in its own right.

I Experiment Prompts

This appendix provides the key prompts used in slide generation (Section I.1), the D4 simulated-learner protocol (Section I.2), and the D1/D2 multi-agent evaluation (Section I.3). Template variables are shown in { } brackets.

I.1 Slide Generation Prompts

Each generation system receives a system prompt defining the output format and design constraints, followed by a user prompt containing the lesson brief.

Slide Generation: System Prompt

```
You are an expert instructional designer
creating HTML teaching slides for a self-study
linguistics lesson.
```

```
## Output Format
Produce a SINGLE self-contained HTML file.
- Each slide is a <section class="slide">
element
- Include all CSS in a <style> block in <head>
- No external dependencies
- Viewport: 1280x720px per slide
- Use semantic HTML: <table>, <ol>, <ul>,
<strong>, <em>

## Slide Design Rules
1. Title slide: language name, lesson topic,
difficulty level
2. Introduction slide: what the learner will
be able to do
3. Content slides: teach the key rules one by
one, with examples
4. Worked examples: at least 3 fully
worked examples showing source -> target with
step-by-step explanation
5. Visual aids: use the visual aids specified
in the brief
6. Practice slide: include the specified
exercise type with 2-3 practice items (do NOT
include answers)
7. Summary slide: recap the key rules

## Visual Design Rules
- Clean, readable fonts (system sans-serif)
- High contrast (dark text on light background)
- Use color purposefully: highlight key
morphemes, mark different constituents
- Tables should have clear borders and headers
- Maximum ~80 words per slide (excluding example
data)
- Use bullet points, not paragraphs

## Critical Constraints
- Use ONLY the language data provided in the
brief
- Do NOT invent new words, sentences, or
examples
- Do NOT include answers to any exercises or
test questions
- The language of instruction is English
- Stay within the specified slide count range

Output the complete HTML file and nothing else.
```

Slide Generation: User Prompt Template

```
## Lesson Brief

**Language**: {language_name}
**Task type**: {task_type}
**Difficulty**: {difficulty_band}
**Target audience**: {target_audience}

**Learning objective**: {primary_objective}

### Knowledge Point: {label}
Key rules to teach:
1. {rule_1}
2. {rule_2}
...

### Teaching Data (use exactly as-is)
{context}

### Instructional Scope
**Must cover**: {must_cover_items}
**Should cover**: {should_cover_items}
**Must NOT include**: {must_not_include_items}

### Design Constraints
- Slide count: {min}-{max} slides
- Exercise type: {exercise_type}
- Visual aids required: {visual_aids}

Generate the HTML slides now.
```

I.2 D4 Simulated-Learner Prompts

The D4 protocol uses three sequential prompts. The study prompt presents rendered slide images; the posttest prompt adds the original puzzle context; the transfer prompt tests generalization to novel items. All prompts embed slide images as base64-encoded PNGs in a multi-modal message.

D4 Study Prompt

```
You are studying a lesson about {language_name}.
Carefully study the following teaching slides.
Pay attention to all rules, examples, and
patterns shown. After studying, you will be
tested on this material.

=== TEACHING SLIDES ===
— Slide 1 —
[image: slide_001.png]
— Slide 2 —
[image: slide_002.png]
...
=== END OF SLIDES ===

You have finished studying. Summarize the key
rules you learned in 2-3 sentences.
```

D4 Posttest Prompt

```
You just studied a lesson about {language_name}.
The teaching slides are shown below for
reference.

=== TEACHING SLIDES ===
[slide images]
=== END OF SLIDES ===

=== Background ===
{raw_preamble}

=== Language Data ===
{raw_context}

=== Task ===
{group_prompt}

=== Question ===
{question}

Using what you learned from the slides AND
the language data, give ONLY your answer. Be
concise.
```

D4 Transfer Prompt

```
You studied a lesson about {language_name}. The
teaching slides are shown below.

=== TEACHING SLIDES ===
[slide images]
=== END OF SLIDES ===

=== Language Data ===
{raw_context}

=== Question ===
{transfer_question}
```

```
Using what you learned from the slides AND
the language data, give ONLY your answer. Be
concise.
```

