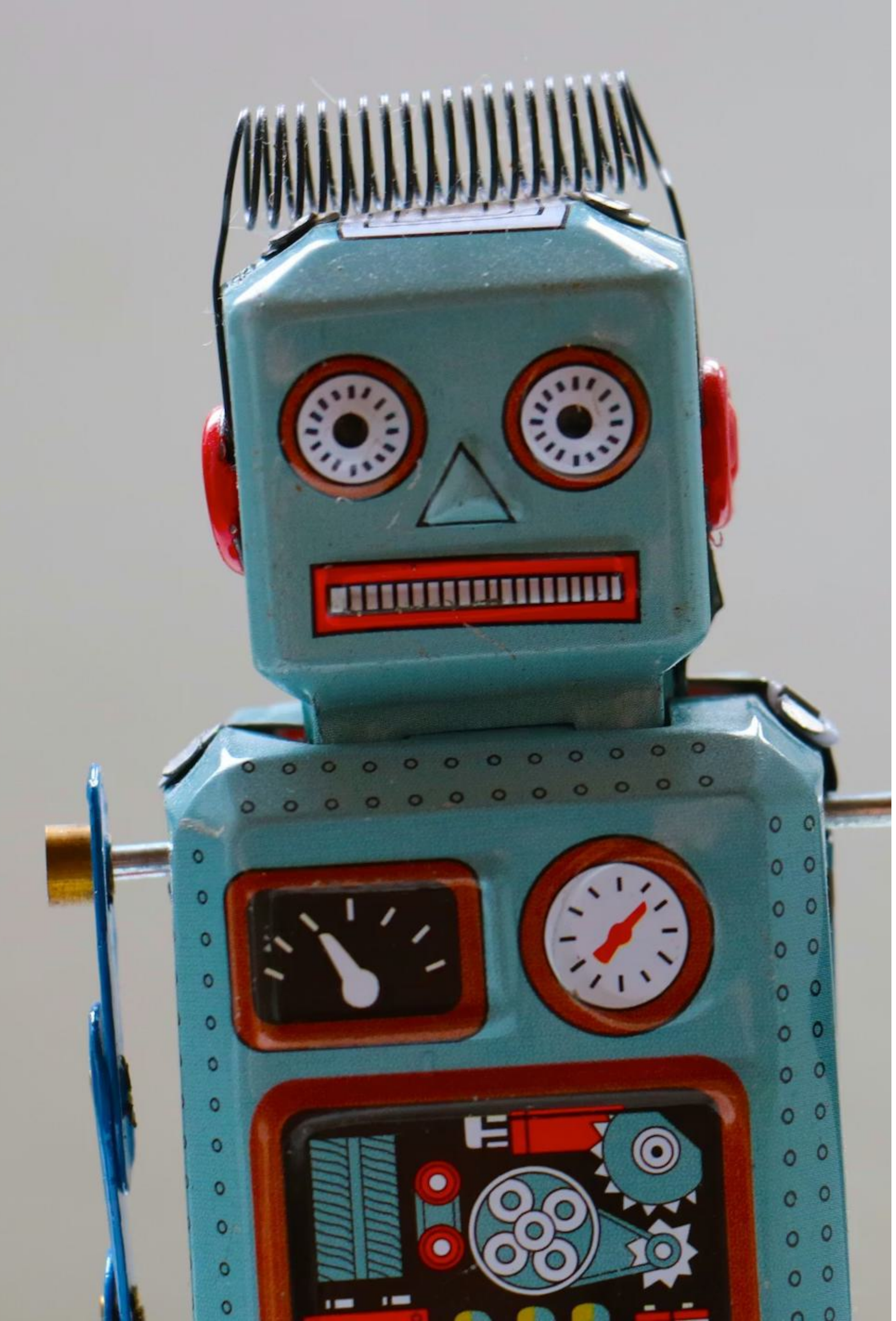
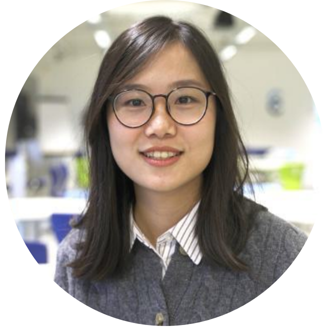




Conversational AI for spoken L2 development: meta-analysis and insights for assessment

Serge Bibauw
University of Louvain

Zhaori Wang
KU Leuven & UCLouvain

Automated L2 Speaking Assessment (AL2SA)
Grenoble, 6 March 2026

A conversation with Claude earlier this week...

