OXFORD UNIVERSITY PRESS

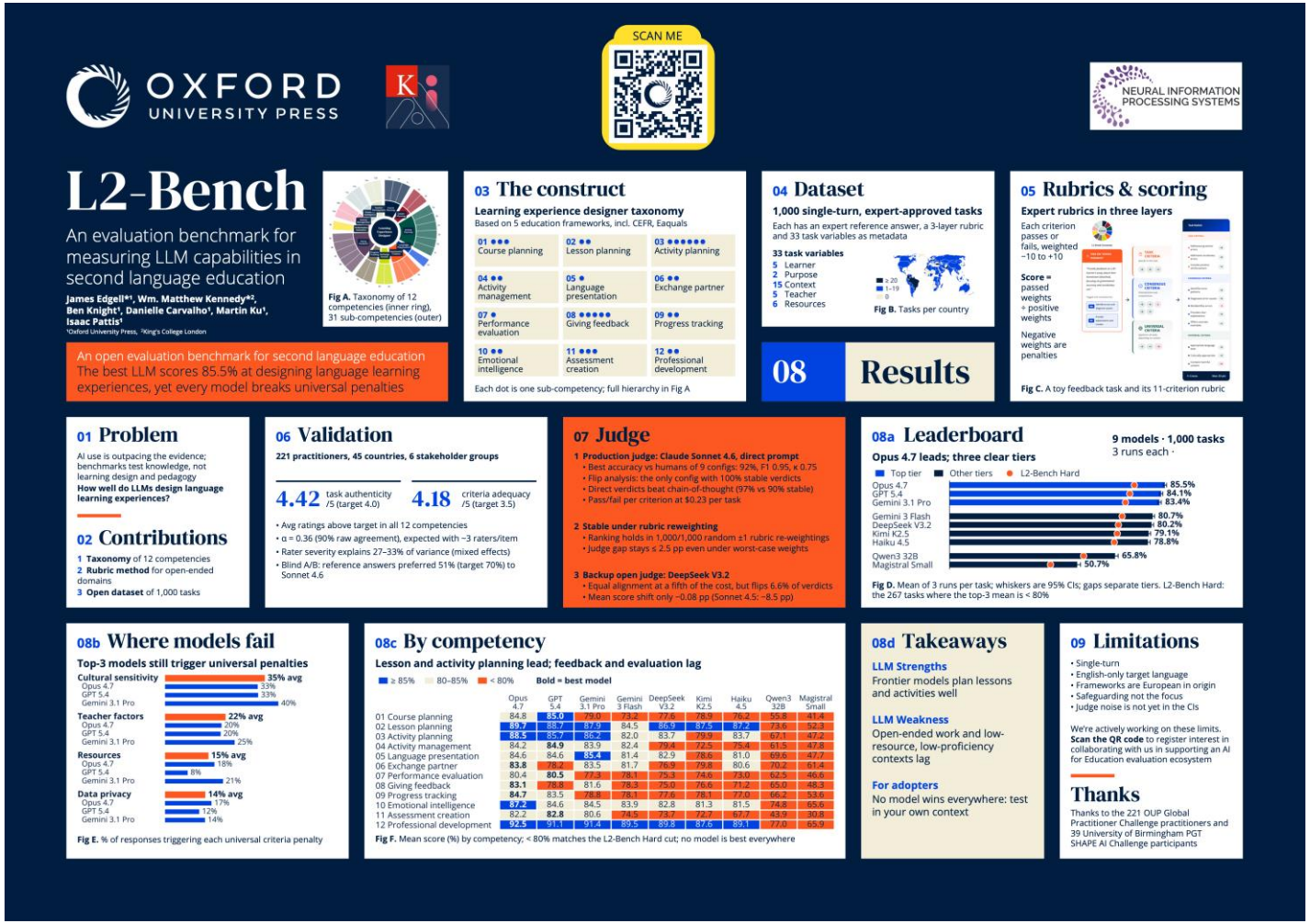
NEURAL INFORMATION PROCESSING SYSTEMS

L2-Bench

An Evaluation Benchmark for Measuring LLM Capabilities in Second Language Education

James Edgell (OUP) presenting

13th November 2026



Who are we?

Inter-disciplinary team of Data Scientists, Pedagogy Experts, AI Engineers, AI Evaluation Science Researchers



James Edgell



Ben Knight



Isaac Pattis



Danielle Carvalho



Martin Ku



Wm Matthew Kennedy

OXFORD UNIVERSITY PRESS

1. Why AI Evaluations in (Language) Education Matter

Benchmarks Drive AI Development and Adoption

Why this matters? Benchmarks drive AIED adoption

Benchmarks are essential for practitioners, organisations and developers, especially in education

For Researchers

Providing signal about where model performance should be improved

For Institutions

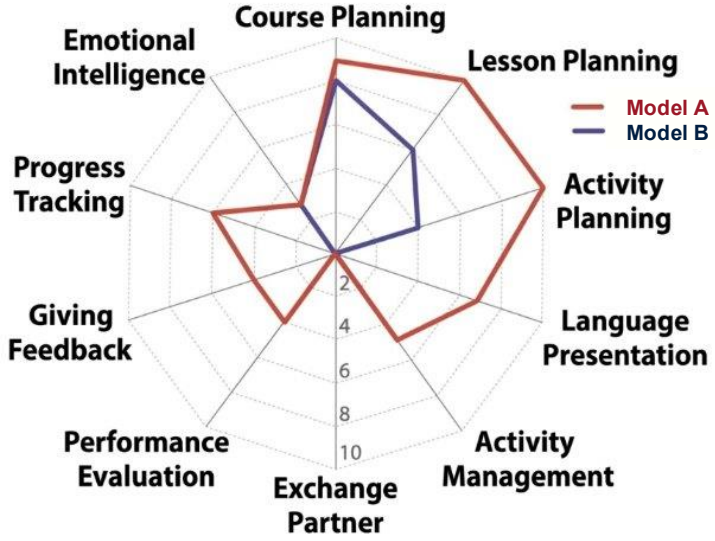
Informed decision-making when selecting AI tools for schools

For Educators

Practical AI capability assessment for specific teaching scenarios

Which model is best in language education?