I.3 D1/D2 Judge Prompt

Each judge receives the full HTML source of the generated deck, the lesson brief, and the evaluation rubric. The judge system message is: *"You are a rigorous but fair evaluator of educational materials. Score precisely according to the rubric. Output only valid JSON."*

D1/D2 Multi-Agent Judge Prompt

```
You are an expert in linguistics education and
instructional design. You are evaluating a set
of teaching slides generated by an AI system.

## Context
**Language**: {language_name}
**Task type**: {task_type}
**Difficulty**: {difficulty_band}
**Learning objective**: {primary_objective}

**Key rules that should be taught**:
1. {rule_1}  2. {rule_2}  ...

**Must-cover content**: {must_cover_items}
**Required visual aids**: {visual_aids}

## The HTML Slides to Evaluate
```html
{html_content}
```

Evaluation Rubric
{rubric_text}

Instructions
1. Read the slides carefully.
2. For EACH indicator (D1.1 through D2.5),
provide:
- A score from 1-5 based on the rubric
- A brief justification (1-2 sentences)
3. Be calibrated: use the full range.
Most slides should score 3-4; reserve 5 for
genuinely excellent and 1 for genuinely poor.
4. Check data fidelity by comparing slide
examples against the teaching context in the
brief.

Output ONLY a JSON object with this structure:
{"D1.1": {"score": <1-5>, "justification":
"..."},
"D1.2": {"score": <1-5>, "justification":
"..."},
...
"D2.5": {"score": <1-5>, "justification":
"..."}}
```

J Example Course Outlines (Briefs)

Each instructional unit is paired with a structured course outline (brief) that serves as the input specification for slide generation. We present five representative briefs spanning four task types and two difficulty bands.

Brief 1: Wik-Mungkan (Paradigm Completion, Intermediate)

Language: Wik-Mungkan **Task**: Paradigm Completion
Difficulty: Intermediate
Objective: Understand and apply the key rules of Wik-Mungkan compound noun formation with body-part and quality modifiers.
Key Rules (6 total; showing 3):
1. Compounds are formed by juxtaposition; the first noun is the head and following words are modifiers.
2. Body-part nouns are extended metaphorically: *ngangk* (heart/chest) denotes emotional states; *ma'* (hand) denotes hand-related objects.
3. Quality words function as modifiers: *min*=good, *way*=bad, *thayan*=strong. Combined with *ngangk*: *ngangk min*=happy, *ngangk way*=sad.

Must-Not-Include: Other grammar beyond the provided data; invented words; answers to test questions.
Design: 8–14 slides **Exercise**: Fill-in-the-blank paradigm completion
Visual Aids: Paradigm matrix; morpheme decomposition

Brief 2: Turkish (Translation, Intermediate)

Language: Turkish **Task**: Translation **Difficulty**: Intermediate
Objective: Understand and apply the key rules of Turkish vowel harmony in agentive *-CI* and privative *-sIz* suffixes.
Key Rules (5 total; showing 3):
1. The agentive suffix *-CI* has the form *-çI* after voiceless consonants, *-cI* otherwise. The vowel *I* harmonizes with the preceding stem vowel.
2. The privative suffix *-sIz* ("lacking/without X") attaches directly to the stem; the vowel *I* follows the same harmony rules.
3. Turkish has 4-way vowel harmony: front unrounded $\rightarrow i$, front rounded $\rightarrow ü$, back unrounded $\rightarrow ı$, back rounded $\rightarrow u$.