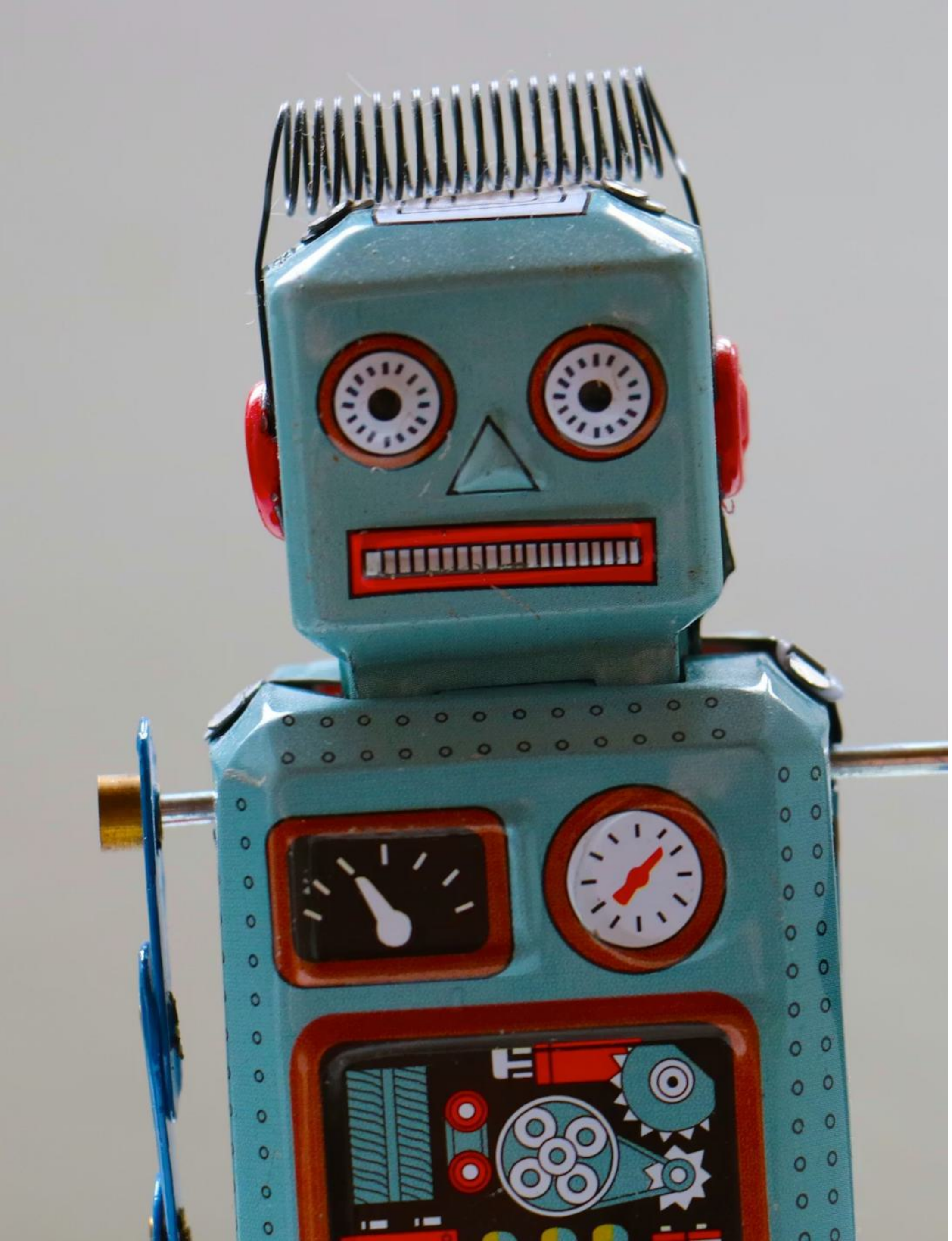
Bonjour Claude. Je veux améliorer mon anglais. Je vais rencontrer des collègues qui font de la recherche. Je veux pratiquer mon anglais avec toi.

Bonjour ! Of course, I'd be happy to help you practice your English. [...]
Let's start — tell me in English about the research topic you'll be discussing with your colleagues. Don't worry about mistakes; I'll help you along the way.

I think didactic of pronunciation

Great start! A small correction: in English, we'd say "the didactics of pronunciation" or more commonly "pronunciation teaching" or "pronunciation pedagogy." The word "didactics" exists in English but sounds quite formal — most anglophone researchers would say "teaching" or "pedagogy."
Now, can you tell me more? For example, what specific aspect of pronunciation teaching are you working on?