Sketch of ELT capabilities of Model A vs Model B



No standardized AI benchmarks exist in Language education

Language learning has unique pedagogical challenges relative to other educational domains

Means / Goal

Language is both the means **and** the goal to learning.

Skills Heavy

Balance between **knowledge** and skills: language education is much more weighted towards **skills** and real-time fluency.

Interpersonal

Language is a **social** interpersonal skill where the 'right' answer can vary for every exchange we have.

Intercultural

Language is an intercultural activity which requires us to understand and communicate with people who may have **very different assumptions**.

Language learning **#4 topic** in **Microsoft Copilot** usage 2025

Costa-Gomes et al. (2025)

\+ among **top use-cases** for Claude

Tamkin et al. (2025)

OXFORD UNIVERSITY PRESS

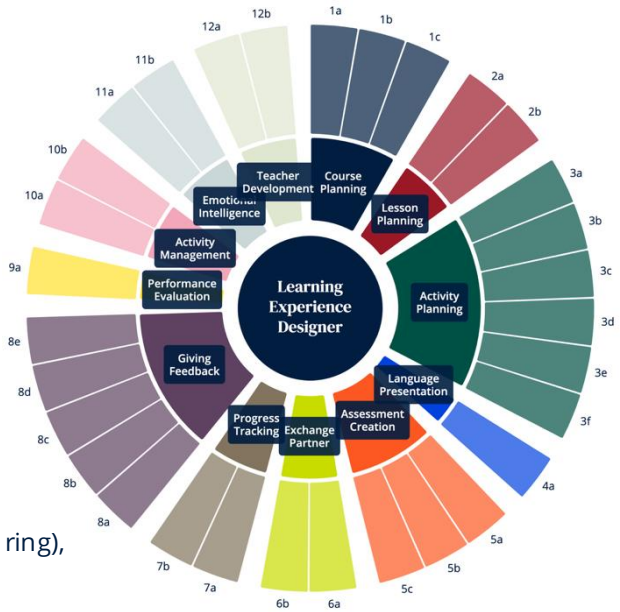
2. L2-Bench Construct and Results

L2-Bench is an Open AI Benchmark for Second Language (L2) Education

Contribution 1: L2-Bench Competency Taxonomy

Evaluate AI's capabilities as a "learning experience designer"

- Covers roles that intentionally design the conditions that shape how people learn in second language education:
  - Teaching, materials development, learning design, assessment development and use, teacher development
- Combines teaching competency frameworks (BC 2025, EAQUALS 2015), learning science frameworks (OxPoLL 2026)



L2-Bench competency construct for a learning experience designer in second language education.

Taxonomy of 12 competencies (inner ring), 31 sub-competencies (outer)

01 ●●● Course planning	02 ●● Lesson planning	03 ●●●●●● Activity planning
04 ●● Activity management	05 ● Language presentation	06 ●● Exchange partner
07 ● Performance evaluation	08 ●●●●● Giving feedback	09 ●● Progress tracking
10 ●● Emotional intelligence	11 ●●● Assessment creation	12 ●● Professional development

Each dot is one sub-competency.

Contribution 2: 1,000 Open Tasks with Validation

Lesson planning
I'm teaching A1 level children (ages 6-7) in Bangkok, Thailand. Tomorrow's lesson is about family members and I want to create hands-on materials that young children can manipulate and use. I need resources for a 50-minute lesson that includes visuals, activities, and possibly songs or games. Can you help me put together a plan for the lesson, and where I can get any resources I need?

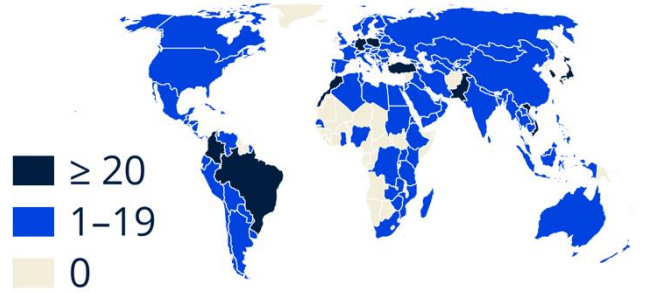
Giving feedback
I'm a B1 student from Cambodia and I just finished a speaking test where I described my favorite place in my country. The examiner recorded it. I want to improve before my final exam next month. Can you listen to my recording and tell me what I need to work on? Here's what I said: [speaking_favorite_place_assessment.md]

L2-Bench consists of 1,000 rubric-scored task-response pairs

Real-world task questions (single turn) are created against each competency and **validated by 221 practitioners**:

- L2 English as target language (EFL focus, not ESL)
- 1,000+ tasks **publicly released** (incl. ref answers, rubric criteria, **task variables** e.g. L1 geography)

Tasks per L1 country



Contribution 3: Systematic 3 Layer Criteria System for Task Rubric

Begin with a sub-competency's 'Consensus Criteria'