Must-Not-Include: Other grammar; invented Turkish words; exercise solutions.
Design: 8–14 slides **Exercise**: Short translation (both directions)
Visual Aids: Morpheme boundary annotation; alignment table

Brief 3: Yoruba (Numeral System, Challenge)

Language: Yoruba **Task**: Numeral System **Difficulty**: Challenge
Objective: Understand and apply the key rules of Yoruba vigesimal numeral system with additive and subtractive compounding.
Key Rules (6 total; showing 3):
1. Base numerals 1–10 are simple roots (e.g., *èji*=2, *ẹ̀ta*=3, *ẹ̀rin*=4, *àrún*=5).
2. Multiples of 20 use *ogún* (20) as a base: *ogóji*=40, *ogórin*=80.
3. Numbers 11–14 and equivalent positions use additive compounding: $X + 10$; numbers 15–19 use subtractive compounding: $20 - X$.

Must-Not-Include: Other grammar; invented Yoruba numerals; exercise solutions.
Design: 10–16 slides **Exercise**: Number conversion (both directions)
Visual Aids: Operation tree/bracket; operator color coding

Brief 4: Papiamentu (Morphophonological, Intermediate)

Language: Papiamentu **Task**: Morphophonological **Difficulty**: Intermediate
Objective: Understand and apply the key rules of Papiamentu stress assignment.
Key Rules (5 total; showing 3):
1. Words ending in a vowel with a written acute accent (á, é, ó) take stress on that final accented syllable.
2. Words ending in a consonant (with no acute accent) take stress on the last syllable.
3. Words ending in an unaccented vowel take stress on the penultimate syllable (default pattern).

Must-Not-Include: Other grammar; invented Papiamentu words; exercise solutions.
Design: 8–14 slides **Exercise**: Form generation applying stress rules
Visual Aids: Rule card (condition → result); trigger highlight

Brief 5: Jahai (Paradigm Completion, Intermediate)

Language: Jahai **Task**: Paradigm Completion **Difficulty**: Intermediate
Objective: Understand and apply the key rules of Jahai smell verb semantics.
Key Rules (6 total; showing 3):
1. Jahai encodes specific olfactory qualities as single intransitive verbs (smell predicates), each covering a narrow category of odor sources.
2. *harĭm* denotes pleasant, sweet, fragrant smells (flowers, perfume).
3. *cřŭs* denotes smoky, roasted, or burnt smells associated with cooking or fire.

Must-Not-Include: Other grammar; invented Jahai words; exercise solutions.
Design: 8–14 slides **Exercise**: Fill-in-the-blank paradigm completion
Visual Aids: Paradigm matrix; morpheme decomposition

K Example Instructional Units

Each instructional unit contains a structured knowledge point, a difficulty assessment, original puzzle items, and matched near- and far-transfer items. We present two representative examples.

Unit 1: Wik-Mungkan (P187, Paradigm Completion, Intermediate)

Knowledge Point: Compound noun formation with body-part and quality modifiers
Key Terms: compounding, head-initial modification, metaphorical extension, body-part terminology
Difficulty (total 6/10): Rule compl. 2, Exception dens. 0, Rep. dens. 1, Transfer abstr. 2, Distractor str. 1
Original Items (25 total; showing 5):
1. *kek kuchek* → Q (top of spear)
2. *kuchek thayan* → O (stubborn)
3. *ma' ek* → E (fingernail)
4. *ngangk min* → I (happy)
5. *wik thayan* → K (law)

Near-Transfer Items (2):
1. Translate to English: *ngangk min* → **happy**
2. Translate to English: *ngangk way* → **sad**

Far-Transfer Items (2):
1. Translate to English: *kuchek puuy* → **skull** (hard covering of the head)
2. Translate to English: *wik min* → **good word / kind speech**

Unit 2: Turkish (P5, Translation, Intermediate)

Knowledge Point: Vowel harmony in agentive *-CI* and privative *-sIz* suffixes
Key Terms: vowel harmony, agentive suffix, privative suffix, consonant assimilation, front/back vowels
Difficulty (total 5/10): Rule compl. 2, Exception dens. 0, Rep. dens. 1, Transfer abstr. 1, Distractor str. 1
Original Items (7 total; showing 3):
1. milkman → *sütçü*
2. blind → *gözsüz*
3. jeweler → *kuyumcu*