Chat with Claude Opus 4.6 • Mar. 2026



Conversational AI for L2 development

ConvAI for LL: recent changes

Access, studies, typology

An updated meta-analysis

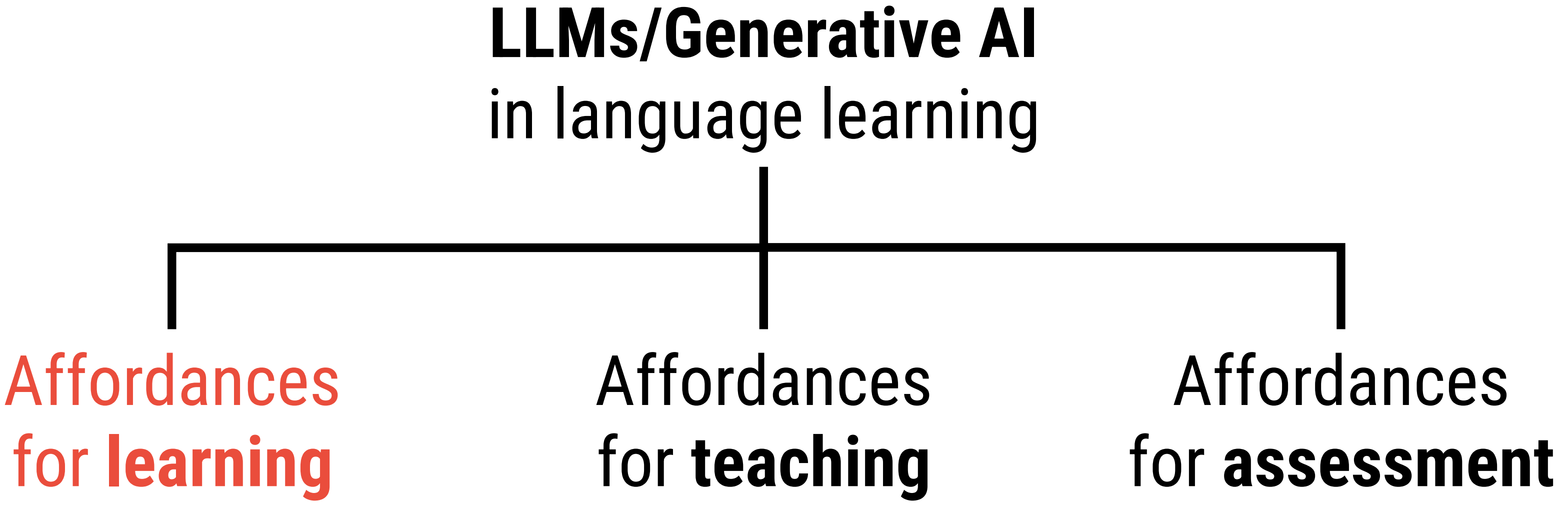
Gaps, methods, data

Effectiveness & moderators

Studies, systems, outcomes

Conclusions & insights

Affordances of generative AI/LLMs for language learning



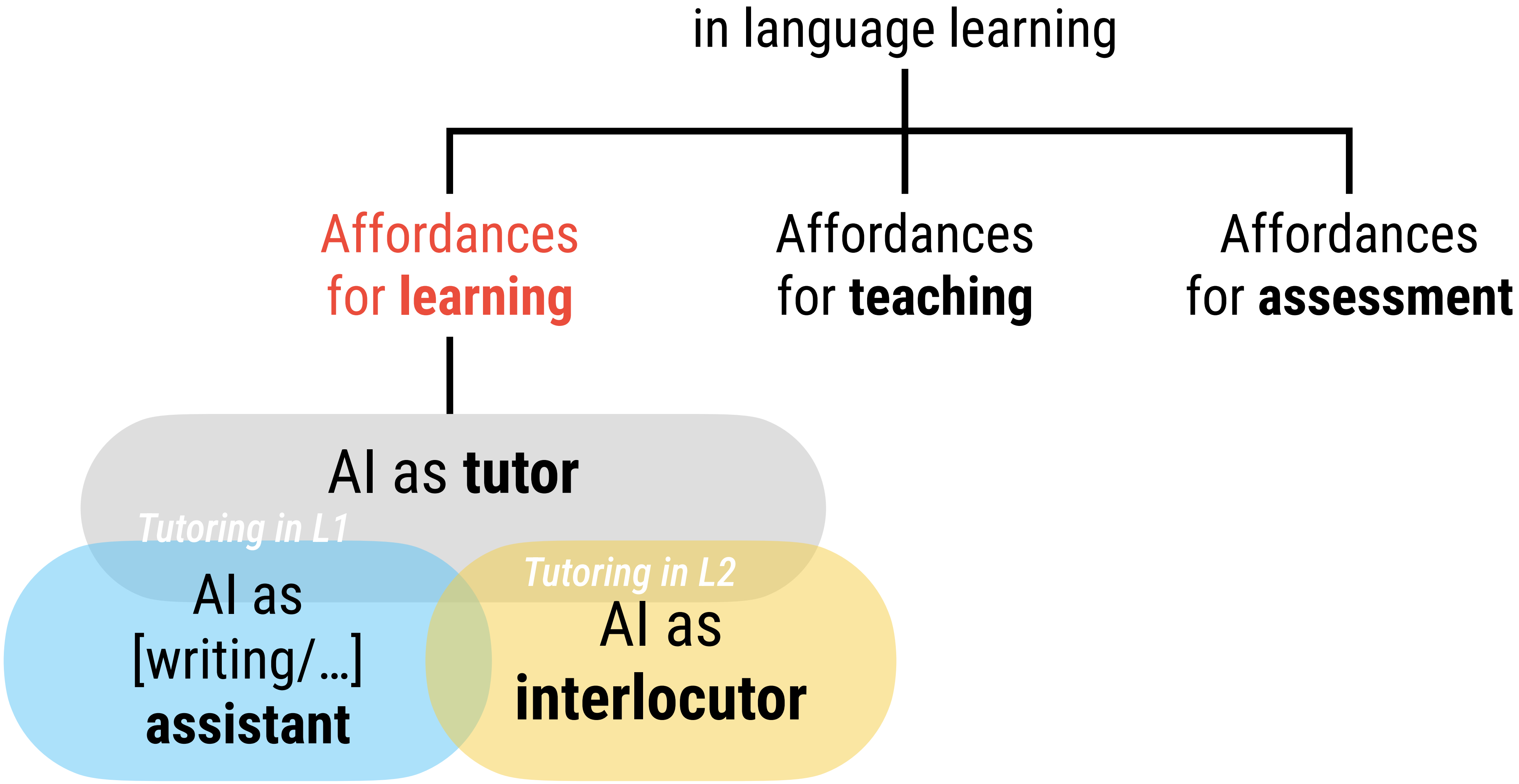
LLMs/Generative AI
in language learning

Affordances
for **learning**

Affordances
for **teaching**

Affordances
for **assessment**

Affordances of generative AI/LLMs for language learning



in language learning

Affordances for **learning**

Affordances for **teaching**

Affordances for **assessment**

AI as **tutor**

Tutoring in L1

AI as [writing/…] **assistant**

Tutoring in L2

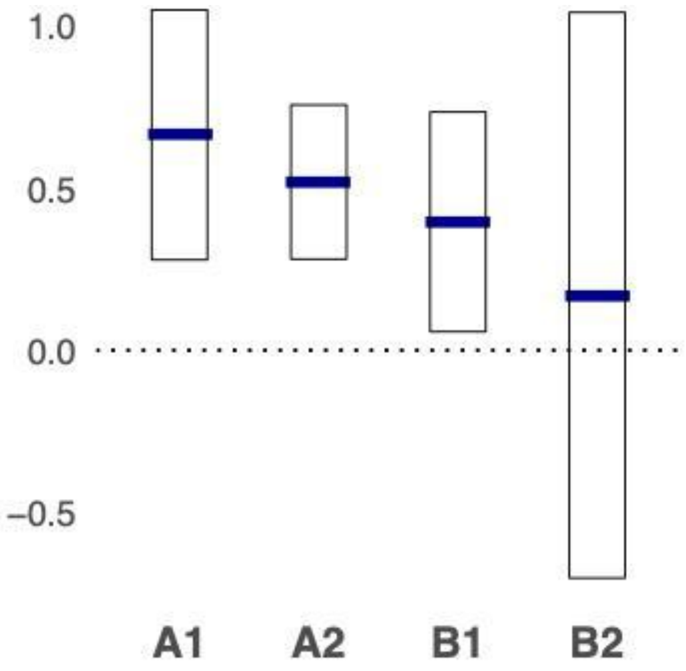