STEP 1: COMPETENCY TAXONOMY



Learning Experience Designer Competencies

Our taxonomy defines what effe
to do in language education, orga
competencies and sub-competer

12 Competencies

30 Sub-competencies

STEP 3: TASK RUBRIC ASSEMBLY

STEP 2: SUB-COMPETENCIES & CONSENSUS

Each s



Giving Feedback

08a

08b

08c

08d

TASK ON "GIVING FEEDBACK"

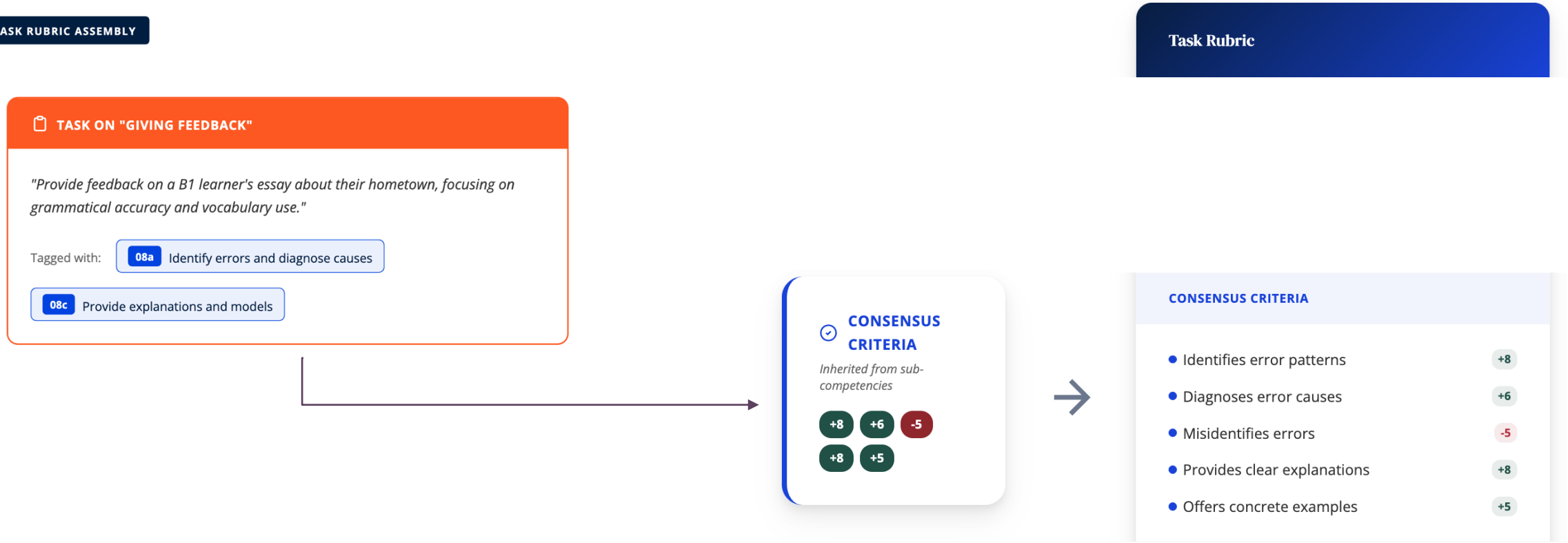
"Provide feedback on a B1 learner's essay about their hometown, focusing on grammatical accuracy and vocabulary use."

Tagged with: 08a Identify errors and diagnose causes

08c Provide explanations and models

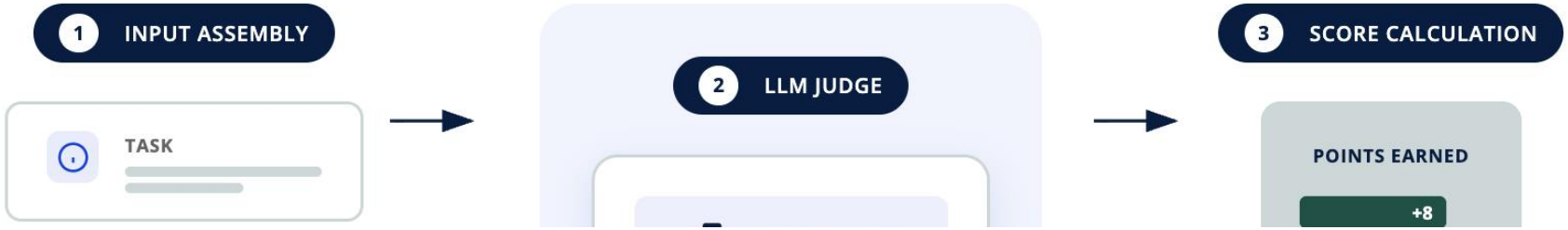
Task Rubrics also include ‘Task Criteria’ and ‘Universal Criteria’

