

# Towards an Evaluation Methodology for AI in Second Language Education: Lessons Learned from Developing L2-Bench

James Edgell<sup>1\*</sup>, Wm. Matthew Kennedy<sup>2\*</sup>  
Isaac Pattis<sup>1</sup>, Ben Knight<sup>1</sup>, Danielle Carvalho<sup>1</sup>, Elizabeth Wonnacott<sup>3</sup>

<sup>1</sup>Oxford University Press

<sup>2</sup>Oxford Internet Institute, University of Oxford

<sup>3</sup>Department of Education, University of Oxford

Correspondence to: L2-Bench Team (elt-bench@oup.com)

## Abstract

The rapid adoption of large language models in AI-powered language education has created an urgent need for evaluations that assess pedagogical effectiveness, particularly in language learning—one of the most common LLM use cases (Tamkin et al. 2024; Costa-Gomes et al. 2025). With only narrowly defined task-specific evaluations of AI system capabilities in second language (L2) education existing in the literature, we require more holistic approaches in this AI for education space. To address this gap, we describe the iteration of the methodology we developed to build L2-Bench, a novel, context-specific evaluation benchmark grounded in a validated “language learning experience designer” construct to assess AI capabilities across L2 education contexts. Our methodology integrates pedagogical theory, sociotechnical AI evaluation methods, and operationalizes a hierarchical taxonomy to structure an expert-curated dataset of over 1,000 authentic rubric-scored task-response pairs with measurement and scoring pipeline. We report the results of a pilot validation exercise (N = 39) on an initial sample of our dataset (tasks were validated as authentic [M = 4.23/5], but criteria scores were lower [M = 3.94], with universally poor inter-annotator agreement despite good internal consistency), alongside the experimental design for our follow-up practitioner data validation study as we iterate and scale to the full dataset. Ultimately, this research not only offers methodological lessons towards a more context-specific AI evaluations ecosystem, but also works towards better design of reproducible evaluations for AI systems deployed to educational contexts.

## Introduction

Access to language education is a human right (World Conference on Linguistic Rights 1996). Large Language Models (LLMs) are rapidly being integrated into language learning products used by millions of learners worldwide. However, no widely recognised benchmarks exist for evaluating AI capabilities in education beyond narrow lesson generation (Clark et al. 2020) or knowledge of pedagogical concepts (Lelièvre et al. 2025). Although important aspects, learning experience design requires knowing when and how to use these capabilities in practice, reflecting the diverse educational scenarios encountered in real-world settings.

\*These authors contributed equally.

The absence of comprehensive evaluation frameworks for AI in education creates several critical challenges. Educators and institutions lack systematic methods to assess AI capabilities for specific teaching scenarios, leading to uninformed adoption decisions. Product developers cannot rigorously validate their AI implementations against pedagogical best practices. Most importantly, the lack of standardized evaluation impedes the development of more effective AI-powered educational systems.

In this paper, we present consequential lessons learned in the construction of L2-Bench, an evaluation benchmark assessing AI performance in second language learning experience design. We discuss our iterative efforts to develop and validate artefacts produced by our methods, and in doing so, we make three contributions:

1. **A novel methodology** for developing AI evaluation benchmarks specific to the unique norms, values, and dynamics of educational spaces.
2. **A hierarchical taxonomy** of 12 competencies and 31 subcompetencies that comprise a L2 learning experience design construct, grounded in established pedagogical framework
3. **Lessons learned in the development of our novel methodology**, including positive and negative results of our efforts to validate our core technical components: taxonomy, measures, and an initial sample (325 items) of our planned 1,300 item dataset.

Overall, we hope that our contributions help move AIED evaluations in general beyond narrow accuracy metrics toward more rigorous and context-sensitive assessment of AI capabilities in educational spaces.

## Related work

The speed and scale of AI system deployment have outpaced the development of evaluation methodologies for novel technologies and real-world contexts (Feffer et al. 2025; Bean et al. 2025; Reuel et al. 2024; Schwartz et al. 2025), prompting calls for greater rigor, systematization, and the incorporation of social scientific methods (Butler et al. 2024; Weidinger et al. 2025; Olteanu et al. 2025).

This evaluations crisis is acute in education, where no generally accepted holistic evaluations for AI in education exist despite rapid adoption (Digital Education Council 2024; Costa-Gomes et al. 2025). While some evaluations assess performance in highly rules-based domains (e.g., mathematics or computer science), these are not readily transferable to more open-ended educational settings. Conversely, benchmarks claiming to measure broad constructs such as knowledge or reasoning risk importing inappropriate measurement techniques into educational contexts. These limitations create conditions for substantial, compounding risk (Bastani et al. 2024; Kennedy and Vargas Campos 2026).

Evaluations for AI in education are only beginning to emerge. A growing body of scholarship suggests AI use may yield marginal learning benefits but at a cost to engagement (Pardos and Bhandari 2023; Nie et al. 2025). Google DeepMind’s LearnLM team has produced a series of evaluations during the development of a tutorial chatbot (Jurenka et al. 2024) and, in partnership with Eedi, conducted a small (N = 165) RCT reporting a 5.5% improvement in independent problem solving over human tutoring alone (93.6% probability of a genuine improvement) (Team et al. 2025). However, these evaluations remain proprietary and focus narrowly on tutorial interaction, privileging instruction-following over holistic pedagogical adaptivity (LearnLM Team and Google 2025).

Other efforts are noteworthy but limited. Xu et al. (2025)’s general-purpose AI education evaluation lacks grounding in widely accepted pedagogical frameworks. Oak National Academy released a benchmark dataset for AI-generated educational content safety, though its scope is restricted to safety concerns (Clark et al. 2025). Kennedy and Vargas Campos (2026)’s taxonomy of AI harms in education advances context-specific evaluation but has yet to be widely operationalized. Shetye (2024) qualitative analysis of Khanmigo using Chapelle (2001)’s CALL framework offers useful insights but relies on personal experience and thus remains anecdotal.

## Building L2-Bench

### Theory and design

A gap remains. Current evaluations insufficiently cover educational contexts, particularly second-language learning, and frequently misgeneralize tutorial interaction to group instruction. We present the first AI evaluation methodology for language learning to our knowledge. It is also among the first *holistic* evaluations of AI in educational contexts that evaluates AI models at both the instance- and systemic- level (Solaiman et al.).

Context-specificity is key. Language education is different from most other areas of education: it requires learners to acquire more implicit knowledge and proceduralised skills than other subjects; it comprises both the target of learning and the means of learning; it is heavily influenced by affective factors (motivation, identity, anxiety, confidence, and willingness to communicate) (Papi and Khajavy 2023); and it is fundamentally shaped by each learner’s own social and cultural experience (Poehner and Lantolf 2024). As a result,

language is never “solved” (DeKeyser and Suzuki 2025).

Furthermore, classroom learning is not the aggregation of individual interactions; knowledge is produced through social interaction (Bandura 1977; Kennedy and Vargas Campos 2024). Because AI systems shape learning both directly and indirectly through instructional materials, evaluation must extend beyond subject-matter knowledge to capture learning experience design (Knight et al. 2026; Kennedy and Vargas Campos 2026). Indeed, our evaluation is more interested in assessing LLM capabilities in designing experiences that promote the “doing,” not diagramming, of language (Searle 1996; Austin 1975).

As the contours of language learning are unique, we aim to produce evaluation artefacts that are each representative of the peculiar “vernacular” (Kennedy and Vargas Campos 2024) of language learning. We draw upon three frameworks common to UK, EU, and global language learning design: the Council of Europe’s Common European Framework of Reference for Languages (CEFR) (Council of Europe 2001), the Equals Framework for Language Teacher Training and Development (European Association for Quality Language Services 2016) and the British Council’s Continuing Professional Development Framework (British Council 2025). We also engage experts in a series of practitioner validation exercises.

The pedagogical frameworks utilized in this study correspond to the leading UK/EU frameworks for teaching these varieties of English. We utilize these frameworks here to demonstrate the importance of these components to the proposed methodology. In future implementations of this method, other global frameworks of importance could likewise be used to produce evaluation tasks representative to the target teaching context. Note that Equals reflects global L2 education best practices. Lastly, we should restate that our benchmark initially intends to assess model performance specifically on EFL education (English as a Foreign Language), in the medium of US or UK varieties of English (see Appendix C.1 for more). We do so because these varieties of English represent the most-often taught second languages across the world (Blanco 2025). We recognize that models do not perform equally well across World English varieties (Smart et al. 2024).

### Components

**Taxonomy** To evaluate AI capabilities in language education learning design, we first define a “learning experience designer in second language education” which encompasses the range of roles that intentionally design the conditions that shape how people learn: teachers, materials developers (content or assessment creators), learning designers, and teacher trainers (see Appendix C.1 for glossary). We then define the construct as a hierarchical competency taxonomy that articulates the capabilities required for effective “learning experience design” in L2 education.

Our cross-disciplinary team initially developed the taxonomy under three constraints: (1) understandability – avoiding deep hierarchies that obscure interpretation; (2) independence – reducing ambiguity in task classification while acknowledging overlap; and (3) practitioner credibility –

we draw upon several established pedagogical frameworks adapted interactions with AI systems.

It is useful to distinguish between two complementary frameworks for language learning: a knowledge framework representing what learners need in order to learn (i.e. capturing the science of how language learners learn), and a competency framework representing the application of that knowledge (i.e. what practitioners need to do to apply knowledge effectively). They are critical distinctions: L2-Bench is based on the latter; we do not merely benchmark pedagogical knowledge as existing evaluations already show saturation on knowledge-based tasks (Lelièvre et al. 2025).

The current L2-Bench taxonomy comprises 12 competencies and 31 sub-competencies organized into 12 main competencies across a two-level hierarchy (Figure 1) that span the full scope of a **second language learning experience designer in second language education** construct.

Table 1: L2-Bench taxonomy competencies.

Number	Competency
1	Course Planning
2	Lesson Planning
3	Activity Planning
4	Language Presentation
5	Activity Management
6	Exchange Partner
7	Performance Evaluation
8	Giving Feedback
9	Progress Tracking
10	Emotional Intelligence
11	Assessment Creation
12	Professional Development

Each competency comprises one to six sub-competencies that capture specific capabilities assessable through “consensus criteria” (Components). For example, to be competent in “Giving Feedback” requires four sub-competencies: identifying errors and diagnosing their causes; prioritizing areas for feedback; providing explanations, models, or hints; and providing improvement activities. See Appendix C.2 for the full L2-Bench competency taxonomy.

Validation of the taxonomy proceeded through multiple stages: iterative review during dataset construction of the tasks and criteria (Components); the pilot validation (Pilot Validation Exercise) providing empirical signal on whether tasks designed around the taxonomy measured coherent constructs; and forthcoming practitioner validation with representative stakeholder groups (Future Work).

**Dataset** The quality of an evaluation benchmark hinges on the questions it asks. To this end, we operationalize our competency taxonomy to design task-response pairs that assess capabilities against each competency, targeting over 1,000 tasks to create a high-quality dataset with meaningful variation. Task design follows four heuristics adapted from the UK AI Safety Institute’s guidance on question-answer pair development (UK AI Security Institute 2024) and recently

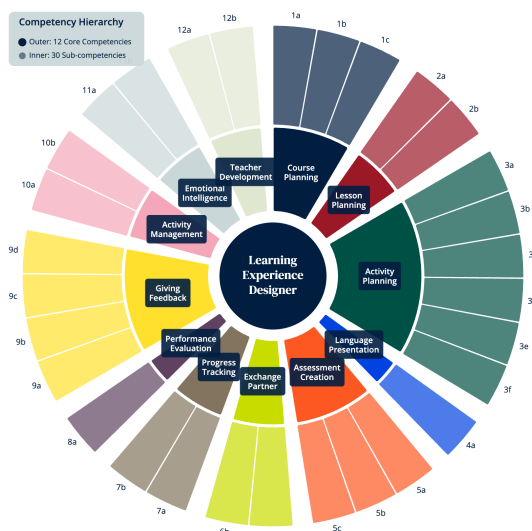


Figure 1: L2-Bench Taxonomy of Competencies - sunburst visualization showing the 12 competencies and 30 sub-competencies of a “learning experience designer in second language education.” See Appendix C.2 for large format figure.

proposed AICALL frameworks (Bahari, Han, and Strzelecki 2025):

- **Relevance:** Tasks are authentic real-world scenarios that directly relate to “learning experience designer competencies”, therefore requiring open responses.
- **Perspective:** Tasks consider multiple stakeholder viewpoints with full coverage of educational contexts.
- **Clarity:** Tasks include sufficient guidance to enable appropriate responses while avoiding ambiguity that could lead to inconsistent scoring.
- **Originality:** Tasks test application of pedagogical knowledge to new situations rather than relying on memory.