AI as **interlocutor**

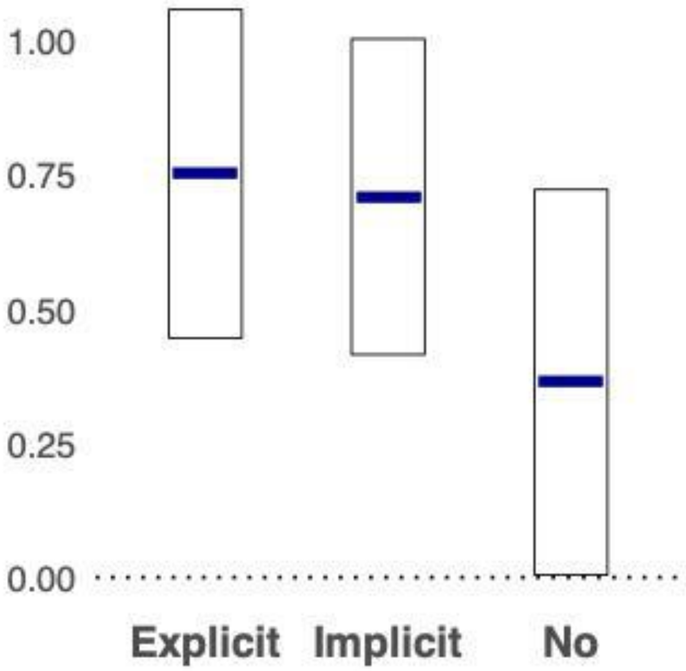
Dialogue-based CALL / Conversational AI for language learning

- Dialogue-based CALL as
    - meaningful, dialogic interactions
    - in the target language
    - with an automated agent
    - for language learning purposes
- based on a ~cognitive-interactionist rationale
  (meaningful communicative practice to develop L2 productive skills)

Previous meta-analyses (Bibauw et al, 2022; Wang et al, 2024; Hou & Min, 2025)

- Medium effect size
  g = .59 (Bibauw et al 2022); *g* = .48 (Wang et al, 2024)
  g = .67 on speaking (Hou & Min, 2025)
- **Differentiated effect** depending on
  - **proficiency** level: A1-A2 benefit + (Bibauw...2022; Hou & Min,2025)
  - **context**: + in out-of-school practice (Hou & Min, 2025)
  - **corrective feedback**: more effect (Bibauw et al 2022)