STEP 3: TASK RUBRIC ASSEMBLY

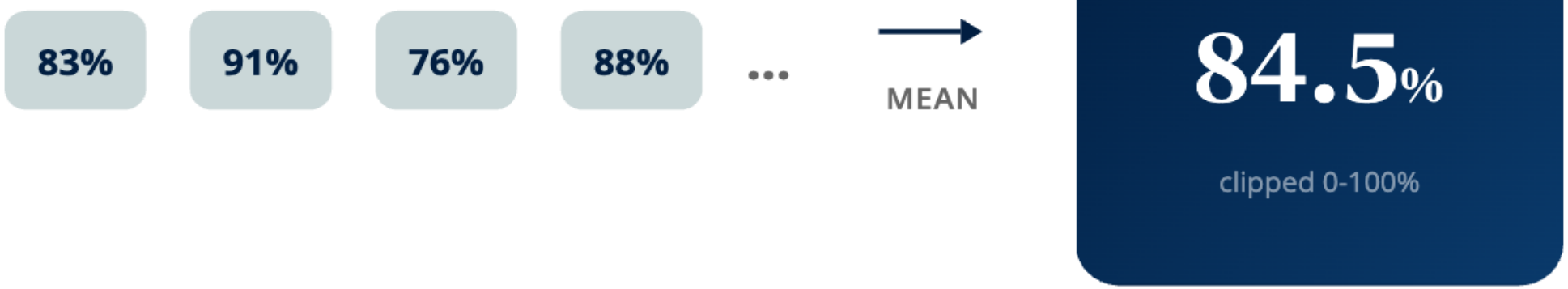


OXFORD UNIVERSITY PRESS

Score 9 models (GPT, Claude, Gemini etc) on all 1,000 L2-Bench tasks

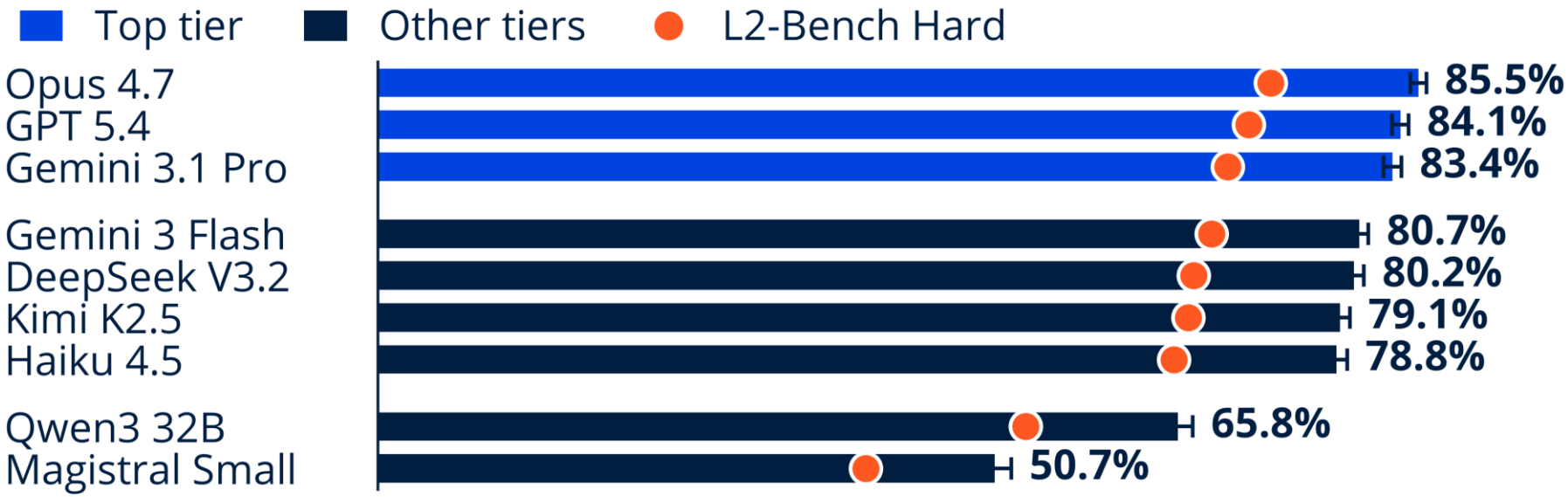


Aggregate Across All Tasks



The frontier models cluster in the low-to-mid 80s

"Hard" tasks indicate there is headroom for model improvements



Mean of 3 runs per task; whiskers are 95% CIs; gaps separate tiers. L2-Bench Hard: the 267 tasks where the top-3 mean is < 80%

Top tier: Opus 4.7 85.5%, GPT 5.4 84.1%, Gemini 3.1 Pro 83.4%

Mid tier: Gemini 3 Flash, DeepSeek V3.2, Kimi K2.5, Haiku 4.5: 79 to 81%

L2-Bench "Hard": the 267 tasks where all 3 top tier models scored under 80% (they fall to 73.4%, 71.6%, 69.9%)

OXFORD UNIVERSITY PRESS

Per competency L2-Bench scores reveal strengths and weaknesses

Structured tasks are strongest, and the top model is not the winner everywhere

Strong competencies: Teacher development (86%), Lesson planning (82%)

Weak competencies: Creating assessments (68%), Giving feedback (73%)

Model competency strengths: Gemini 3 Flash beats GPT 5.4 at conversational exchange