We constrained tasks to single-turn conversations (task-response pairs) to enable simplicity in data creation and evaluation. Multi-turn alternatives involve difficult experimental setup (recruiting users; contrived scenarios of experts role-playing as users; or unvalidated synthetic users) as well as requiring both turn-level and conversation-level evaluations that increase complexity. We acknowledge this constraint trades off against real-world relevance (learning interactions are inherently multi-turn), and we actively plan to expand into multi-turn dataset production. However, we note that many *learning design* tasks are appropriately represented as single turn interaction (e.g. an instructor prompting an LLM for a quick introductory activity). And furthermore, beginning with the simplest possible form of interaction (adjacency pairs) provides a stable baseline, which not only aids iterative development but also project feasibility.

The dataset comprises tasks distributed across the 12 competencies that span diverse teaching contexts (“task vari-

ables”): geographic regions (Far East, South-East Asia, Middle East, Latin America, Europe, Africa), learner profiles (ages from 4–6 to 26+, CEFR levels A1–C2, primary school to corporate training settings), and learning aims (academic, professional, exam preparation, travel, cultural).

To ensure scalability of generation and maintain experimental control, L2-Bench items are produced through a hybrid human-AI authoring approach that is modelled on publishing workflows:

1. “design”: language pedagogy experts create hand-crafted task exemplars and establish prompt templates for both task and reference answer creation
2. “draft”: state-of-the-art foundation models with agent scaffolding generate candidate tasks, task criteria, and reference answers using the prompt templates and examples
3. “review”: experts iteratively refine generated content, with modifications triggering regeneration cycles
4. “approval”: a separate expert validates the reviewed items for pedagogical soundness
5. “publish”: items are published in version-controlled benchmark dataset release (see Appendix D.1).

To create authentic practitioner scenarios, some tasks include “resources” (documents in markdown or CSV), however since our goal is not to benchmark tool-handling phenomena, we simply append these resources in-context. Representative task examples are provided in Appendix D.2.

Dataset validation proceeds through multiple stages: internal error analysis examines tasks for unrealistic scenarios, sociocultural biases, and scoring ambiguity; pilot data validation (Pilot Validation Exercise) provides quantitative signal on task authenticity and criteria quality for 325 task-response pairs, as well as qualitative reports that inform the next stage of iterative refinements; and future practitioner data validation (Future Work) following scaling to the full 1,000+ item dataset.

**Measures** Since tasks are designed to elicit open responses, evaluation cannot rely on standard accuracy metrics. We employ rubric-based measurement common to high-stakes, contextual domains (e.g. medical diagnosis) (Arora et al. 2025).

**Binary Criteria.** Scoring rubrics are composed of weighted criteria, where each criterion outlines what an ideal response should include or avoid. We employ binary pass/fail decisions for each criterion rather than Likert scales for two reasons: (1) calibrating both human annotators and LLM-Judges (“auto-scorers”) on multi-point scales where the differences between adjacent points are subtle is challenging (Yan 2025); (2) binary labels yield faster, more consistent human annotations, and simpler auto-scorer alignment due to clear decision boundaries. Criteria are assigned point values from -10 to +10 based on importance, with negative points for undesirable responses. Final scores are computed by summing points for criteria met and dividing by the maximum possible score.

**Criteria System.** The hierarchical structure of the competency framework enables granular assessment while maintaining connection to broader pedagogical competencies. We define three types of criteria:

**Consensus Criteria.** Defined per sub-competency with expert agreement, capturing essential requirements for competent performance. Tasks inherit consensus criteria from their tagged sub-competencies. For example, any task tagged with sub-competency 08a (identifying errors and diagnosing causes) includes the consensus criterion: “Estimates the likely causes of the error - e.g. gaps in knowledge, or skill proficiency” (weight: +5).

**Task Criteria.** Criteria specific to individual tasks that capture context-specific requirements (designed to be unique).

**Universal Criteria.** Criteria applied to all tasks, with weightings conditional on task context, for example, CEFR-level language appropriateness (weight: +9 for learner-facing responses, +2 for teacher-only responses) and child-safety considerations (weight: -10 for offensive content, -5 for sensitive content requiring teacher guidance).

A task rubric is therefore composed of task criteria, consensus criteria, and universal criteria, all designed to be independent of one another (see Appendix C.2 and Appendix C.3). Reference answers are designed to score maximally against each task rubric, providing guidance for both human and automated scoring.

**Scoring pipeline** We do not want to underestimate what models can do, but instead elicit their best response, and therefore every task has a system prompt to capture the implicit context a practitioner with expertise in the underlying competency may have to perform the task, without revealing task scoring rubrics. We design system prompts using best practices in capability elicitation (UK AI Safety Institute 2024) pairing each task with its own unique prompt template of:

1. role establishment
2. domain expertise
3. contextual framing (based on the “task variables”)
4. optional chain-of-thought reasoning.

We recognise that optimal elicitation techniques vary across models, but we accept this trade-off vs unelicited prompts that may underestimate capability or even favor certain models over others due to inherent prompt biases (Abbas, Waggoner, and Olive 2025).

We use open or cloud-hosted API endpoint models for all generation and scoring pipelines in order to minimise leakage to model providers and therefore mitigate benchmark saturation and contamination. A model receives the task-specific system prompt and the task itself (with any “resources”) as input and generates an open-response.

Scoring open-responses against rubrics at scale requires automated approaches; we therefore employ LLM-as-a-Judge (auto-scorer) as our scoring mechanism for assessing

open-ended generations against our task rubrics. Having deconstructed our rubrics into binary decisions, we can present simpler judgments to the auto-scorer by providing it with a single criterion at a time (Yan 2025), whilst also mitigating (but not solving) documented limitations in LLM-as-a-Judge literature, such as verbosity bias (preferring longer responses) and judge bias (favouring responses from the same model family) (UK AI Safety Institute 2025). We will employ techniques to test for known failure modes that our research design implicates most clearly, such as pairwise testing (Liu et al. 2024) and playing favorites (Spiliopoulou et al. 2025).

We develop our auto-scorer in two phases: (1) initially, we use Claude Sonnet-4.5 with thinking as the auto-scorer foundation model for our scoring pipeline (outputting reasoning traces to enable subsequent meta-analysis), then (2) we will later collect scores from human experts in our future practitioner data validation and use this data to optimise the auto-scorer by varying models and prompts.

Currently, we employ reference-guided prompting for our auto-scorer, whereby the auto-scorer receives the task-response pair along with the original task context, task metadata (competency, sub-competency, task variables), the reference "gold standard" answer, and a single criterion from the task rubric to score against. This approach assumes that the reference answer is an objective solution. Of course, this does not hold in many educational contexts, and so we temper our expectations for inter-judge agreement in our future practitioner data validation (Future Work).

Our scoring pipeline begins with generating a response per task. For each response, our auto-scorer determines whether the response meets each criterion; if the criterion is met, full points are given, otherwise no points are given. We then get the total points for a given task-response pair by summing the point values for criteria met and dividing by the maximum possible score to produce the task score (a task score can be negative if more negative points were assigned to it than positive points). For each task, we calculate an AI system's overall L2-Bench score by taking the mean of these task-level scores, clipped between 0–100

To ensure trust in our future L2-Bench leaderboard rankings, we will conduct uncertainty quantification that distinguishes genuine performance gaps from measurement noise (Miller 2024). We will therefore generate  $n = 3$  responses per task and compute the mean score, reporting standard errors alongside point estimates to enable statistical inference on model differences rather than relying solely on rank ordering (see Appendix E for details).

### Pilot Validation Exercise

Prior to full-scale development of L2-Bench, we conducted a pilot validation study (IRB-exempt) which served two purposes:

1. **gathering early feedback** to improve the competency construct and dataset design before scaling to the full 1,000+ tasks
2. **informing best practices** for a future global practitioner data validation (Future Work), such as establishing eval-

uation design parameters, identifying challenging competencies, and calibration guidelines.

### Methods

Participants ( $N = 39$ ) were recruited from a leading UK university in collaboration with the university's careers network team. Recruitment proceeded via an intranet announcement for voluntary work experience that included an application form detailing: eligibility (postgraduates enrolled in taught masters programmes in Arts, Humanities and Social Sciences), time commitments (2-3 hours per week over six weeks) and the study format (a team-based 'challenge').

While this participant group was chosen primarily for availability within our development timelines, their critical thinking backgrounds and diverse perspectives were deemed appropriate to scrutinise our benchmark components to identify patterns that would allow us to iterate. Furthermore, despite not holding practitioner roles that would be best placed to validate L2-Bench (see Section 5), 32 of the participants were international students with L2-English proficiency, 8 had prior teaching experience, and 4 were enrolled on MA Education programmes, all relevant experiences direct or adjacent to L2 education that could be reasonably drawn upon for initial validation signal on elements of our benchmark.

Participants were initially put into 8 teams of 6 stratified by L2-English and teaching experience to balance expertise levels (see Appendix A.2 for complete team composition), with the challenge format (involving prize incentives) encouraging teams to operate independently as cluster-level evaluators ( $N = 8$ ), minimising leakage between samples. Teams were assigned 325 task-response pairs in randomised order across each team to ensure broad coverage across all competencies. Data (complete with all task metadata, reference answer) was provided in both Excel format and on an internally-hosted instance of the open-source Langfuse annotation platform (which was also used to collect item-level ratings and free-text annotations) to accommodate different working preferences. Task responses were generated with Claude Sonnet-3.7 via Amazon Bedrock API with default settings ( $T = 1.0$ , top-p = 0.999, top-k = 250).

To measure dataset validity, we encouraged participants to rate task authenticity and criteria adequacy for each task item in the dataset on 5-point Likert scales (we used rather than 7-point scales to enable faster ratings without sacrificing reliability (Preston and Colman 2000)):

- **Authenticity score:** Does the task represent an authentic L2 educational scenario?
- **Criteria score:** Do the task rubric criteria evaluate the AI response well enough to distinguish good vs poor responses?

Beyond individual ratings, teams produced written reports identifying systematic issues in task design and broader competency construct and proposing revisions. Teams received five hours of dedicated in-person training, with two 2-hour sessions dedicated to study orientation and calibration (objectives, background context, interface, practice task validation examples, norming discussions) within the first

two weeks, plus one additional hour per team for focused question-and-answer support in the third week.

## Results

**Response rates and coverage** The study collected 1,128 ratings across 325 unique tasks, achieving an overall response rate of 43% (1,128 of 2,632 possible rating opportunities). Response rates varied considerably across teams, ranging from 15% to 89% (see Appendix A.3), reflecting a combination of team dynamics, the substantial time required per task (teams estimated 10–20 minutes for the average task) and poor UI experience on the annotation platform (teams preferred the spreadsheet interface). Critically, there was 100% task coverage (all tasks received at least one rating), with 77% of tasks receiving three or more independent ratings - the minimum typically required for meaningful reliability analysis. Response patterns varied systematically across competency domains, revealing which areas evaluators found more accessible (see Appendix A.4). Certain responses were missing but not at random, with specialised competencies showing notably higher skip rates: "Evaluate student's performance" (69%) and "Support professional development" (63%). Given teams had varying levels of experience with AI tools and pedagogical assessment, this self-selection pattern may indicate that only evaluators feeling qualified responded, potentially concentrating expertise in those ratings.