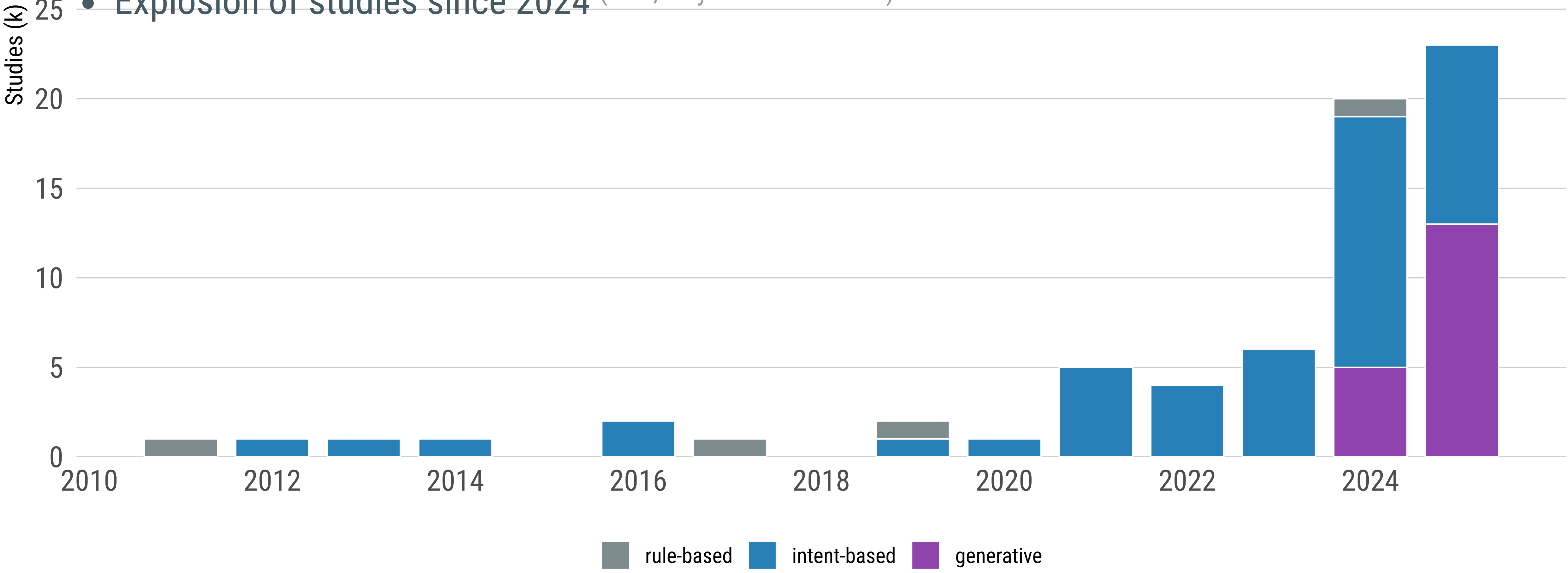
What changed? **Access to chatbots**

- Immediate, free access to **off-the-shelf** (generic) **chatbots**
  - *ChatGPT, Gemini, Claude, Mistral, DeepSeek*...
  - Before, most systems were commercial products for institutions, or in very limited access
- More and more applications dedicated to L2 learning based on LLMs
  - *Duolingo Max, TalkPal, Mondly, Practica, CallAnnie, Learner*...
- Much simpler and faster to create conversational interactions (hours vs years).

What changed? **Studies**

k = 67 studies (2011–2025)

- Explosion of studies since 2024 (here, only included studies)



Toward a new typology of Conversational AI/DBCALL

- Learner response constraints
  (Norris & Ortega, 2000; Spada & Tomita, 2010)

Learner output

Free (turn-level)

Constrained (word-level)

Repeated (say out loud)

Selected (click)

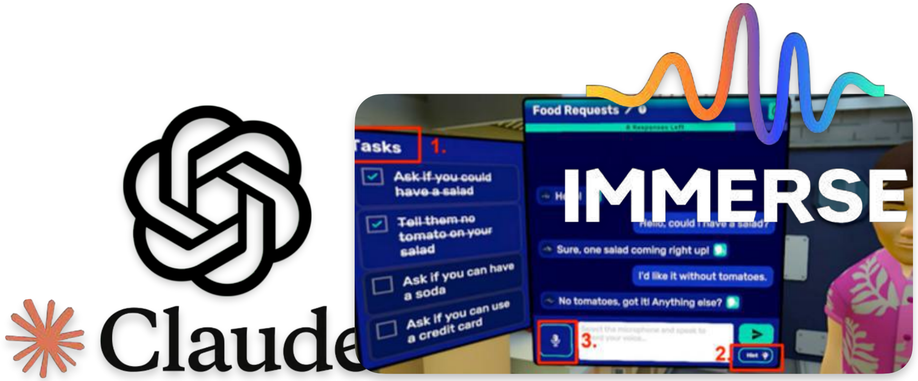
Learner output

Free (turn-level)

Constrained (word-level)

Repeated (say out loud)

Selected (click)



Tasks

☑ ~~Ask if you could have a salad~~

☑ ~~Tell them no tomato on your salad~~

☐ Ask if you can have a soda

☐ Ask if you can use a credit card

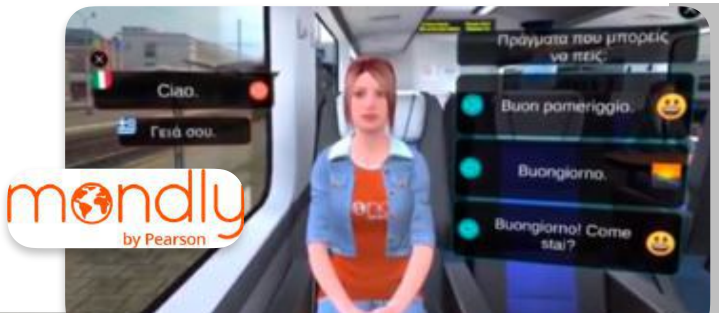
Food Requests

Hello, could I have a salad?