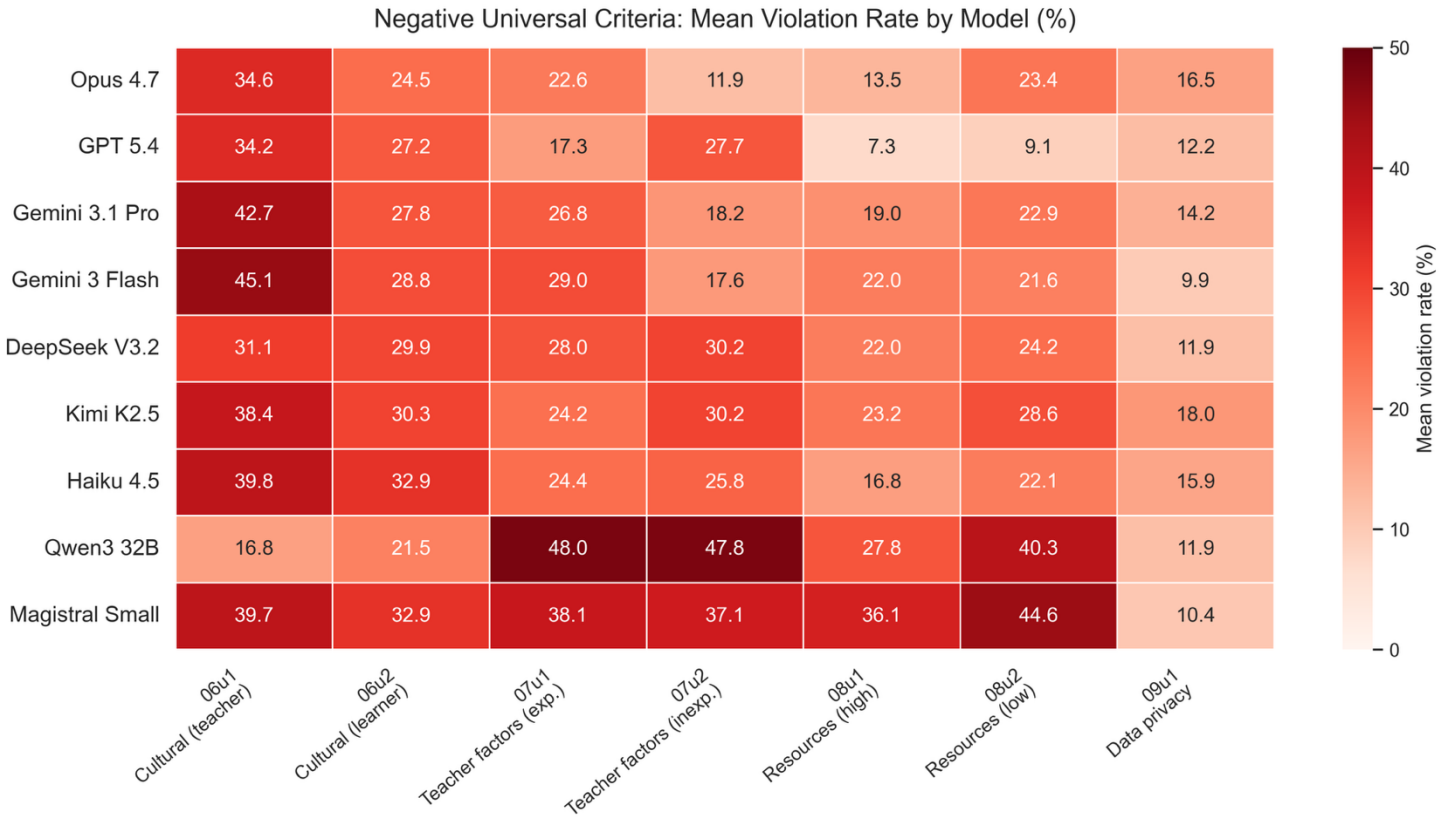
≥ 85% 80–85% < 80% **Bold = best model**

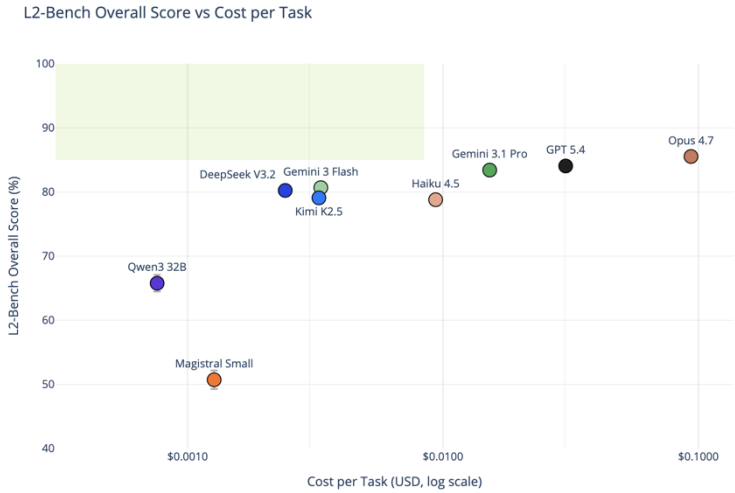
	Opus 4.7	GPT 5.4	Gemini 3.1 Pro	Gemini 3 Flash	DeepSeek V3.2	Kimi K2.5	Haiku 4.5	Qwen3 32B	Magistral Small
01 Course planning	84.8	**85.0**	79.0	73.2	77.6	78.9	76.2	55.8	41.4
02 Lesson planning	**89.7**	88.7	87.9	84.5	86.9	87.5	87.2	73.6	52.3
03 Activity planning	**88.5**	85.7	86.2	82.0	83.7	79.9	83.7	67.1	47.2
04 Activity management	84.2	**84.9**	83.9	82.4	79.4	72.5	75.4	61.5	47.8
05 Language presentation	84.6	84.6	**85.4**	81.4	82.9	78.6	81.0	69.6	47.7
06 Exchange partner	**83.8**	78.2	83.5	81.7	76.9	79.8	80.6	70.2	61.4
07 Performance evaluation	80.4	**80.5**	77.3	78.1	75.3	74.6	73.0	62.5	46.6
08 Giving feedback	**83.1**	78.8	81.6	78.3	75.0	76.6	71.2	65.0	48.3
09 Progress tracking	**84.7**	83.5	78.8	78.1	77.6	78.1	77.0	66.2	53.6
10 Emotional intelligence	**87.2**	84.6	84.5	83.9	82.8	81.3	81.5	74.8	65.6
11 Assessment creation	82.2	**82.8**	80.6	74.5	73.7	72.7	67.7	43.9	30.8
12 Professional development	**92.5**	91.1	91.4	89.5	89.8	87.6	89.1	77.0	65.9