**Dataset validity** After converting the 5-point Likert scales (where 1='Strongly Disagree' through 5='Strongly Agree'), evaluators broadly endorsed benchmark tasks as representing realistic teaching scenarios, with authenticity scores averaging mean  $M = 4.24$  (95% CI [4.19, 4.30]). This offers some evidence that our task design methodology reflects authentic teaching practice, albeit the UK-based university evaluators limits generalisability to global L2-educational contexts. However, criteria scores were lower ( $M = 3.93$ , 95% CI [3.87, 3.98]). The Wilcoxon signed-rank test confirmed this gap ( $W = 42,690$ ,  $p < 0.001$ ,  $d = 0.28$ ), indicating that while participants agreed that the tasks are realistic, they are less sure that the criteria are sufficient to evaluate pedagogical quality of AI responses. Table 2 gives the statistical summary of the pilot data validation (see Appendix A.1 for complete statistical methodology).

Performance varied meaningfully across competency domains, with pedagogically complex competencies such as "Present language learning points" ( $M = 3.69$ ) and "Giving feedback" ( $M = 3.77$ ) scoring lowest on criteria scores despite high authenticity ratings (see Figure 2).

We measured inter-annotator agreement on criteria (hereafter criteria IAA) using Krippendorff's Alpha to handle incomplete data (see Appendix A.1). Criteria IAA scores were universally poor across all competencies (see Figure 3). Two-thirds (8/12) showed negative values, indicating systematic disagreement where evaluators diverged more than expected by chance. The maximum alpha achieved was just 0.10 (Course Planning), far below the 0.80 threshold required for reliable conclusions (Krippendorff 2004).

We also conducted criteria internal item consistency

Table 2: Summary statistics from pilot data validation. Auth=Authenticity; Crit=Criteria; M=Mean; CI=95% confidence interval; SD=Standard deviation; IAA=Inter-annotator agreement (Krippendorff's  $\alpha$ ) for criteria; IIC=Internal item consistency (Cronbach's  $\alpha$ ) for criteria.

Metric	Value
Tasks	325
Ratings	1,128
Skip %	57%
Auth M	4.24
Auth CI	[4.19, 4.30]
Crit M	3.93
Crit CI	[3.87, 3.98]
Crit SD	0.99
IAA ( $\alpha$ )	-0.01
IIC ( $\alpha$ )	0.95

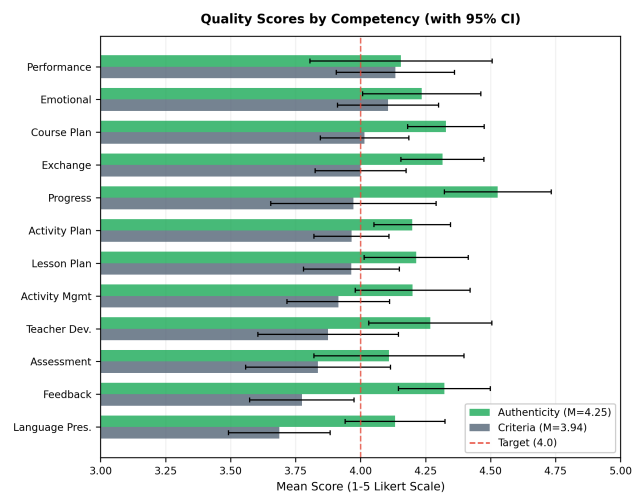


Figure 2: Criteria and authenticity scores per competency with 95% CIs.

(hereafter criteria IIC) analysis using Cronbach's Alpha to assess whether tasks within each competency measured a coherent underlying construct (see Appendix A.1 for details). Most competencies demonstrated at least low-to-moderate criteria IIC values ( $\alpha \geq 0.40$ ) on criteria scores, with overall criteria IIC achieving excellent reliability ( $\alpha = 0.95$ ) (see Figure 2).

This pattern of poor criteria IAA with low-to-excellent criteria IIC is consequential. It provides initial evidence that tasks designed around the competency taxonomy are *measuring the same underlying concept*, even if evaluators *apply systematically different standards* to assess the suitability of proposed ways of measuring that concept.

We can explain this result primarily via calibration: ultimately evaluators were postgraduate students, not experienced practitioners, meaning that evaluators brought their own implicit standards for "good" performance. Future validation exercises with professional practitioners will shed

Criteria Inter-Annotator Agreement (IAA) and Inter-Item Consistency (IIC) by Competency

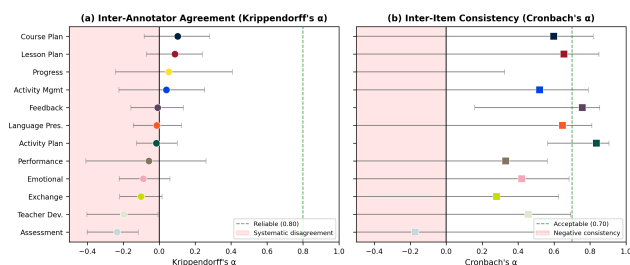


Figure 3: Side-by-side panels showing (a) Krippendorff's  $\alpha$  (criterion IAA) and (b) Cronbach's  $\alpha$  (criterion IIC) by competency for criteria scores with 95% CI.

more light on this phenomenon (Future Work). At the same time, we felt it important to report our negative results against criteria to demonstrate to those working in allied fields that they should exercise caution when transferring rubric-based evaluations methodologies across domains. Ultimately, we think this also factors into the misalignment between poor criteria IAA on highly-rated criteria IIC items.

The "Giving Feedback" competency warrants mention as it illustrates the distinction between construct validity problems and evaluator calibration problems. This competency showed the highest score variance ( $SD = 1.13$ ) and the second-lowest mean criteria score ( $M = 3.77$ ) despite high authenticity ( $M = 4.32$ ). Yet it exhibited the second-highest Cronbach's alpha ( $\alpha = 0.76$ ) alongside negative Krippendorff's alpha ( $\alpha = -0.01$ ), following the overall trend that tasks within "Giving Feedback" measure a coherent construct, but that evaluators apply systematically different standards when judging feedback criteria.

Nevertheless, despite poor criteria IAA scores, moderate to excellent criteria IIC scores from an adequate sample of one of our core stakeholder groups (postgraduate learners) provide **promising early signal that the L2-Bench construct is sound**. We propose future and larger-scale validation work to confirm these findings (Future Work).

## Iteration

Beyond quantitative analysis, teams produced written reports identifying systematic issues across L2-Bench components. Prior to the scaling up to the full 1,000+ task dataset, we may use this feedback to make revisions in three areas:

**Competency taxonomy.** Several sub-competencies were identified as oversimplified. In other areas, practitioners observed "inter-competency" elements: capabilities that appear important in multiple competencies. We will determine whether these represent distinct competencies or cross-cutting capabilities requiring separate treatment. We hypothesize that the method (of producing a taxonomy based on language-specific pedagogies) will scale to other pedagogies. However, we also believe each of these is, ultimately, an empirical question and one we will consider in future work.

**Evaluation criteria.** Evaluators noted that some criteria are more abstract or more context-sensitive than others; we plan to implement variable consensus criteria weightings based on task context and expand universal criteria to address sensitivities that vary by geographic and cultural context.

**Task design.** We will expand on our task variable approach for systematic task design, creating a framework of the context dimensions that impact on language learning tasks and responses – from linguistic context, to educational, resource availability and social factors. Additionally, we will standardise task wording for consistency where equivalent.

## Future Work

The tasks of a "learning experience designer in second language education" span multiple roles (teaching, content creation, assessment, learning design, professional development), but no single person or role is sufficiently skilled in all areas to validate these independently. Even if such a person existed, the dynamics of global pedagogy vary significantly from person to person, day to day, learning problem to learning problem. Therefore, we need data validation that brings together multiple stakeholder groups:

- **Language learning practitioners:** Experienced practitioners that design the conditions in which people learn
- **Language teachers:** L2 teachers, particularly those who engage with pedagogical position papers and research
- **Language assessment specialists:** Professionals who design and validate L2 assessment tests
- **Researchers:** Academic researchers in L2 education
- **EdTech professionals:** Professionals who design, implement or administrate applications for L2 education
- **Learners:** Advanced adult L2 learners for language fluency requirements and ethics considerations.

To this end, we will build on our pilot data validation to conduct a global "practitioner data validation" study, recruiting volunteers across the above stakeholder groups from our institution network whilst accounting for practical constraints in both representation (see Appendix B2) and ethical considerations (see Impact Statement). Practitioners will be assigned to "practitioner groups" (groups) matching their expertise through a short pre-screening survey, and these groups are then mapped to primary competencies from the L2-Bench competency construct which will be used to filter which tasks they will validate (see Appendix B2 Table 8 for group definitions and competency coverage). This stratified approach ensures specialist coverage for challenging competencies, while enabling cross-validation by practitioners with diverse professional perspectives.

The practitioner data validation will involve the full L2-Bench dataset of 1,300 task-response pairs (1,000 pairs to be released, 250 hold-out pairs for auto-scorer optimisation and benchmark saturation detection, and allowing for up to 50 pairs that may be excluded following validation (see Appendix B.5)), addressing three primary research objectives (ROs):

Table 3: Research objectives (RO), statistical methods, success criteria and power status ( $\checkmark$  = well-powered at 80%) for practitioner validation. IAA and IIC computed separately for authenticity and criteria scores.

RO	Analysis	Methods	Target	In Pilot	Power
RO1	Authenticity M	One-sample $t$ -test	$>4.0/5$	$\checkmark$	$\checkmark$
RO1	Criteria M	One-sample $t$ -test	$>3.5/5$	$\checkmark$	$\checkmark$
RO1	IAA	Fleiss' $\kappa$	$>0.20$	Kripp's $\alpha$	$\checkmark$
RO1	IIC	Cronbach's $\alpha$	$>0.70$		$\checkmark$
RO1	Auth vs Criteria gap	Wilcoxon signed-rank	—	$\checkmark$	$\checkmark$
RO1	Mixed effects model	Likelihood ratio test	$d=0.30, p < 0.05$	—	$\sim 45\%$
RO2	A/B preference	Binomial test	$\sim 70\%$ ref, $p < 0.05$	—	$\checkmark$
RO3	Inter-judge agreement	Cohen's $\kappa$	$>0.60$	—	$\checkmark$
RO3	Auto-scorer sensitivity	Recall	$\geq 0.80$	—	$\checkmark$
RO4	Group differences	Mann-Whitney $U$	$r=0.3$	—	$\sim 68-92\%$

**RO1. Measuring dataset validity** by establishing practitioner agreement on task authenticity and criteria adequacy.

**RO2. Measuring answer quality** by conducting blind comparisons of AI-generated responses to tasks versus our reference answers.

**RO3. Measuring auto-scorer validity** by collecting practitioner scores on "AI answers" against the task rubric and establishing inter-rater agreement.

We will also explore whether agreement varies systematically by practitioner expertise on competencies where we have diverse professional coverage (RO4).

While RO1 mirrors our approach to the pilot data validation (Pilot Validation Exercise), RO2 and RO3 are motivated by our desires to measure the reference answers produced in our task item production process (Components) and to explore the extent of practitioner bias against AI answers over reference answers. Moreover, the data collected for RO3 will be used to optimize our auto-scorer before dataset release.

For each task-response pair shown to the practitioner, the validation workflow proceeds sequentially through 2-3 stages:

1. practitioners rate task authenticity and criteria adequacy on 5-point Likert scales
2. they complete blind A/B comparisons where two answers – one AI-generated, one reference – are presented in randomized order; practitioners indicate preference with optional comments, and
3. after revealing the "AI answer" (which is randomized to be either the AI response or our reference answer), some validators score the "AI answer" against the task rubric on each criterion as Pass/Fail with optional comments, enabling measurement of inter-rater agreement with the auto-scorer.

The AI response will be generated by a single frontier model using the model provider's recommended default settings.

To achieve well-powered analysis, we seek 5 validity raters per task (Part 1+2) and 3 judge raters per task (Part 3). We analyze results using Fleiss' Kappa for inter-annotator agreement (IAA), Cronbach's Alpha for internal consistency

(IIC), Cohen's Kappa for inter-judge agreement (IJA), and mixed-effects models to test whether ratings differ systematically across competencies while controlling for rater and item effects. Table 2 summarizes our research objectives, statistical methods, and success criteria (full statistical methods and power analysis are detailed in Appendix B.3).