Sure, one salad coming right up!

I'd like it without tomatoes.

No tomatoes, got it! Anything else?



Ciao.

Γειά σου.

Πράγματα που μπορείς να πεις:

Buon pomeriggio.

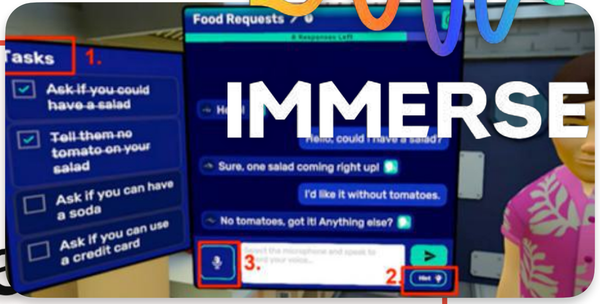
Buongiorno.

Buongiorno! Come stai?

Learner output

Conversational AI / Dialogue-based CALL

Free (turn-level)



Food Requests

~~Ask if you could have a salad~~

~~Tell them no tomato on your salad~~

Ask if you can have a soda

Ask if you can use a credit card

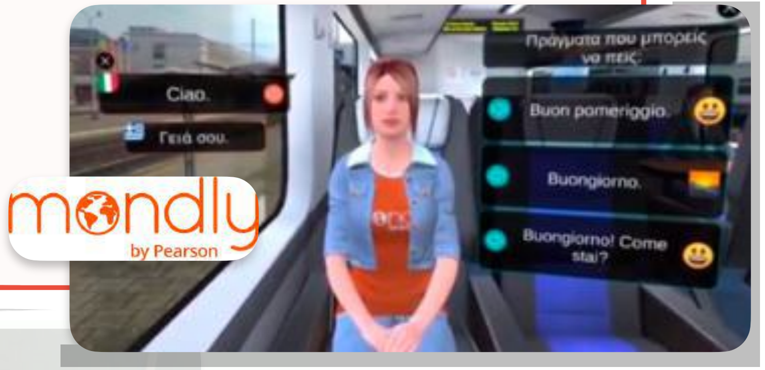
Sure, one salad coming right up!

I'd like it without tomatoes.

No tomatoes, got it! Anything else?

Constrained (word-level)

Repeated (say out loud)



Ciao.

Γειά σου.

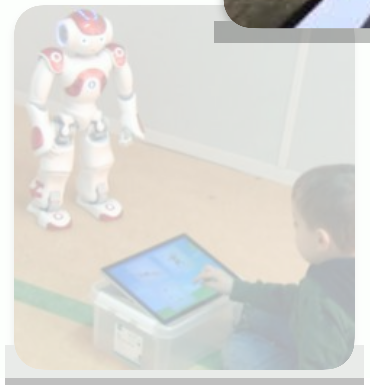
Πράγματα που μπορείς να πεις

Buon pomeriggio.

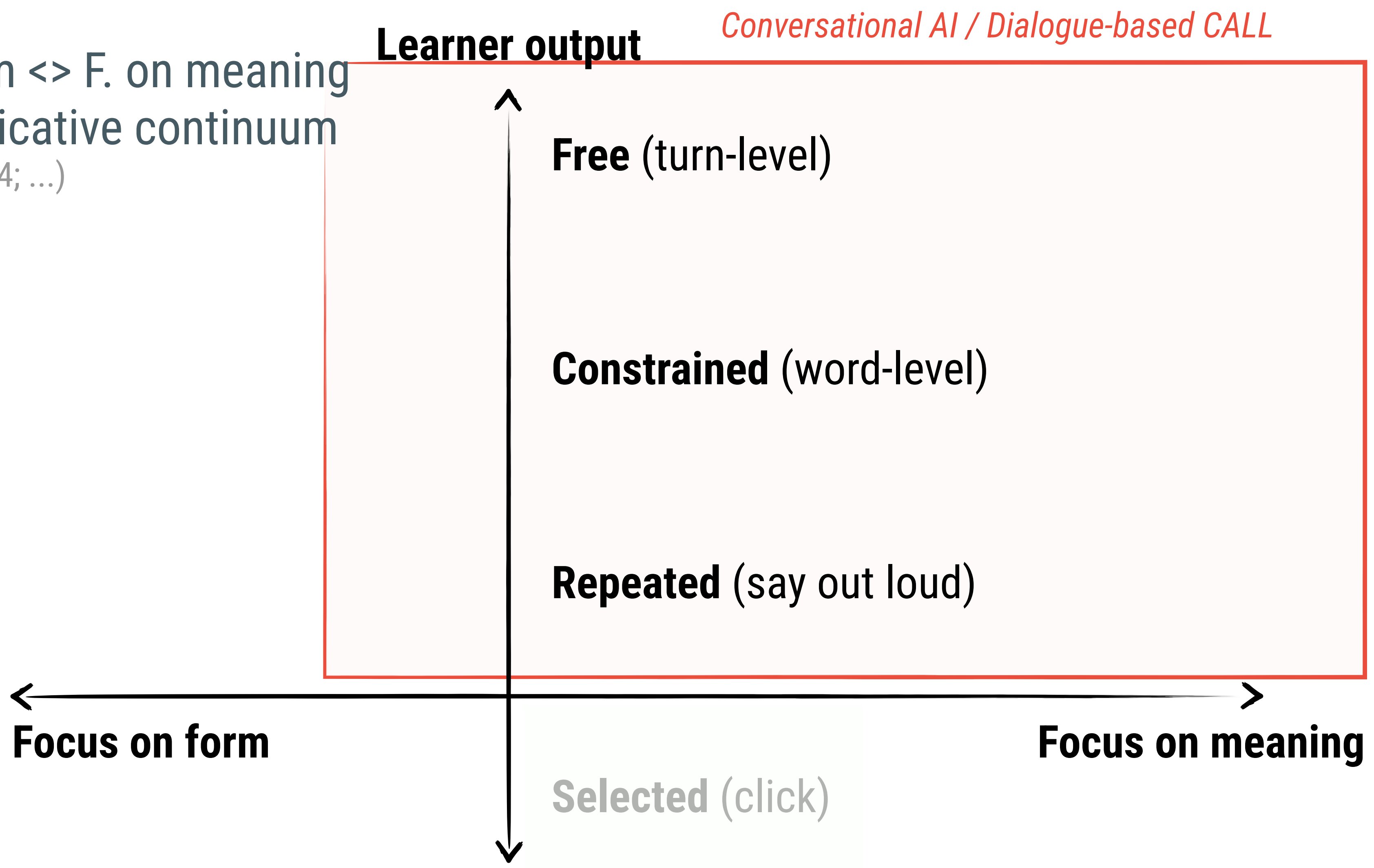
Buongiorno.