Mean score per competency; models ordered by overall rank

Where the risk concentrates

Cheaper models degrade fastest in exactly the contexts we care most about





Every model triggers universal penalties:

- Cultural sensitivity, teacher factors resource assumptions, data privacy

The top tier varies little across **resource levels.** However, **small models** do not:

- Qwen3 32B: 28% for low vs 40% for high

i.e. *does the cheaper model only work for* ***already-advantaged learners?***

OXFORD UNIVERSITY PRESS

3. L2-Bench Validation

OUP Global Practitioner Validation for L2-Bench

OXFORD UNIVERSITY PRESS

Homes 2019

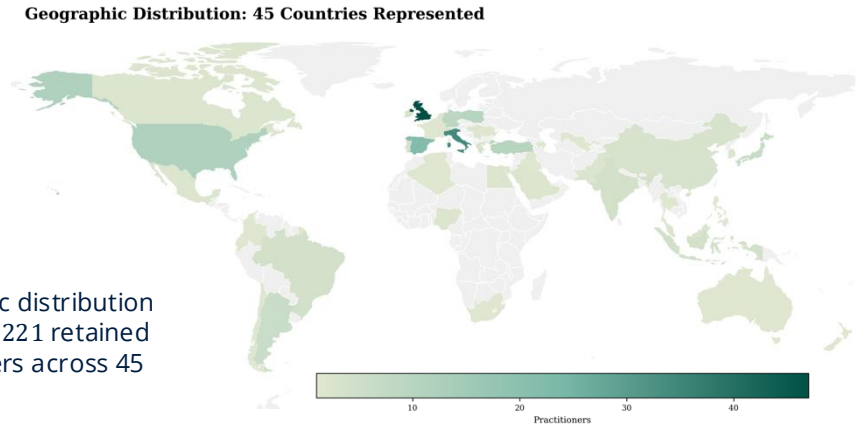
No One Role Can Validate the Construct

221 Practitioners from 6 stakeholder groups over 45 countries

Stakeholders **Practitioner Groups**

Six stakeholder groups validate L2-Bench through their unique professional perspectives

- Language Learning Practitioners
- Language Teachers
- Assessment Specialists
- Researchers
- EdTech Professionals
- Learners



Geographic distribution of the $N = 221$ retained practitioners across 45 countries

Join the L2-Bench OUP Global Practitioner Challenge

Help shape the future of AI evaluations in second language (L2) education

Call for Participants

- Are you an education practitioner that **shapes how people learn**?
- Interested in evaluating how well AI designs learning experiences?
- Join our research study and contribute your expertise!

What You'll Do

Complete 9 to 27 short online tasks in 1 week.

Rate AI evaluations for authenticity and adequacy.

Compare AI vs. "gold-standard" responses.

Why Join?

Influence AI *for* Education.

Contribute to impactful, peer-reviewed research.

Combine our academic heritage with AI evaluations across the org.

SCAN ME

L2-Bench Model Leaderboard

2 1 3

Key Dates

Week of **March 2-8** AoE.