Based on pilot results, we anticipate authenticity ratings exceeding  $M = 4.0/5$  and criteria adequacy exceeding  $M = 3.5/5$ , strong  $IIC > 0.70$ , and variable Fleiss  $IAA > 0.20$  across competencies - pedagogical quality is contingent upon many factors, including teaching philosophy and experience, learner needs, and the learning problem in question.

We hypothesize reference answers will be systematically preferred over AI answers when that is true ( $\sim 70\%$  preference for reference), but when the AI answer is disguised as a Reference answer, we suspect a smaller systematic preference for the "Reference answer", accounting for practitioner bias against AI answers. Furthermore, we set a pragmatic target of inter-judge agreement (IJA) for fair agreement ( $\kappa > 0.20$ ), acknowledging the limitations of our initial reference-guided auto-scorer, recognising that many scenarios in language learning lack clear "correct" answers. For instance, when giving feedback, multiple lexical or semantic formulations may be acceptable, counterfactual learning techniques might be employed, and appropriacy will depend on unstated contextual factors (Knight et al. 2026), as well as accounting for practitioner fatigue (Yan 2025).

Given the substantial practitioner time commitment required to power our study ( $\sim 2,400$  hours), we document risk mitigation strategies in Appendix B.4, including over-recruitment buffers and reduced-coverage fallback designs that preserve statistical validity. In addition to validating L2-Bench, we note that we also use the results from the practitioner study to apply a task-response pair exclusion protocol (Appendix B.5) to remove validation outliers before the final dataset release (Truong et al. 2025).

## Conclusion

We have introduced L2-Bench, an evaluation benchmark intended to assess AI system capabilities for performing tasks entailed in quality learning experience design in second language education. Our paper primarily reported our methodology, detailed key artefacts (e.g. our taxonomy, measures,

and aspects of our pilot dataset) and validation exercises in the hopes that these methods can be of broader use to the evaluations community as well as those developing AI systems for educational contexts, including but not limited to language learning. Ultimately, our work demonstrates the feasibility of creating rigorous, scalable evaluations that bridge AI capabilities with learning science theory, contributing to efforts within the AI evaluation community to move beyond narrow accuracy metrics toward more rigorous, context-sensitive assessment of AI capabilities.

## Acknowledgements

We would like to thank all participants of the University of Birmingham ShapeAI Challenge for their hard work and valuable contributions to AI research. We would like to particularly thank the winners and runner-ups of the challenge: Venkata Vyjayanthi Pedapati (Vy), Yernur Niyetkaliyev, Aparajitha Magnesh, Manh Nguyen (Leo), Niamh Evans, Hsin-Yun Ho (Sydney), Sofia Muñoz, Saniya Saheer, Taiki Shimosakai, Yang Yu.

We would also like to thank Dr. Liam Knight for his help in bringing the challenge from idea to reality, and for his tireless support in facilitating the challenge and assisting all participants for over 6 weeks.

## Impact Statement

### IS1. Human Subjects Research for Data Validation

Both the pilot validation study and the forthcoming practitioner validation were designed in accordance with established research ethics principles. Participation is voluntary with explicit right to withdraw at any stage without penalty. Informed consent covers study purpose, procedures, time commitment, data handling, and anonymisation. All data collection complies with UK GDPR requirements; ratings are anonymised before analysis and stored securely following institutional data governance policies with 5-year post-publication retention. A formal legal agreement between participating institutions and/or individuals establishes clear provisions for intellectual property, confidentiality (5-year non-disclosure with specific AI tool restrictions to prevent data contamination that could compromise benchmark validity), and data protection.

Both studies are structured as voluntary programmes. The pilot study was conducted in partnership with a UK university's careers network team as a voluntary challenge programme designed to provide meaningful professional development with non-monetary incentives, and a prize incentive for the winning team. The practitioner study will be conducted predominantly on a voluntary research basis with non-monetary incentives. All incentives are subject to compliance review to ensure fair recognition for participant contributions.

The practitioner study additionally employs a blind A/B comparison protocol where participants compare two answers – one AI-generated, one reference – to which we then reveal the "AI answer" (which is randomized to be either the AI response or the reference answer). This methodologically necessary blinding is mitigated through full debriefing after

completion, scientific rationale for the design (Future Work), and the opportunity to withdraw data post-debrief. We also address practitioner burden through flexible scheduling and clear time-per-task communication.

### IS2. Broader Impacts

We anticipate this work will contribute positively to the AI evaluation ecosystem by providing an open-source benchmark and methodology that enables more rigorous assessment of AI capabilities in educational contexts.

We remain attentive to several concerns. First, benchmarks assessing AI capabilities in education could, in principle, be repurposed to evaluate human educators; we note explicitly that our work focuses solely on AI system assessment and we have no products or interests in teacher evaluation. Second, our reliance on predominantly European frameworks (CEFR, Equals) may embed cultural assumptions that limit generalisability; we plan to partner with regional language education organisations to address this in future iterations. Third, leaderboard rankings could inadvertently incentivise benchmark-specific optimisation rather than genuine pedagogical improvement.

Unanticipated consequences may arise from applications of our competency taxonomy or evaluation methodology in ways we have not foreseen. We encourage researchers building on this work to consider the potential for dual-use applications, to examine their own assumptions about pedagogical quality, and to implement appropriate safeguards when deploying evaluation frameworks in educational contexts. Likewise, we are actively exploring how our benchmark can evolve to account for multi-turn nature of authentic pedagogical relationships.

Once we are satisfied with the effectiveness of L2-Bench as described here, we plan to expand the benchmark itself. Planned expansion includes (1) increasing the number of tasks to maintain task representativeness and depth and establishing regular update intervals; (2) exploring how best to include support for multi-turn interactions; (3) expanding coverage to image and audio modalities (which are common in language learning design); and (4) including more globally diverse practitioner communities and language education frameworks. We intend to achieve this latter goal primarily through leveraging extensive institutional relationships with communities of practice (broadly construed) across the world.

We also note that L2-Bench is the first evaluation benchmark in a planned AI-for-education evaluation ecosystem. We dwell on our evaluation methodology because we intend to scale our methods to support evaluation in other domains for which usage data indicates evaluation needs are particularly acute—for instance, in legal pedagogy.

### IS3. Generative AI Statement

Generative AI tools were used in several aspects of this research.

For core methodology, Claude models (Anthropic) via Amazon Bedrock API served three functions: (1) Claude Sonnet-4.1 and Sonnet-4.5 with Claude Code initially generated candidate tasks, criteria, and reference answers dur-

ing hybrid human-AI dataset authoring (Components); (2) Claude Sonnet-3.7 generated AI responses for pilot validation (Methods); and (3) Claude Sonnet-4.5 with extended thinking serves as the auto-scoring foundation model (Components). All AI-generated content underwent expert review and validation.

For research support, GenAI assisted with: reviewing experimental design and identifying methodological improvements; research on statistical methods and their implementation; reviewing data processing pipelines and analysis iterations; and iterating on data visualisations.

For manuscript preparation, GenAI assisted with: LaTeX table and equation formatting, grammar and spelling review, and website development that resulted in figure creation.

All substantive research decisions, interpretations, and conclusions remain solely the responsibility of the authors.

## References

- Abbas, A.; Waggoner, C.; and Olive, J. 2025. Developing and Maintaining an Open-Source Repository of AI Evaluations: Challenges and Insights. arXiv:2507.06893.
- Arora, R. K.; Wei, J.; Hicks, R. S.; Bowman, P.; Quiñonero-Candela, J.; Tsimpourlas, F.; Sharman, M.; Shah, M.; Vallone, A.; Beutel, A.; Heidecke, J.; and Singhal, K. 2025. HealthBench: Evaluating Large Language Models Towards Improved Human Health. arXiv:2505.08775.
- Austin, J. L. 1975. *How to Do Things with Words*. Oxford, UK: Oxford University Press, 2 edition.
- Bahari, A.; Han, F.; and Strzelecki, A. 2025. Integrating CALL and AIALL for an Interactive Pedagogical Model of Language Learning. *Education and Information Technologies*, 30: 14305–14333.
- Bandura, A. 1977. *Social Learning Theory*. Englewood Cliffs, NJ: Prentice-Hall.
- Bastani, H.; Bastani, O.; Sungu, A.; Ge, H.; Kabakci, O.; and Mariman, R. 2024. Generative AI Can Harm Learning. SSRN Working Paper, DOI: 10.2139/ssrn.4895486.
- Bean, A. M.; Kearns, R. O.; Romanou, A.; Hafner, F. S.; Mayne, H.; Batzner, J.; Foroutan, N.; Schmitz, C.; Korgul, K.; Batra, H.; Deb, O.; Beharry, E.; Emde, C.; Foster, T.; Gausen, A.; Grandury, M.; Han, S.; Hofmann, V.; Ibrahim, L.; Kim, H.; Kirk, H. R.; Lin, F.; Liu, G. K.-M.; Luettgau, L.; Magomere, J.; Rystrøm, J.; Sotnikova, A.; Yang, Y.; Zhao, Y.; Bibi, A.; Bosselut, A.; Clark, R.; Cohan, A.; Foerster, J.; Gal, Y.; Hale, S. A.; Raji, I. D.; Summerfield, C.; Torr, P. H. S.; Ududec, C.; Rocher, L.; and Mahdi, A. 2025. Measuring what Matters: Construct Validity in Large Language Model Benchmarks. arXiv:2511.04703.
- Blanco, C. 2025. 2025 Duolingo Language Report. Technical report, Duolingo.
- British Council. 2025. Teaching for Success: Continuing Professional Development (CPD) for Teachers. <https://www.teachingenglish.org.uk/professional-development/teachers>. Accessed: 2025.
- Butler, J.; Vorvoreanu, M.; Janßen, R.; Sellen, A.; Immorlica, N.; Hecht, B.; and Teevan, J. 2024. Microsoft New Future of Work Report 2024. Technical Report MSR-TR-2024-56, Microsoft Research.
- Chapelle, C. A. 2001. *Computer Applications in Second Language Acquisition*. Cambridge: Cambridge University Press.
- Clark, H.-B.; Dowland, M.; Benton, L.; Budai, R.; Keskin, I. K.; Searle, E.; Gregory, M.; Hodiernne, M.; Gayne, W.; and Roberts, J. 2025. Auto-Evaluation: A Critical Measure in Driving Improvements in Quality and Safety of AI-Generated Lesson Resources. Technical report, The AI + Open Education Initiative.
- Clark, J. H.; Choi, E.; Collins, M.; Garrette, D.; Kwiatkowski, T.; Nikolaev, V.; and Palomaki, J. 2020. TyDi QA: A Benchmark for Information-Seeking Question Answering in Typologically Diverse Languages. *Transactions of the Association for Computational Linguistics*, 8: 454–470.
- Costa-Gomes, B.; Chen, S.; Hsueh, C.-H.; Morgan, D.; Schoenegger, P.; Shah, Y.; Way, S.; Zhu, Y.; Spielman, S.; Suleyman, M.; and Bhaskar, M. 2025. It’s about time: The Copilot Usage Report 2025: The Temporal and Modal Dynamics of Copilot Usage. Preprint.
- Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Council of Europe.
- DeKeyser, R.; and Suzuki, Y. 2025. *Skill Acquisition Theory*, 26. London: Routledge, 4th edition.
- Digital Education Council. 2024. Digital Education Council Global AI Student Survey 2024. Online report. Published August 2, 2024.
- European Association for Quality Language Services. 2016. The Equals Framework for Language Teacher Training and Development. Accessed: 2016.
- Feffer, M.; Sinha, A.; Deng, W. H.; Lipton, Z. C.; and Heidari, H. 2025. *Red-Teaming for Generative AI: Silver Bullet or Security Theater?*, 421–437. AAAI Press.
- Jurenka, I.; Kunesch, M.; McKee, K. R.; Gillick, D.; Zhu, S.; Wiltberger, S.; Phal, S. M.; Hermann, K.; Kasenberg, D.; Bhoopchand, A.; Anand, A.; Pîslar, M.; Chan, S.; Wang, L.; She, J.; Mahmoudieh, P.; Rysbek, A.; Ko, W.-J.; Huber, A.; Wiltshire, B.; Elidan, G.; Rabin, R.; Rubinovitz, J.; Pitaru, A.; McAllister, M.; Wilkowski, J.; Choi, D.; Engelberg, R.; Hackmon, L.; Levin, A.; Griffin, R.; Sears, M.; Bar, F.; Mesar, M.; Jabbour, M.; Chaudhry, A.; Cohan, J.; Thiagarajan, S.; Levine, N.; Brown, B.; Gorur, D.; Grant, S.; Hashimshoni, R.; Weidinger, L.; Hu, J.; Chen, D.; Dolecki, K.; Akbulut, C.; Bileschi, M.; Culp, L.; Dong, W.-X.; Marchal, N.; Deman, K. V.; Misra, H. B.; Duah, M.; Ambar, M.; Caciularu, A.; Lefdal, S.; Summerfield, C.; An, J.; Kamieny, P.-A.; Mohdi, A.; Strinopoulos, T.; Hale, A.; Anderson, W.; Cobo, L. C.; Efron, N.; Ananda, M.; Mohamed, S.; Heymans, M.; Ghahramani, Z.; Matias, Y.; Gomes, B.; and Ibrahim, L. 2024. Towards Responsible Development of Generative AI for Education: An Evaluation-Driven Approach. arXiv:2407.12687.
- Kennedy, W. M.; and Vargas Campos, D. 2024. Vernacularizing Taxonomies of Harm Is Essential for Operationalizing