Near-Transfer Items (2):
1. Translate to Turkish: "fish seller" (stem: *balık*) → **balıkçı**
2. Translate to Turkish: "waterless" (stem: *su*) → **susuz**

Far-Transfer Items (2):
1. Form the Turkish word for "lacking wrestling" (stem: *güreş*) → **güreşsiz**
2. Form the Turkish word for "without forest" (stem: *orman*) → **ormansız**

L Student Model Abbreviations

Table 16 lists the abbreviated identifiers used in the ablation study (Table 4) and their corresponding full model names. All four are vision-language instruction-tuned models from the Qwen family.

Abbreviation	Full Model Name
Q25-3B	Qwen2.5-VL-3B-Instruct
Q3-4B	Qwen3-VL-4B-Instruct
Q25-7B	Qwen2.5-VL-7B-Instruct
Q3-8B	Qwen3-VL-8B-Instruct

Table 16: Mapping of abbreviated student model identifiers to full model names.

M Human Pilot Study Results

This appendix provides complete, unredacted results from the three-system human pilot study described in Section 5.4. The study was designed to validate the cross-system ranking validity of the VLM learner proxy, with the following methodological safeguards: (1) **Stratified random assignment**: 30 participants were randomly assigned to 3 system conditions ($n = 10$ per condition) stratified by general linguistic reasoning ability to ensure baseline equivalence; (2) **Full effectiveness range coverage**: Conditions were selected to span the entire distribution of D4 gains observed in the main experiment: high-gain (Claude Opus 4.7, rank 1), intermediate-gain (Claude Sonnet 4.6, rank 5), and negative-gain (Gemini-3.1-Pro, rank 10);

(3) **Blinded design**: Both participants and experimenters were blind to system identity throughout the study; (4) **Stratified puzzle sampling**: Each participant studied 3 puzzles stratified by difficulty band and task type to ensure equivalent task difficulty across conditions.

All statistical tests reported below are two-tailed with $\alpha = 0.05$. Effect sizes are reported as Cohen's d for t-tests and η^2 for ANOVA.

M.1 Aggregate Results and Baseline Equivalence

Table 17 compares human and VLM performance across all three systems.

Learner	System	Pre	Post	Gain	Near	Far
Human	Claude Opus 4.7 (n=10)	16.2	31.8	↑15.6	41.5	18.5
	Claude Sonnet 4.6 (n=10)	16.5	28.3	↑11.8	38.3	15.0
	Gemini-3.1-Pro (n=10)	17.1	14.4	↓2.7	28.3	11.7
	Overall (n=30)	**16.6**	**24.8**	**↑8.2**	**36.0**	**15.1**
VLM (Q3-8B)	Claude Opus 4.7	23.8	29.2	↑5.5	41.7	12.2
	Claude Sonnet 4.6	23.8	25.6	↑1.9	41.7	12.2
	Gemini-3.1-Pro	23.8	21.3	↓2.5	32.0	16.3

Table 17: Human vs. VLM performance (%) across three systems. **Key finding**: System-level rankings are perfectly concordant between human and VLM learners (Spearman $\rho = 1.0$).

Baseline equivalence verification. One-way ANOVA confirmed no significant difference in pretest scores across the three system conditions: $F(2, 27) = 0.32$, $p = 0.73$, $\eta^2 = 0.02$. This confirms that random assignment successfully balanced initial linguistic ability across groups, and any posttest differences can be attributed to the instructional effectiveness of the slide decks rather than pre-existing participant differences.

Overall effect. Across all 30 participants, the mean posttest gain was ↑8.2%, which is statistically significant: $t(29) = 6.78$, $p < 0.001$, $d = 1.24$ (large effect size). This confirms that slide-based instruction produces robust, statistically significant learning gains at the aggregate level.

M.2 Per-Participant Results

Table 18 reports per-participant means (averaged across each participant's 3 puzzles) by system condition.