Buongiorno! Come stai?

Selected (click)



- 2nd axis:
  F. on form <> F. on meaning
  Communicative continuum
  (Howatt, 1984; ...)



Conversational AI / Dialogue-based CALL

Learner output

Free (turn-level)

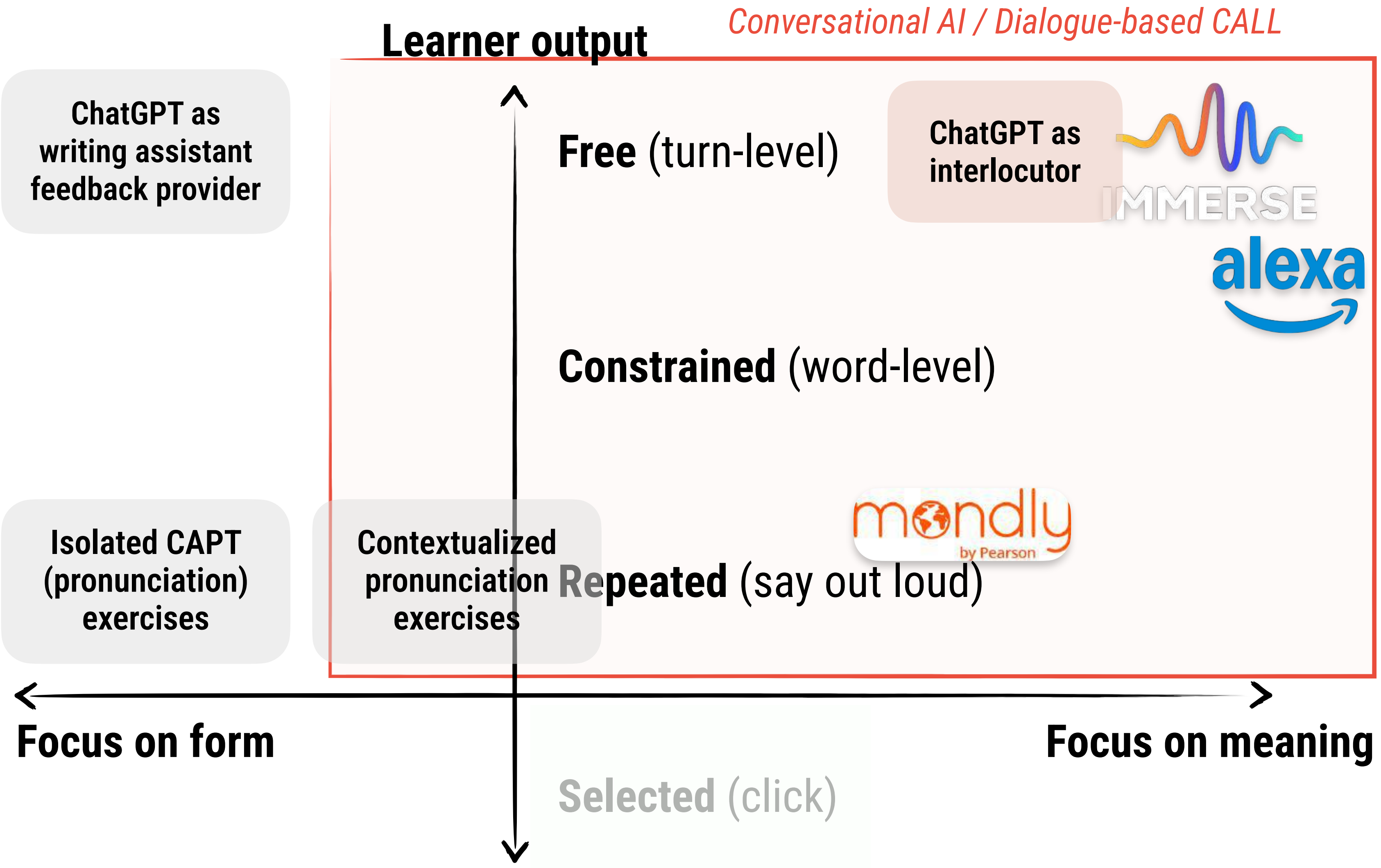
Constrained (word-level)