- Holistic AI Safety. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 698–710.
- Kennedy, W. M.; and Vargas Campos, D. 2026. A Vernacularized Taxonomy of Harms for AI in Education. In Holmes, W.; and Pelletier, C., eds., *Handbook of Critical Studies in AI for Education*. Edward Elgar. Forthcoming.
- Knight, B.; Kennedy, W. M.; Carvalho, D.; Pattis, I.; and Edgell, J. 2026. Ceci n'est pas une explication: Evaluating Explanation Failures as Explainability Pitfalls in Language Learning Systems. arXiv:2604.26145.
- Krippendorff, K. 2004. *Content Analysis: An Introduction to Its Methodology*. Thousand Oaks, CA: Sage Publications, 2 edition.
- LearnLM Team; and Google. 2025. Evaluating Gemini in an Arena for Learning. ArXiv:2505.24477v1 [cs.CY].
- Lelièvre, M.; Waldock, A.; Liu, M.; Aspillaga, N. V.; Mackintosh, A.; Portela, M. J. O.; Lee, J.; Atherton, P.; Ince, R. A. A.; and Garrod, O. G. B. 2025. Benchmarking the Pedagogical Knowledge of Large Language Models. arXiv:2506.18710.
- Liu, Y.; Zhou, H.; Guo, Z.; Shareghi, E.; Vulić, I.; Korhonen, A.; and Collier, N. 2024. Aligning with Human Judgement: The Role of Pairwise Preference in Large Language Model Evaluators. In *First Conference on Language Modeling*.
- Miller, E. 2024. Adding Error Bars to Evals: A Statistical Approach to Language Model Evaluations. arXiv:2411.00640.
- Nie, A.; Chandak, Y.; Suzara, M.; Malik, A.; Woodrow, J.; Peng, M.; Sahami, M.; Brunskill, E.; and Piech, C. 2025. The GPT Surprise: Offering Large Language Model Chat in a Massive Coding Class Reduced Engagement But May Increase Adopters' Exam Performances. In *Proceedings of the Twelfth ACM Conference on Learning @ Scale, L@S '25*, 376–380. New York, NY, USA: Association for Computing Machinery. ISBN 9798400712913.
- Olteanu, A.; Blodgett, S. L.; Balayn, A.; Wang, A.; Diaz, F.; du Pin Calmon, F.; Mitchell, M.; Ekstrand, M.; Binns, R.; and Barocas, S. 2025. Rigor in AI: Doing Rigorous AI Work Requires a Broader, Responsible AI-Informed Conception of Rigor. arXiv:2506.14652.
- Papi, M.; and Khajavy, G. H. 2023. Second Language Anxiety: Construct, Effects, and Sources. *Annual Review of Applied Linguistics*, 43: 127–139.
- Pardos, Z. A.; and Bhandari, S. 2023. Learning gain differences between ChatGPT and human tutor generated algebra hints. arXiv:2302.06871.
- Poehner, M. E.; and Lantolf, J. P. 2024. *Sociocultural Theory and Second Language Developmental Education*. Cambridge: Cambridge University Press.
- Preston, C. C.; and Colman, A. M. 2000. Optimal Number of Response Categories in Rating Scales: Reliability, Validity, Discriminating Power, and Respondent Preferences. *Acta Psychologica*, 104(1): 1–15.
- Reuel, A.; Hardy, A.; Smith, C.; Lamparth, M.; Hardy, M.; and Kochenderfer, M. J. 2024. BetterBench: Assessing AI Benchmarks, Uncovering Issues, and Establishing Best Practices. arXiv:2411.12990.
- Schwartz, R.; Chowdhury, R.; Kundu, A.; Frase, H.; Fadaee, M.; David, T.; Waters, G.; Taik, A.; Briggs, M.; Hall, P.; Jain, S.; Yee, K.; Thomas, S.; Bhandari, S.; Duncan, P.; Thompson, A.; Carlyle, M.; Lu, Q.; Holmes, M.; and Skeadas, T. 2025. Reality Check: A New Evaluation Ecosystem Is Necessary to Understand AI's Real World Effects. arXiv:2505.18893.
- Searle, J. R. 1996. What is Language : Some Preliminary Remarks. In Giovagnoli, R., ed., *Etica E Politica*, 173–202. Clarendon Press.
- Shetye, S. 2024. An Evaluation of Khanmigo, a Generative AI Tool, as a Computer-Assisted Language Learning App. *Studies in Applied Linguistics & TESOL*, 24(1): 38–53.
- Smart, A.; Hutchinson, B.; Amugongo, L. M.; Dikker, S.; Zito, A.; Ebinama, A.; Wudiri, Z.; Wang, D.; van Liemt, E.; Sedoc, J.; Olojo, S.; Uwakwe, S.; Wornyo, E.; Schmergalunder, S.; and Smith-Loud, J. 2024. Socially Responsible Data for Large Multilingual Language Models. arXiv:2409.05247.
- Solaiman, I.; Talat, Z.; Agnew, W.; Ahmad, L.; Baker, D. K.; Blodgett, S. L.; Chen, C.; Daumé, I.; Hal; Dodge, J.; Duan, I.; Evans, E.; Friedrich, F.; Ghosh, A.; Gohar, U.; Hooker, S.; Jernite, Y.; Kalluri, P. R.; Leidinger, A.; Lusoli, A.; Lin, M.; Lin, X.; Luccioni, S.; Mickel, J.; Mitchell, M.; Newman, J.; Ovalle, A.; Png, M.-T.; Singh, S.; Strait, A.; Struppek, L.; Subramonian, A.; and Vassilev, A. ????. Evaluating the Social Impact of Generative AI Systems. In Hacker, P.; Engel, A.; Hammer, S.; and Mittelstadt, B., eds., *The Oxford Handbook of the Foundations and Regulation of Generative AI*. Oxford University Press. ISBN 9780198940272.
- Spiliopoulou, E.; Fogliato, R.; Burnsky, H.; Soliman, T.; Ma, J.; Horwood, G.; and Ballesteros, M. 2025. Play Favorites: A Statistical Method to Measure Self-Bias in LLM-as-a-Judge. arXiv:2508.06709.
- Tamkin, A.; McCain, M.; Handa, K.; Durmus, E.; Lovitt, L.; Rathi, A.; Huang, S.; Mountfield, A.; Hong, J.; Ritchie, S.; Stern, M.; Clarke, B.; Goldberg, L.; Summers, T. R.; Mueller, J.; McEachen, W.; Mitchell, W.; Carter, S.; Clark, J.; Kaplan, J.; and Ganguli, D. 2024. Clio: Privacy-Preserving Insights into Real-World AI Use. <https://arxiv.org/abs/2412.13678>. ArXiv:2412.13678.
- Team, L.; Eedi, ; Wang, A.; Rysbek, A.; Huber, A.; Nambiar, A.; Kenolty, A.; Caulfield, B.; Lilley-Draper, B.; Groot, B.; Veprek, B.; Burdett, C.; Willis, C.; Barton, C.; Smith, D.; Mu, G.; Walters, H.; Jurenka, I.; Hulls, I.; Stalley-Moores, J.; Caton, J.; Wilkowski, J.; Alarakyia, K.; McKee, K. R.; McCafferty, L.; Dalton, L.; Kunesch, M.; Malubay, P.; Kidson, R.; Wells, R.; Wheeler, S.; Wiltberger, S.; Mohamed, S.; Woodhead, S.; and Brazão, V. 2025. AI tutoring can safely and effectively support students: An exploratory RCT in UK classrooms. arXiv:2512.23633.
- Truong, S.; Tu, Y.; Hardy, M.; Reuel, A.; Tang, Z.; Burapachep, J.; Perera, J.; Uwakwe, C.; Domingue, B.; Haber, N.; and Koyejo, S. 2025. Fantastic Bugs and Where to Find Them in AI Benchmarks. arXiv:2511.16842.
- UK AI Safety Institute. 2024. Elicitation of AI responses protocol. Technical report, UK AI Safety Institute.

UK AI Safety Institute. 2025. LLM Judges on Trial: A New Statistical Framework to Assess Autograders. Accessed January 2026.

UK AI Security Institute. 2024. Early Insights from Developing Question-Answer Evaluations for Frontier AI. <https://www.aisi.gov.uk/blog/early-insights-from-developing-question-answer-evaluations-for-frontier-ai>. Accessed: 2026-03-17.

Weidinger, L.; Raji, I. D.; Wallach, H.; Mitchell, M.; Wang, A.; Salaudeen, O.; Bommasani, R.; Ganguli, D.; Koyejo, S.; and Isaac, W. 2025. Toward an Evaluation Science for Generative AI Systems. arXiv:2503.05336.

World Conference on Linguistic Rights. 1996. Universal Declaration of Linguistic Rights.

Xu, B.; Bai, Y.; Sun, H.; Lin, Y.; Liu, S.; Liang, X.; Li, Y.; Gao, Y.; and Huang, H. 2025. EduBench: A Comprehensive Benchmarking Dataset for Evaluating Large Language Models in Diverse Educational Scenarios. arXiv:2505.16160.

Yan, Z. 2025. Product Evals in Three Simple Steps. <https://eugeneyan.com/writing/product-evals/>. Accessed January 2026.

## Appendices

### Pilot Study

#### Statistical Methods

Our pilot analysis first involved extracting raw scores from the Langfuse annotation platform and converting them to 1–5 Likert scale (where 1=Strongly Disagree through 5=Strongly Agree). We then employed statistical methods appropriate for ordinal Likert-scale data and sparse rating matrices:

**Inter-annotator agreement (IAA)** We used Krippendorff’s Alpha to compute IAA rather than Fleiss’s Kappa because it handles missing data (albeit limiting precision):

$$\alpha = 1 - \frac{D_o}{D_e} \quad (1)$$

where  $D_o$  is observed disagreement and  $D_e$  is expected disagreement under chance; for ordinal data, disagreement weights are  $(i - j)^2$  for categories  $i$  and  $j$ .

Alpha was computed to respect the ordinal nature of Likert scales. Items required ratings from  $N \geq 2$  evaluators for inclusion. Alpha can be strongly affected by prevalence/skew: in our study, the marginal distribution of ratings was highly concentrated with most ratings clustering around 4–5 (see Appendix A.3), so we expect some depression of Krippendorff’s alpha (modest disagreement appears large relative to the low expected chance disagreement baseline). Our confidence intervals should be interpreted cautiously as an uncertainty heuristic rather than a theoretically exact interval since our implementation differs from Krippendorff’s proposed bootstrap (see below).