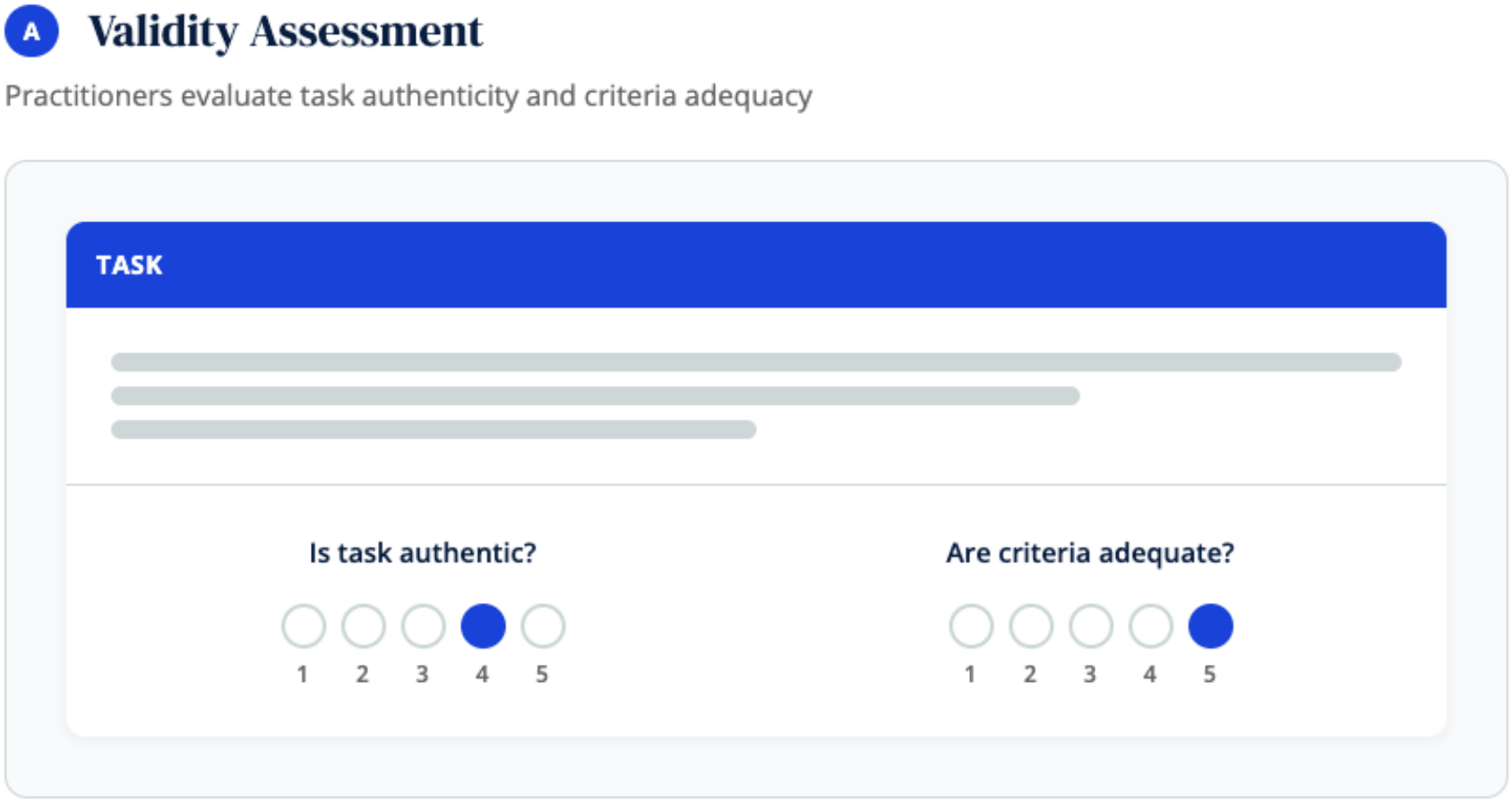
Register by **Tue 24 Feb.**

Ready to take on the challenge?

Scan the QR code to complete the expression of interest form

RO1: Dataset validity measured by 2x 5-point Likert questions

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



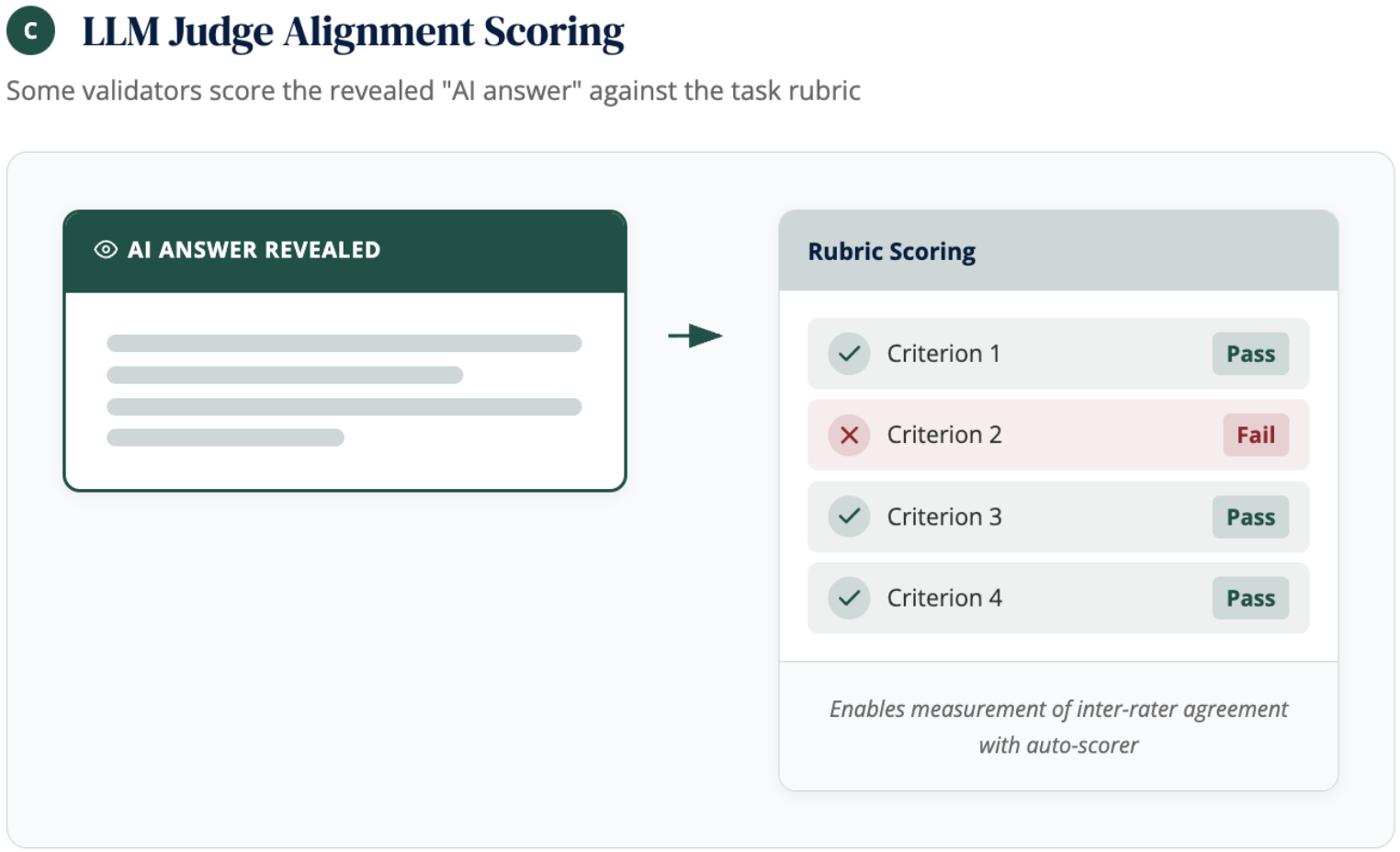
RO2: Measuring answer quality through preference voting

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



RO3: Measuring LLM-as-a-judge alignment with human scoring

RO3. Measuring LLM-as-a-Judge validity by collecting practitioner scores on “AI answers” against the task rubric.