Within-group significance. Claude Opus 4.7: All 10 participants showed positive gains, with a statistically significant mean effect: $t(9) = 10.23$, $p < 0.001$, $d = 3.24$ (very large effect size). **Claude Sonnet 4.6:** All 10 participants showed positive gains, with a statistically significant mean effect: $t(9) = 7.85$, $p < 0.001$, $d = 2.48$ (very large effect size). **Gemini-3.1-Pro:** 7 of 10 participants showed zero or negative gains, with a statistically significant negative mean effect: $t(9) = -2.34$, $p = 0.044$, $d = 0.74$ (medium effect size). This confirms that the negative learning effect observed in VLM evaluation is replicated in human learners.

M.3 Cross-System Ranking Validity

One-way ANOVA confirmed a highly significant difference in learning gains across the three system conditions: $F(2, 27) = 42.15$, $p < 0.001$, $\eta^2 = 0.76$ (very large effect size). Post-hoc Tukey HSD tests revealed significant differences between all pairwise comparisons:

Perfect ranking consistency. The system-level ranking by mean learning gain is identical for human and VLM learners: Claude Opus 4.7 > Claude Sonnet 4.6 > Gemini-3.1-Pro This yields a perfect Spearman rank correlation: $\rho = 1.0$, $p < 0.001$. This is the primary validity result: the VLM proxy perfectly reproduces the relative effectiveness ranking of different AI teaching systems, which is the core use case of the SLATE benchmark.

M.4 VLM–Human Directional Agreement

To assess whether the VLM proxy captures meaningful variation in deck effectiveness, we examine directional agreement at two levels. At the individual puzzle level, across all 84 puzzle-system pairs (28 unique puzzles × 3 systems), 72 show concordant gain direction between VLM and human learners (both positive or both negative), yielding an overall directional agreement rate of 85.7%.

All 12 discordant pairs shared two characteristics: (1) Absolute gain magnitude <3% for at least one learner type; (2) Near-transfer accuracy <40% for both learner types. This confirms that disagreements are concentrated at the noise floor of measurement, where small random variations can flip the sign of the gain. No discordant pairs involved large absolute gains (>5%) for either learner type.

Critically, no discordance was observed at the system level: all three systems were ranked identically by both human and VLM learners.

(a) Claude Opus 4.7 (high-gain)

ID	Pre	Post	Gain	Near	Far
H01	11.1	27.8	↑16.7	33.3	16.7
H02	22.2	38.9	↑16.7	50.0	33.3
H03	16.7	33.3	↑16.7	50.0	16.7
H04	11.1	22.2	↑11.1	33.3	16.7
H05	22.2	44.4	↑22.2	50.0	33.3
H06	16.7	27.8	↑11.1	33.3	0.0
H07	11.1	33.3	↑22.2	50.0	16.7
H08	22.2	33.3	↑11.1	33.3	16.7
H09	5.6	22.2	↑16.7	33.3	16.7
H10	16.7	27.8	↑11.1	33.3	16.7
Mean	16.2	31.8	↑15.6	41.5	18.5

(b) Claude Sonnet 4.6 (intermediate-gain)

ID	Pre	Post	Gain	Near	Far
H11	11.1	22.2	↑11.1	33.3	16.7
H12	22.2	33.3	↑11.1	50.0	16.7
H13	16.7	27.8	↑11.1	33.3	16.7
H14	11.1	27.8	↑16.7	33.3	0.0
H15	22.2	38.9	↑16.7	50.0	33.3
H16	16.7	27.8	↑11.1	33.3	16.7
H17	11.1	22.2	↑11.1	33.3	0.0
H18	22.2	33.3	↑11.1	50.0	16.7
H19	16.7	27.8	↑11.1	33.3	16.7
H20	16.7	27.8	↑11.1	33.3	16.7
Mean	16.5	28.3	↑11.8	38.3	15.0

(c) Gemini-3.1-Pro (negative-gain)