**Internal-item consistency (IIC)** We used Cronbach’s alpha using the standard variance formula to compute ICC:

$$\alpha = \frac{k}{k - 1} \left( 1 - \frac{\sum_{i=1}^k \sigma_i^2}{\sigma_T^2} \right) \quad (2)$$

where  $k$  is the number of items,  $\sigma_i^2$  is the variance of item  $i$ , and  $\sigma_T^2$  is the variance of total scores.

Due to high missing data (~50% across competencies), and to avoid dropping items that would change the construct, we filtered to items with  $N \geq 3$  raters and used mean imputation for remaining missing values. We acknowledge this may bias alpha upward (mean imputation reduces item variances which can inflate inter-item correlations). However, we note that standard Cronbach’s alpha assumes interval-level data, and we would therefore expect ordinal-appropriate methods (e.g. Spearman instead of Pearson correlations) would yield slightly higher alpha estimates. Finally, Cronbach’s alpha assumes exchangeable respondents, but team-level effects may violate this assumption, meaning IIC confidence intervals should be treated as heuristic rather than exact. In any case, given our overall alpha of 0.95 already indicates excellent reliability, this methodological limitation does not affect our substantive conclusions about internal consistency.

**Confidence intervals (CI)** For statistics without known sampling distributions (e.g. Krippendorff’s  $\alpha$ , Cronbach’s  $\alpha$ ), we used bootstrap resampling ( $n = 500$  iterations, percentile method) to provide a heuristic uncertainty band:

For statistic  $\theta$ , resample items  $B$  times with replacement, compute  $\theta_b^*$  for each resample:

$$\text{CI} = [\theta_{(0.025)}^*, \theta_{(0.975)}^*] \quad (3)$$

For descriptive statistics (means), we used parametric t-distribution CIs given their known asymptotic properties.

**Group comparisons** All inferential tests used non-parametric methods appropriate for ordinal data: Wilcoxon signed-rank for paired comparisons (authenticity vs criteria scores):

$$W = \sum_{i: d_i > 0} R_i \quad (4)$$

with  $d_i$  being the paired difference and  $R_i$  the rank of  $|d_i|$ .

For interpretability, we report effect sizes as Cohen’s  $d$  for paired samples (acknowledging that  $d$  incorrectly assumes interval data):

$$d = \frac{\bar{d}}{s_d} \quad (5)$$

## Practitioner Data Validation

### Design Parameters

Table 7 summarizes the key design parameters for the practitioner data validation. The design targets 100% dataset coverage to enable comprehensive validity assessment and well-powered statistical analysis (see Table 3 and Appendix B.3).

Table 4: Validation design parameters. Valid=Validity scoring tasks; IJA=Inter-judge agreement tasks; /Task=Raters per task; Time=Time per task; Hours=Total hours; Days=Person-days (7h/day);  $N$ =Practitioners required; Tasks/Comp=Tasks per competency.

Dataset	Valid	IJA	Valid/Task	IJA/Task	Valid Time	IJA Time	Hours	Days	$N$	Tasks/Comp
1,300	1,300	1,300	5	3	~15 min	~10 min	~2,389h	~342	~710	~108

### Practitioner Group Definitions

Table 8 summarizes practitioner group definitions, where practitioners are recruited from institutional networks and global practitioner communities to ensure diverse professional perspectives and geographic representation.

Practitioner counts ( $N$ ) are derived from the total hours required for well-powered analysis (~2,389h, see Appendix B.1), divided by expected time commitment per practitioner. Time estimates vary by group accounting for role expertise to competency mapping, professional availability and pilot findings:

- We expect classroom teachers (Group C) contribute ~1 hour given teaching schedules, while content specialists and learners (Groups A, B, F) contribute ~7 hours given their professional focus on materials development.
- We apply a ~15% over-recruitment buffer to account for expected dropout and scheduling variability.
- Time estimates are also informed by pilot findings (Methods), adjusted for expected efficiency gains from (a) domain expertise enabling faster task completion, and (b) improved annotation interface design.

Table 5: Practitioner group definitions. Hours per person vary by professional availability and role focus.

Group	Description	$N$	Hrs/Pers	Hours	Competencies
A: Content	Materials developers, curriculum designers	~150	~7.0h	~1,055h	01, 02, 03, 04, 05, 08
B: Assessment	Test developers, evaluation experts	~18	~7.0h	~125h	06, 07, 09, 11
C: Teaching	Active classroom teachers	~60	~1.0h	~60h	01–10, 12
D: Generalists	EdTech professionals, administrators	~360	~1.4h	~508h	06, 08
E: Academic	Researchers, teacher trainers	~60	~3.5h	~211h	05, 06, 07, 09, 10, 12
F: Learner	Advanced L2 users	~62	~7.0h	~432h	06, 08
<b>Total</b>		<b>~710</b>		<b>~2,389h</b>	

Table 9 shows competency coverage mapping, indicating which practitioner groups serve as primary and backup reviewers for each competency. Competencies 07 and 09 have expanded coverage through Teaching and Academic groups; only competency 11 requires assessment specialists exclusively.

Table 6: Competency coverage mapping by reviewer group.

Competency	Primary	Backup
01–03. Planning	A, C	D
04. Running Activities	A, C	—
05. Language Learning	A, C, E	—
06. Exchange Partner	All	—
07. Performance Eval	C, B, E	A, D
08. Giving Feedback	A, C, D, F	—
09. Progress Tracking	C, B, E	A
10. Emotional Intel	C, E	A, D, F
11. Assessment Creation	B	A (trained)
12. Professional Dev	E, C	—

## Statistical Methods and Power Analysis

This section details our statistical framework for the practitioner data validation. Where methods overlap with the pilot study (Appendix A.1), we note key differences and refer to Appendix A.1 for foundational explanations.

**Inter-annotator agreement (IAA)** We use Fleiss' Kappa for IAA rather than Krippendorff's Alpha (used in the pilot, Appendix A.1) because the practitioner study achieves complete rating matrices within each task (5 raters per task for each rating type):

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e} \quad (6)$$

where  $\bar{P}$  is the mean proportion of agreeing rater pairs and  $\bar{P}_e = \sum_j p_j^2$  is expected agreement by chance.

**Internal item consistency (IIC)** As in the pilot (Appendix A.1), we use Cronbach's Alpha to assess whether tasks within each competency measure a coherent underlying construct.

**Confidence intervals (CI)** As in the pilot (Appendix A.1), for statistics without closed-form distributions (Fleiss'  $\kappa$ , Cronbach's  $\alpha$ ), we use bootstrap resampling to provide uncertainty estimates.

**Descriptive comparisons** As in the pilot (Appendix A.1), we use Wilcoxon signed-rank tests appropriate for ordinal data, reporting effect sizes as Cohen's  $d$  for interpretability while acknowledging this incorrectly assumes interval data.

**Inter-judge agreement (IJA)** We assess agreement between human practitioners and our auto-scorer using Cohen's Kappa (two rater agreement) on binary Pass/Fail decisions per criterion:

$$\kappa = \frac{P_o - P_e}{1 - P_e} \quad (7)$$

where  $P_o$  is observed agreement and  $P_e$  is expected agreement by chance.

**Auto-scorer sensitivity** We focus on auto-scorer sensitivity (recall) for detecting failures:

$$\text{Sensitivity} = \frac{TP}{TP + FN} \quad (8)$$

We prioritise sensitivity over precision because false negatives (auto-scorer passes a poor AI response) risk exposing learners to inadequate pedagogical content, whereas false positives (auto-scorer fails an acceptable AI response) result in conservative flagging that can be resolved through human review without learner impact.

**A/B preference analysis** For blind answer comparisons, we use a binomial test (one tailed) against the null hypothesis of no preference ( $p = 0.50$ ):

$$p\text{-value} = \sum_{x=k}^n \binom{n}{x} p_0^x (1 - p_0)^{n-x} \quad (9)$$

where  $k$  is observed reference preferences,  $n$  is total trials, and  $p_0 = 0.50$ .

**Mixed effects model for competency comparisons** To test whether ratings differ systematically across the 12 competencies, we fit a linear mixed effects model:

$$y_{ijk} = \beta_0 + \beta_c \cdot \text{comp}_j + u_i + v_k + \varepsilon_{ijk} \quad (10)$$

where  $y_{ijk}$  is the rating given by rater  $i$  on item  $k$  within competency  $j$ ,  $\beta_0$  is the grand intercept,  $\beta_c$  represents the fixed effect coefficients for competency (with one reference level),  $u_i \sim N(0, \sigma_u^2)$  is the random intercept for rater  $i$ ,  $v_k \sim N(0, \sigma_v^2)$  is the random intercept for item  $k$ , and  $\varepsilon_{ijk} \sim N(0, \sigma^2)$  is the residual error.

Significance is assessed via likelihood ratio test (LRT) comparing the full model (with competency as a fixed effect) against an intercept-only model:

$$\Lambda = -2 \ln \left( \frac{L_{\text{red}}}{L_{\text{full}}} \right) \sim \chi_{df}^2 \quad (11)$$

where  $L_{\text{red}}$  is the likelihood of the reduced model,  $L_{\text{full}}$  is the likelihood of the full model, and  $df$  is the difference in parameters.

We conduct post-hoc pairwise comparisons between all competency pairs (66 for our case of 12 competencies) using estimated marginal means from the fitted model. To control the family-wise error rate (FWER), we apply the Bonferroni correction:

$$\alpha_{\text{adj}} = \frac{\alpha}{m} \quad (12)$$

where  $\alpha$  is the nominal significance level and  $m$  is the number of comparisons. Equivalently,  $p_{\text{adj}} = p \times m$ .

**Group comparisons** To explore group comparisons (such as whether ELT specialists rate differently from generalists), we use Mann-Whitney U tests on aggregated per-rater mean scores in order to preserve the assumption of independent observations, reporting the rank-biserial correlation ( $r$ ) as the appropriate non-parametric effect size measure for ordinal data:

$$r = \frac{2U}{n_1 n_2} - 1 \quad (13)$$

where  $U$  is the Mann-Whitney statistic.

**Power analysis** Power calculations ( $\alpha=0.05$ , target power=0.80) were used within self-imposed constraints of L2-bench dataset size and practitioner recruitment practicalities:

- Fleiss' Kappa ( $H_0: \kappa = 0.40$  vs  $H_1: \kappa = 0.60$ ) is well-powered with 103 items per competency with 3+ raters → exceeded
- Cronbach's Alpha ( $H_0: \alpha = 0.50$  vs  $H_1: \alpha = 0.70$ ) is well-powered with 85 subjects (6,500 total ratings) → exceeded
- Cohen's Kappa ( $\kappa = 0.60 \pm 0.15$ ) is well-powered with 100 items (1,300 tasks) → exceeded
- Recall (85%  $\pm$  8%) is well-powered with 150 items → exceeded
- Binomial preference (70% vs 50%) is well-powered with 38 items per competency → exceeded
- Mixed effect model analysis is powered (~45%) to detect small-medium effects ( $d \geq 0.30$ ) between competencies → caution needed
- Competency-level comparisons are powered (~45%) to detect small-medium effects ( $d \geq 0.30$ ) between competencies → caution needed
- Group comparisons (rank-biserial correlation  $r = 0.3$ ) achieve ~68-92% power depending on group sizes → treated as exploratory.

## Contingency Planning

If practitioner recruitment, retention or hours available falls below target levels, the design maintains statistical validity through several mechanisms:

1. Over-recruitment buffer: We recruit practitioners beyond minimum requirements to absorb expected dropout (~15% buffer), with reminder protocols to maintain engagement.
2. Specialist backup training: Content specialists are trained as backup validators for competencies requiring specialist expertise (particularly Assessment Creation), addressing potential bottlenecks in specialist availability.
3. Fallback sampling strategy: If tasks must be reduced to accommodate reduced hours and/or practitioners, tasks are stratified across competencies with priority given to maintaining balanced coverage even if it means reducing the total number of tasks per competency. Both validity and judge assessments will prioritize minimum rater coverage (3+ raters per task) to ensure well-powered statistics, even if total task coverage must be reduced. Therefore the design can fall back to reduced tasks per competency while maintaining minimum rater coverage, preserving reliable inter-rater agreement at the cost of reduced power for competency-level effects.