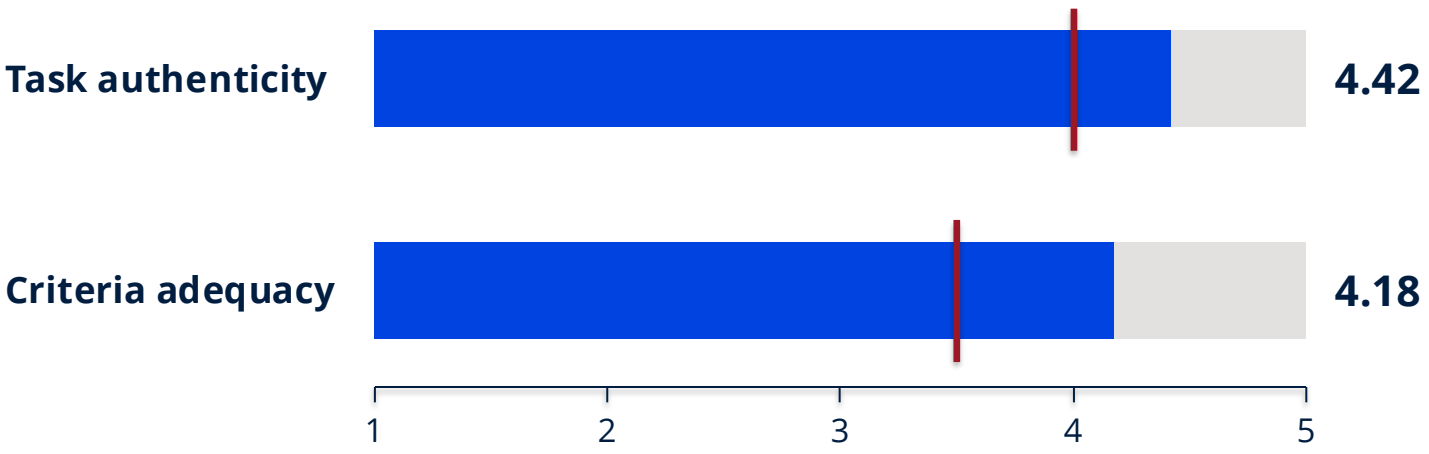
4. Validation Results

Construct and Judge Validation Results

Making claims about construct validity

Construct validity = the testing target (Second Language Learning Experience Design) is :

- stable enough to assess using an automated benchmark
- faithfully represents the target according to the consensus of the community



Mean rating on a 1–5 scale, 95% CI ±0.04. **Red tick = target** (4.0 authenticity, 3.5 adequacy).

How the construct is built

5 practitioner frameworks
CEFR, Eaquals, Cambridge CETF, NILE, British Council CPD

12 competencies, 31 sub-competencies

Caveats

Krippendorff's α = 0.36

27–33% of rating variance is **rater severity**.

Making claims about judge validity

Judge validity = the LLM employed to score model responses against our rubric does so:

- In a **stable** manner (it scores different kinds of responses consistently)
- In a manner that **faithfully agrees** with expert consensus (after accounting for human rating variability)
- Also in a manner that is less certain in **precisely the same areas that** human experts are uncertain

Judge performance and agreement

The production judge (Sonnet 4.6 as a classifier [v1]) achieved Cohen's κ = 0.746 against human rater majority vote (67%) with accuracy of 92.3%, F1 = 0.953 across 326 criteria with matched practitioner validation data.

These results must be interpreted against the human inter-rater reliability ceiling: on criterion-level binary judgments, practitioners achieved Krippendorff's α = 0.362 [95% CI: 0.343, 0.380] with 89.8% raw agreement.

Alternative judge

While the production judge gives the highest + most stable score amongst 15 configurations tested, a cross-family DeepSeek V3.2 judge given the identical prompt (v1) agrees with it to within 0.08 points.

Agreement is also robust to how criteria are weighted: across 1,000 random re-weightings the gap between the two judges remains within 1.06 points.

Making claims about overall measurement validity

Construct validity + judge validity (is strong evidence for) measurement validity

- Recall the great lengths taken to collect various kinds of ratings from a large body of stakeholders – necessary, not sufficient
- L2-Bench is one of the most widely validated evaluation benchmarks in the field
- Still many improvements to make
    - e.g. Multi-turn
    - e.g. Multi-lingual
    - e.g. Safeguarding

AI evaluation is not magic – it is a complex social science measurement challenge

(Wallach et al. 2025)

Where to find more / Questions?

— **benchmarks.elt.edu.oup.com** — leaderboard, blog and dataset links

— **Dataset**: huggingface.co/datasets/OUP/l2-bench

— **Papers**: **results** (arXiv 2607.08842) · **methods** (arXiv 2603.20088) · **judge** (under review)