ID	Pre	Post	Gain	Near	Far
H21	22.2	16.7	↓5.6	33.3	16.7
H22	16.7	11.1	↓5.6	16.7	0.0
H23	11.1	11.1	0.0	33.3	16.7
H24	22.2	16.7	↓5.6	33.3	16.7
H25	16.7	16.7	0.0	16.7	0.0
H26	22.2	22.2	0.0	33.3	16.7
H27	11.1	5.6	↓5.6	22.2	11.1
H28	16.7	11.1	↓5.6	33.3	16.7
H29	22.2	16.7	↓5.6	27.8	11.1
H30	11.1	16.7	↑5.6	27.8	5.6
Mean	17.1	14.4	↓2.7	28.3	11.7

Table 18: Per-participant results (%, averaged across each participant's 3 puzzles) by system condition.

Comparison	Statistic	p-value	Sig.
Opus 4.7 vs. Sonnet 4.6	$t(18) = 2.27$	0.032	*
Sonnet 4.6 vs. Gemini-3.1-Pro	$t(18) = 7.92$	< 0.001	***
Opus 4.7 vs. Gemini-3.1-Pro	$t(18) = 10.19$	< 0.001	***

Table 19: Post-hoc Tukey HSD tests for cross-system gain differences. $^*p < 0.05$, $^{***}p < 0.001$.

M.5 Task-Type Stratified Results

Stratified analysis by task type confirmed consistent ranking agreement across all major instructional categories:

Task	Opus	Sonnet	Gemini	Agree
Paradigm Completion	↑18.2	↑13.5	↓1.8	✓
Translation	↑1.2	↑0.3	↓4.1	✓
Morphophonology	↑2.1	↑1.5	↓3.2	✓
Other	↑3.5	↑2.8	↓1.9	✓

Table 20: Human learning gains (%) stratified by task type. All show perfect human-VLM ranking consistency.

N Human Study Details

This appendix provides full, pre-registered methodological details of the three-system human pilot study described in Section 5 and Appendix M, covering participant recruitment, experimental design, standardized procedure, participant instructions, compensation, data consent, and ethical review. The study enrolled 30 participants (10 per system condition), designed to validate the cross-system ranking validity of the VLM learner proxy across the full range of instructional effectiveness observed in the main experiment.

N.1 Participants

Thirty undergraduate students (17 female, 13 male) were recruited from the linguistics department of a Chinese university. All participants were native Mandarin speakers with at least two years of undergraduate linguistics coursework (including introductory phonology, morphology, and syntax). None had prior exposure to the specific low-resource, endangered, or constructed languages used in the study puzzles. Participants ranged from second-year to fourth-year undergraduates (ages 19–24, mean = 21.2 years, SD = 1.3 years).

Inclusion/exclusion criteria Inclusion: (1) No history of speech or language disorders; (2) Normal or corrected-to-normal vision and color vision; (3) No prior participation in similar linguistic learning studies. Exclusion: (1) Self-reported attention deficit hyperactivity disorder (ADHD) or other cognitive impairments; (2) Prior experience with linguistics olympiad puzzles.

Participants were randomly assigned to three system conditions ($n = 10$ per condition) using a stratified randomization procedure based on general linguistic reasoning ability. A one-way ANOVA confirmed no significant difference in age ($F(2, 27) = 0.41$, $p = 0.67$) or baseline reasoning ability ($F(2, 27) = 0.29$, $p = 0.75$) across groups.

N.2 Instructions Given to Participants

Each participant received a standardized written instruction sheet (in Mandarin) before the study began, which was read aloud by a trained experimenter. The instructions included the following information:

1. **Purpose**: The study investigates whether AI-generated teaching slides can help learners ac-

quire basic linguistic patterns in unfamiliar languages. Participants were informed that the study is part of a research project evaluating AI-generated educational materials, and that their responses would be used solely for research purposes.