These contingencies preserve the core research objectives while adapting to real-world recruitment constraints.

## Task-Response Pair Exclusion Protocol

Following validation, we apply a two-stage exclusion process to apply a systematic exclusion protocol to identify and remove task-response pairs with poor validity signals, using multi-signal statistical flagging combined with expert review rather than rigid automated thresholds. We use deliberately conservative thresholds to flag only genuine validity issues to our expert reviews and avoid over-exclusion in the final open dataset:

**Stage 1: Statistical Flagging** We flag task-response pairs meeting any of the following criteria:

1. Low authenticity: Mean authenticity rating  $< 2.5/5$  (clear practitioner rejection)
2. Low criteria adequacy: Mean criteria adequacy rating  $< 2.5/5$  (clear practitioner rejection)
3. High within-task disagreement: Standard deviation  $> 2.0$  on 5-point scales (indicating systematic confusion about the task)
4. Anomalous A/B preference: Unanimous preference for "AI answer" across all raters (potential reference answer quality issue)
5. Rater annotations: Task flagged by  $\geq 2$  raters with substantive comments indicating systematic problems (e.g., cultural bias, ambiguous scenario, scoring criteria mismatch).

**Stage 2: Expert Review** Flagged items undergo expert review by the development team to determine final exclusion decisions. Experts assess:

- Whether the flagged issue reflects a genuine task problem vs. expected disagreement in pedagogically subjective domains, and if there is consistency with similar items in the same competency
- Whether the item can be revised rather than excluded.

All exclusion decisions are documented with justification, enabling transparency in the final dataset release. Excluded items are retained in a separate archive for methodological analysis.

## Competency Taxonomy

### Glossary of Terms

Table 7: Key terminology used in second language education and L2-Bench

Term	Explanation
L1, L2	L1 = first language or mother tongue; L2 = any additional language learned after the L1.
EAL, ESL	EAL = English as an Additional Language; ESL = English as a Second Language. Both terms are most commonly used when the learner resides in a country where English is the dominant language (e.g., the US or UK) but has a different L1.
EFL	EFL = English as a Foreign Language. Used when the learner is studying English in a context where it is not the dominant language (e.g., Spain or China).
ELT	ELT = English Language Teaching. Refers to the profession of teaching English and encompasses both EAL and EFL contexts.
Learning experience designer in second language education	In L2-Bench, this term encompasses roles that intentionally design the conditions shaping how people learn, including classroom and online teachers, materials developers (content or assessment creators), learning designers, and teacher trainers.
Language acquisition, language learning, language teaching	Language acquisition refers to the largely unconscious process of learning a language through immersion and applies to L1 acquisition and some L2 contexts. Language learning implies a more conscious effort, often supported by a teacher. Language teaching is the purposeful activity of helping a learner acquire the language.
Competencies	A competency is a combination of knowledge, skills, and attitudes required to perform a role or occupational function successfully. In L2-Bench, competencies refer to those required by teachers and other language education practitioners.
Communicative competence	An individual's ability to convey meaning effectively across contexts, encompassing grammatical, sociolinguistic, discourse, and strategic competence.
CEFR	The Common European Framework of Reference for Languages, developed by the Council of Europe and widely recognized as an international standard for language learning.
Target Language, Language of Instruction, English Language Teaching	The Target Language is the language the learner aims to learn; the Language of Instruction is the language used by the teacher. In English Language Teaching, English is always the Target Language, but it may or may not be the Language of Instruction. In L2-Bench, both the Target Language and Language of Instruction are set as English.

## Competencies, Sub-competencies and Consensus Criteria

The L2-Bench competency taxonomy comprises 12 competencies, each with sub-competencies and associated consensus criteria. For each competency below, we present the sub-competency breakdown and the consensus criteria grouped by sub-competency (weights in parentheses).

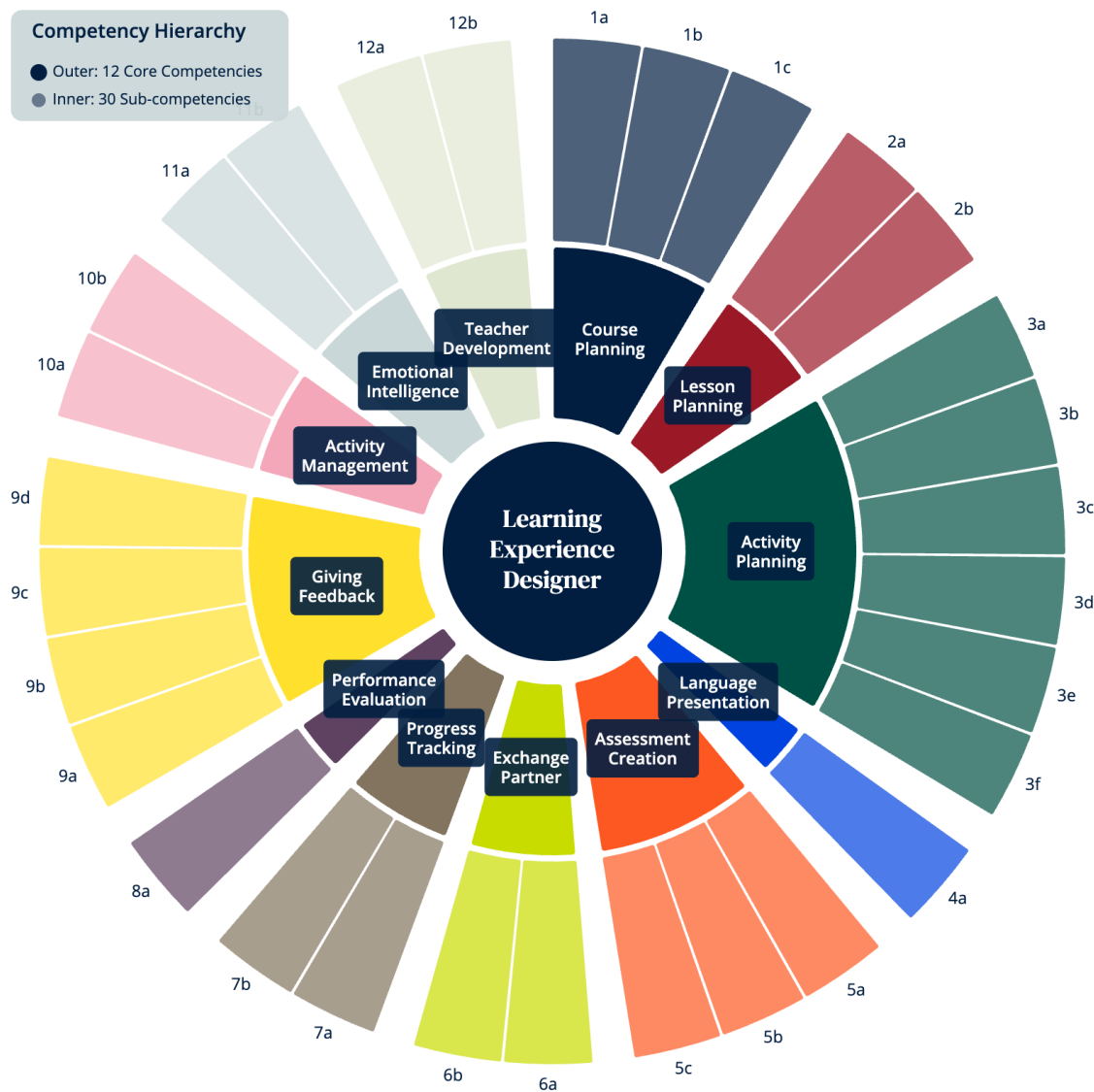


Figure 4: L2-Bench Taxonomy of Competencies - sunburst visualization showing the 12 competencies and 30 sub-competencies of a “learning experience designer in second language education”

### C01: Create a Course Plan Sub-competencies:

- **01a:** Decide which learning goals are most important for students’ aims, context, needs, interests
- **01b:** Organise learning goals into units and lessons
- **01c:** Decide on learning experience design

### Consensus Criteria:

For SC01a (Learning Goals):

- 01a-01: References students’ learning aim(s) (+7)
- 01a-02: References different areas of language learning (+7)

- 01a-03: Selection appropriate for time, level, pace (+7)
- 01a-04: References data from tests, reports, analytics (+4)
- 01a-05: References students' interests (+3)

*For SC01b (Organise Goals):*

- 01b-01: Goals progress in difficulty; complementary goals grouped (+4)
- 01b-02: Plan revisits earlier learning goals (+6)

*For SC01c (Experience Design):*

- 01c-01: Pace, recycling, assessment appropriate for context (+8)
- 

## **C02: Plan a Lesson Sub-competencies:**

- **02a:** Decide on sequence and types of activities for effective learning
- **02b:** Identify or create materials or resources needed

### **Consensus Criteria:**

*For SC02a (Activity Sequence):*

- 02a-01: Includes appropriate pattern (PPP, ESA, TBLT) (+5)
- 02a-02: Activities build knowledge/skills for goal (+6)
- 02a-03: Clear structure for student profile (+8)
- 02a-04: Instructions on setup/management (+9)
- 02a-05: Balances form with function (+6)
- 02a-06: Develops students' learning skills (+5)
- 02a-07: Realistic timings (+6)
- 02a-08: Scaffolding supports proficiency (+7)
- 02a-09: Activities engage students (+8)

*For SC02b (Materials/Resources):*

- 02b-01: Resources appropriate length (+7)
  - 02b-02: Texts realistic for level/context (+7)
- 

## **C03: Plan an Activity Sub-competencies:**

- **03a:** Decide on most suitable type of activity
- **03b:** Provide appropriate level of scaffolding
- **03c:** Identify or create materials or resources
- **03d:** Create key for evaluating responses
- **03e:** Provide instructions on running activity
- **03f:** Integrate with other lesson activities

### **Consensus Criteria:**

*For SC03a–03f:*

- 03a-01: Activity suits goal, stage, type, profile (+8)
  - 03b-01: Scaffolding reduces as learning progresses (+6)
  - 03c-01: Resources match request, appropriate for context (+8)
  - 03c-02: Multiple resources have coherent theme (+5)
  - 03d-01: Indicates accepted answers, rubrics, marks (+7)
  - 03e-01: Clear instructions with timings (+9)
  - 03f-01: Uses key language from previous activities (+6)
-

**C04: Manage Activities Within a Class Sub-competencies:**

- **04a:** Check instructions understood and followed
- **04b:** Organise learners into pairs/groups, assign roles

**Consensus Criteria:**

*For SC04a (Check Instructions):*

- 04a-01: Instructions clear; checking process exists (+7)

*For SC04b (Organise Learners):*

- 04b-01: Learners grouped by principles suiting task (+5)
- 

**C05: Present Language Learning Points Sub-competencies:**

- **05a:** Present language learning points effectively

**Consensus Criteria:**

*For SC05a (Present Language):*

- 05a-01: Language point explained clearly (+10)
  - 05a-02: Appropriate approach (inductive/deductive) (+4)
  - 05a-03: Activates prior knowledge first (+5)
  - 05a-04: Items in meaningful context (+7)
  - 05a-05: Covers meaning, use and form (+8)
  - 05a-06: Visual aids used when needed (+5)
  - 05a-07: Checks learner understanding (+5)
  - 05a-08: If inductive, questions help discovery (+8)
- 

**C06: Act as Conversational Exchange Partner Sub-competencies:**

- **06a:** Respond appropriately for role and context
- **06b:** Identify when learner struggles and respond

**Consensus Criteria:**

*For SC06a (Respond Appropriately):*

- 06a-01: Uses key language appropriately (+6)
- 06a-02: Responds in real time (+8)
- 06a-03: Responds to earlier conversation (+6)

*For SC06b (Identify Struggling):*

- 06b-01: Simplifies or uses L1 when learner struggles (+8)
- 

**C07: Evaluate a Student's Performance Sub-competencies:**

- **07a:** Assign evaluation (general to detailed marks)