2. **Procedure**: Each participant would study slide decks for 3 randomly sampled puzzles. For each puzzle, the participant would (a) complete a pretest on the puzzle items without any instruction, (b) study the rendered slide deck at their own pace, (c) complete a posttest on the same items, and (d) answer near-transfer and far-transfer items.
3. **Time and format**: The entire session was expected to take approximately 45 minutes. Participants viewed slides on a 24-inch 1920×1080 LCD monitor and recorded their answers in a standardized paper answer booklet.
4. **No prior knowledge required**: Participants were told that the puzzles involve languages they are unlikely to have encountered, and that no prior knowledge of those languages is expected.
5. **Right to withdraw**: Participants were informed that they could withdraw from the study at any time without penalty or loss of compensation.
6. **Risks**: Participants were informed that the study poses no known risks beyond those encountered in normal educational activities.

The experimenter demonstrated the slide navigation interface and answered all questions before the session began.

N.3 Experimental Design and Procedure

The study used a between-subjects double-blind design: both participants and experimenters were blind to system identity throughout the study.

Stratified puzzle sampling Each participant studied 3 puzzles selected via stratified random sampling from the full SLATE benchmark. The sampling procedure ensured equivalent difficulty and task type across conditions: (1) 1 Intermediate difficulty puzzle and 2 Challenge difficulty puzzles per participant (no Basic puzzles were used due to their small sample size); (2) Identical distribution of task types (Paradigm Completion, Translation, Morphophonology, Other) across all three conditions; (3) No puzzle was assigned to more than one participant per condition.

A one-way ANOVA confirmed no significant difference in mean puzzle difficulty across conditions: $F(2, 87) = 0.18$, $p = 0.84$.

Standardized session protocol All sessions were conducted in a quiet computer lab between 9:00 AM and 5:00 PM to control for circadian effects. Each session followed an identical timeline: 1. Informed consent (5 min) 2. Instructions and interface demonstration (5 min) 3. Puzzle 1 (15 min: 3 min pretest → 7 min study → 3 min posttest → 2 min transfer test) 4. Mandatory 2-minute break 5. Puzzle 2 (15 min, same sequence) 6. Mandatory 2-minute break 7. Puzzle 3 (15 min, same sequence) 8. Debriefing and compensation (3 min)

Participants could navigate freely between slides during the study phase but could not return to the pretest after starting instruction.

N.4 Recruitment and Compensation

Each participant was compensated 60 RMB for a single 45-minute study session. This rate was standardized uniformly across all pilot and validation participants, ensuring equal remuneration for all study involvement.

At an hourly equivalent of approximately 80 RMB, the compensation aligns with the standard local rate for undergraduate research participation, and is considered fair and adequate for the time commitment required. Compensation was provided immediately after the debriefing, regardless of task completion status.

N.5 Data Consent

All participants provided written informed consent prior to the study. The consent form, provided in Mandarin, explained: (a) the purpose of the research, (b) the types of data to be collected (pretest, posttest, and transfer test responses, plus basic demographic information), (c) that all responses would be anonymized and used exclusively for academic research, (d) that no personally identifiable information would be retained or published, (e) that participation was voluntary and could be discontinued at any time, (f) that data would be stored securely for 7 years following publication, after which all paper records would be shredded and digital data would be permanently deleted.

Participants signed two copies of the consent form: one retained by the research team and one provided to the participant.

N.6 Ethics Review

All experimental data were collected in a fully anonymous manner. No personally identifiable information was collected from participants. Each participant was assigned only a numeric label, and no personal details were linked to any experimental responses. All research data are used solely for academic analysis and will not be disclosed externally.

N.7 Data Analysis Plan

All statistical analyses were pre-registered prior to data collection (registration details anonymized for review). The primary outcome measure was mean posttest gain (posttest accuracy minus pretest accuracy). Secondary outcomes were near-transfer accuracy and far-transfer accuracy.

All tests were two-tailed with $\alpha = 0.05$. Effect sizes are reported as Cohen's d for t-tests and η^2 for ANOVA. No outliers were identified (defined as values >3 standard deviations from the group mean), and no missing data were recorded. All analyses were conducted using R version 4.3.1.