**Consensus Criteria:**

*For SC07a (Assign Evaluation):*

- 07a-01: Evaluation accurate per criteria (+10)
  - 07a-02: Fits required detail level (+6)
-

**C08: Give Feedback Sub-competencies:**

- **08a:** Identify errors and diagnose causes
- **08b:** Prioritise areas needing feedback
- **08c:** Provide explanations, models, hints
- **08d:** Provide improvement activities

**Consensus Criteria:**

*For SC08a (Identify Errors):*

- 08a-01: Estimates likely causes of error (+5)

*For SC08b (Prioritise Feedback):*

- 08b-01: Feedback only for most important areas (+5)

*For SC08c (Provide Explanations):*

- 08c-01: Includes explanations/models/hints (+8)

*For SC08d (Improvement Activities):*

- 08d-01: Points to improvement activities (+7)
- 

**C09: Track Progress Sub-competencies:**

- **09a:** Collect data/samples of learning over time
- **09b:** Analyse progress patterns for learning goals

**Consensus Criteria:**

*For SC09a (Collect Data):*

- 09a-01: Data tagged for different learning goals (+8)

*For SC09b (Analyse Progress):*

- 09b-01: Insights derived from data (+10)
  - 09b-02: Analysis relates to recognised standards (+5)
- 

**C10: Manage Social-Emotional Aspects Sub-competencies:**

- **10a:** Identify/diagnose learner emotional status
- **10b:** Implement interventions for emotional issues

**Consensus Criteria:**

*For SC10a (Identify Emotions):*

- 10a-01: Process for monitoring emotions (+7)
- 10a-02: Able to identify learner emotions (+5)

*For SC10b (Implement Interventions):*

- 10b-01: Shows understanding and empathy (+7)
  - 10b-02: Raises awareness of self-efficacy (+5)
  - 10b-03: Develops self-regulated learning (+4)
  - 10b-04: Responds to emotional issues (+2)
- 

**C11: Create Assessments Sub-competencies:**

- **11a:** Decide on learning goals to assess
- **11b:** Decide on task types and organisation
- **11c:** Create a mark scheme

**Consensus Criteria:**

*For SC11a (Assessment Goals):*

- 11a-01: Goals appropriate for CEFR, aim, context (+7)

*For SC11b (Task Types):*

- 11b-01: Task types appropriate for goal (+10)
- 11b-02: Sequence practical for administration (+8)

*For SC11c (Mark Scheme):*

- 11c-01: Mark scheme with answers, rubrics, thresholds (+10)
  - 11c-02: Rubrics for spoken/written production (+7)
- 

## **C12: Support Professional Development Sub-competencies:**

- **12a:** Evaluate a teacher's activity
- **12b:** Provide advice/guidance on teaching

### **Consensus Criteria:**

*For SC12a (Evaluate Teacher):*

- 12a-01: References engagement, activities, management, materials, response to needs (+10)

*For SC12b (Provide Guidance):*

- 12b-01: Appropriate for teacher's experience (+10)

## **Universal Criteria**

Universal criteria are applied to **all** tasks in L2-Bench, with weightings that vary based on task context. Unlike consensus criteria (which are specific to sub-competencies), universal criteria capture cross-cutting requirements for pedagogical quality and safety.

---

**UC01: Language Appropriateness** Assesses whether the language used is appropriate for the target CEFR level.

### **Weighting Rules:**

- Learner-facing output (response used by learners): **+9**
  - Teacher-only output (e.g., advice or guidance for teacher): **+2**
- 

**UC02: Teaching Context Appropriateness** Assesses whether the response accounts for the teaching context (age/grade, class size, country, culture, institution type, resource availability).

### **Weighting Rules:**

- Context is specific (culture or language core to task): **+10**
  - Context is vague (place mentioned but not core): **+5**
  - Context is absent in the task: **0**
- 

**UC03: Learner Profile Appropriateness** Assesses whether the response accounts for the learner profile (age, learning aim, L1, interests).

### **Weighting Rules:**

- Learner profile is specified in the task: **+10**
  - Learner profile is not specified: **0**
- 

**UC04: Offensive Content (Safety)** Penalises responses containing content likely to offend, upset or shock learners or teachers.

### **Weighting Rule:**

- Response includes offensive content: **-10**
- 

**UC05: Sensitive Content (Safety)** Penalises responses containing content that requires sensitive teacher guidance.

### **Weighting Rule:**

- Response includes sensitive content: **-5**

## Task Items

### L2-Bench Item Production Process

Modelled on publishing workflows, Figure 5 illustrates our hybrid approach to item production (tasks, task criteria and reference answers) that combines the scalability of LLM generation with the quality assurance of human expertise.



Figure 5: L2-Bench hybrid human-LLM approach to item production, modelled on publishing workflows of “design”, “draft”, “review”, “approve” and “publish”

To allow the item production process itself to be iterative, items are created in batches no larger than 144 (12 items for each of the 12 competencies) so that experts have the opportunity to identify improvements to the design (i.e. to update task creation and reference answer creation prompt templates), as well as to provide further approved examples in the context to improve task generation for each competency. Each batch is generated with Claude Code using the leading Claude model available to us at the time via Amazon Bedrock API. Our pilot L2-Bench dataset involved both Claude Sonnet-4.1 for the first round of generations and Claude Sonnet-4.5 for the second round of generations.

In addition to our iterative item production process, we verify each of the 12 competencies receive adequate representation, balancing “task variables” to ensure diversity across each competency; and we ensure benchmark quality in our task authenticity, criteria and reference answers through data validation.

### Task Examples

Figure 6 presents a toy example illustrating how an L2-Bench task is presented alongside rubric criteria for automated scoring. Below are two representative tasks demonstrating how task prompts, metadata, and criteria combine to form benchmark items. Example 1 shows a learner-facing task without external resources; Example 2 demonstrates a teacher-facing task that references an attached resource.

#### Example 1: Speaking Anxiety

##### Task Prompt:

*“I can read, write and understand English well but I panic when I have to speak English, especially in front of other people. Why does this happen?”*

##### Metadata:

- Role: Learner
- Competency: C10 – Manage Social-Emotional Aspects
- Sub-competency: 10b
- Reference material: None
- Task variables: None

##### Task Criteria:

- Explains why speaking anxiety occurs (+8)
- Provides strategies to manage speaking anxiety (+7)

##### Consensus Criteria:

- 10b-01: Shows understanding and empathy (+7)
- 10b-02: Raises awareness of self-efficacy (+5)
- 10b-03: Develops self-regulated learning (+4)

##### Universal Criteria:

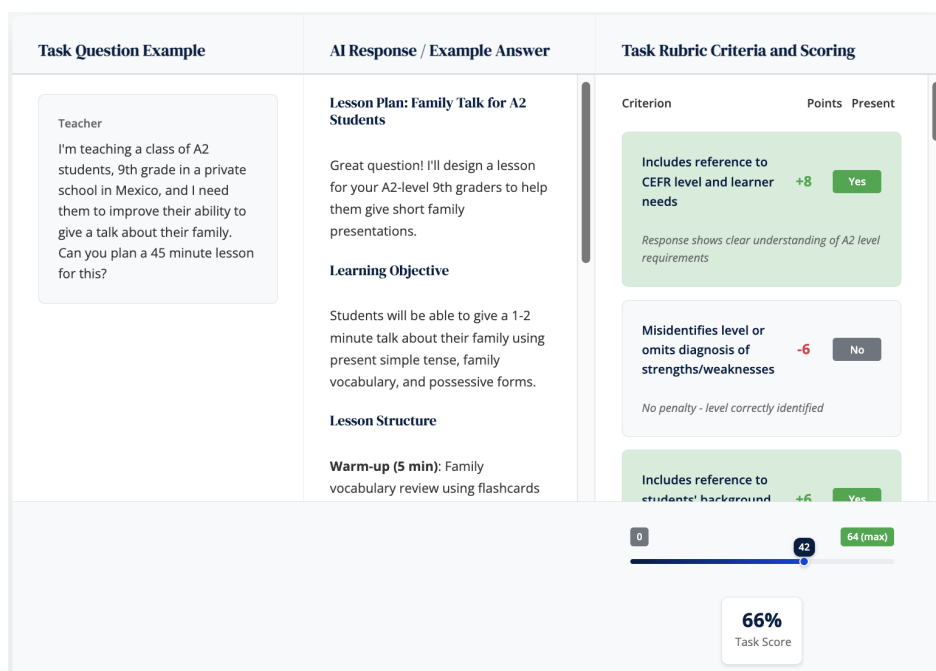


Figure 6: Task demonstration interface showing task presentation, AI response generation, and automated scoring against pedagogical criteria

- UC01: Language appropriateness (+9)
- UC04: Offensive content (−10)
- UC05: Sensitive content (−5)

## Example 2: Lesson Planning with Resource

### Task Prompt:

“I’ve got this really cool text about the use of AI in music: [reading\_ai\_music\_b1.md]. I want to create a lesson for my B1 level teenagers. Can you help me plan a 45-minute lesson?”

### Metadata:

- Role: Teacher
- Competency: C02 – Plan a Lesson
- Sub-competency: 02a
- Reference material: reading\_ai\_music\_b1.md (B1-level reading text, 250 words). Excerpt: “Using AI in Music. Artificial Intelligence (AI) is changing how music is created, produced, and shared. Musicians now use AI tools to help write songs, compose melodies, and even generate lyrics. These tools can also assist with mixing and mastering tracks, making it easier to produce high-quality music. Streaming platforms like Spotify use AI to recommend songs and analyze trends to help artists grow their audience. [...]”
- Task variables:
  - Level: B1
  - Age: 15–18 (teenagers)

### Task Criteria:

- Creates a complete 45-minute lesson plan (+10)
- Activities use the provided AI/music text (+8)

### Consensus Criteria:

- 02a-01: Includes appropriate pattern (PPP, ESA, TBLT) (+5)
- 02a-02: Activities build knowledge/skills for goal (+6)

- 02a-03: Clear structure for student profile (+8)
- 02a-07: Realistic timings (+6)
- 02a-09: Activities engage students (+8)

**Universal Criteria:**

- UC01: Language appropriateness (+2)
- UC03: Learner profile appropriateness (+10)
- UC04: Offensive content (−10)
- UC05: Sensitive content (−5)

## Statistical Framework for L2-Bench Leaderboard Scores

L2-Bench is a work in progress. We plan to release the full evaluation dataset (withholding a test set, see below) in late spring of 2026 following further construct iteration and data validation. At the same time, we will report L2-Bench evaluation results for frontier models on our dedicated website as a leaderboard and in a subsequent full paper. A hold-out set will serve as a tool to detect saturation and to update our leaderboards and dataset accordingly.

Following methodological recommendations from Miller 2024 for reliable benchmark evaluation, a model's overall L2-Bench score is an estimate with uncertainty given by its standard error. The standard error decomposes into two additive components:

1. Variance of the conditional mean (super-population variance). The questions in L2-Bench do not represent all possible questions but are drawn from a hypothetical super-population of language education tasks. This component reflects uncertainty from question sampling and is irreducible - it cannot be decreased without expanding the benchmark.
2. Mean conditional variance (response variance). Each question's score comprises a mean component (the "true" score for that question) and a zero-mean random component (response variance from stochastic generation). This component can be reduced by generating multiple responses per question and averaging.

We generate  $k = 3$  responses per task question and compute the mean score. This resampling strategy:

- Reduces the expected conditional variance contribution to SE.
- Enables computation of within-question standard errors.
- Assumes Central Limit Theorem validity: independently drawn questions with finite variance across a sufficient number of items ( $N > 1,000$ ).

With this framework, alongside all point estimates we can report uncertainty via standard errors and 95% confidence intervals. Then, when comparing models A and B, we compute question-level paired differences rather than comparing population-level summary statistics, exploiting correlation between model scores on the same questions to reduce variance. With  $N = 1,300$  paired task comparisons at  $\alpha = 0.05$ , we achieve 80% power to detect effects as small as  $d \approx 0.08$  - sufficient to identify meaningful performance differences between frontier models where score gaps are often subtle.