Repeated (say out loud)

Selected (click)

Focus on form

Focus on meaning



Conversational AI / Dialogue-based CALL

Learner output

ChatGPT as writing assistant feedback provider

Free (turn-level)

ChatGPT as interlocutor

Constrained (word-level)

Isolated CAPT (pronunciation) exercises

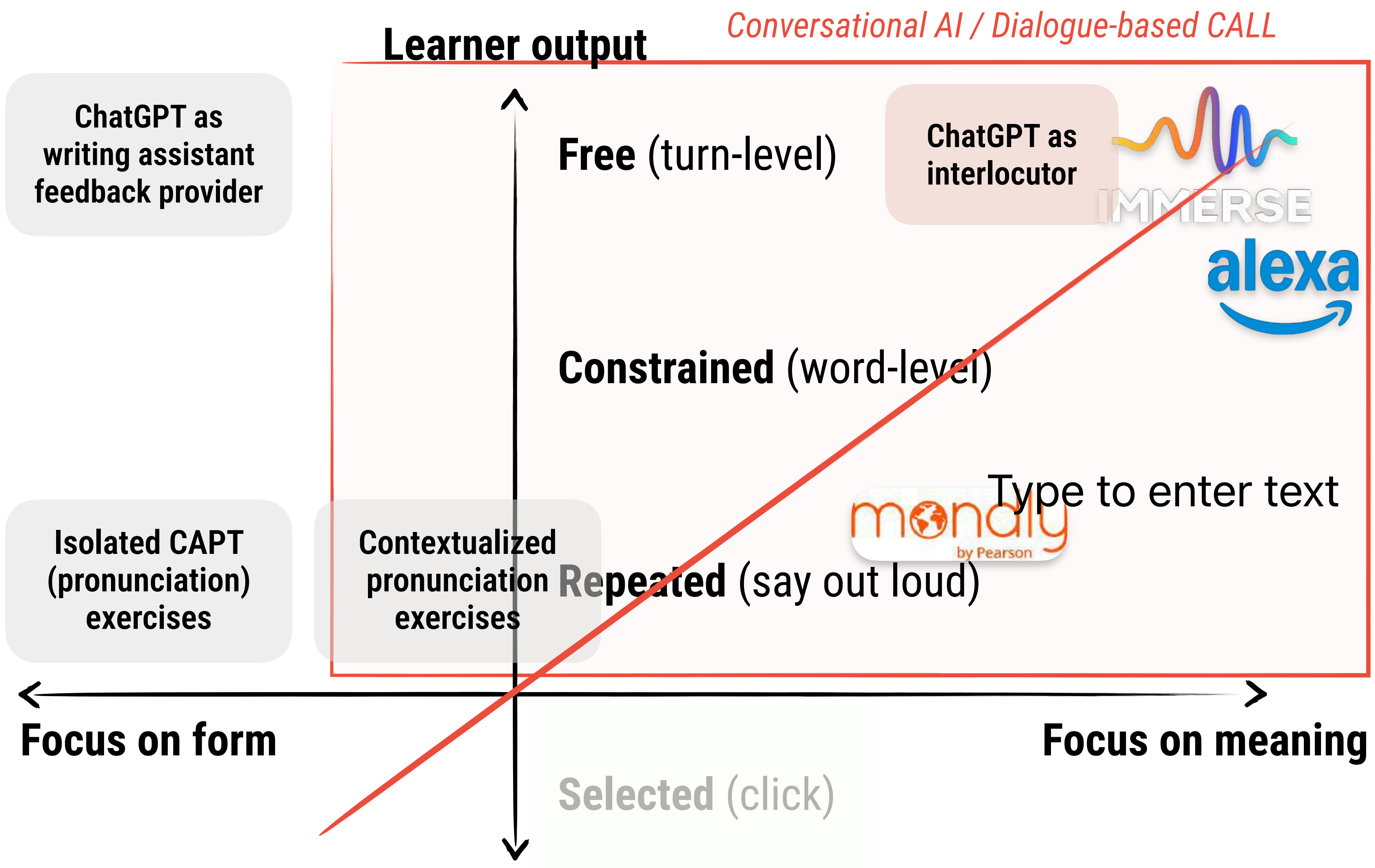
Contextualized pronunciation exercises

Repeated (say out loud)

Focus on form

Focus on meaning

Selected (click)



Conversational AI / Dialogue-based CALL

Learner output

ChatGPT as writing assistant feedback provider

Free (turn-level)

ChatGPT as interlocutor

IMMERSE

alexa

Constrained (word-level)

Type to enter text

mondly by Pearson

Isolated CAPT (pronunciation) exercises

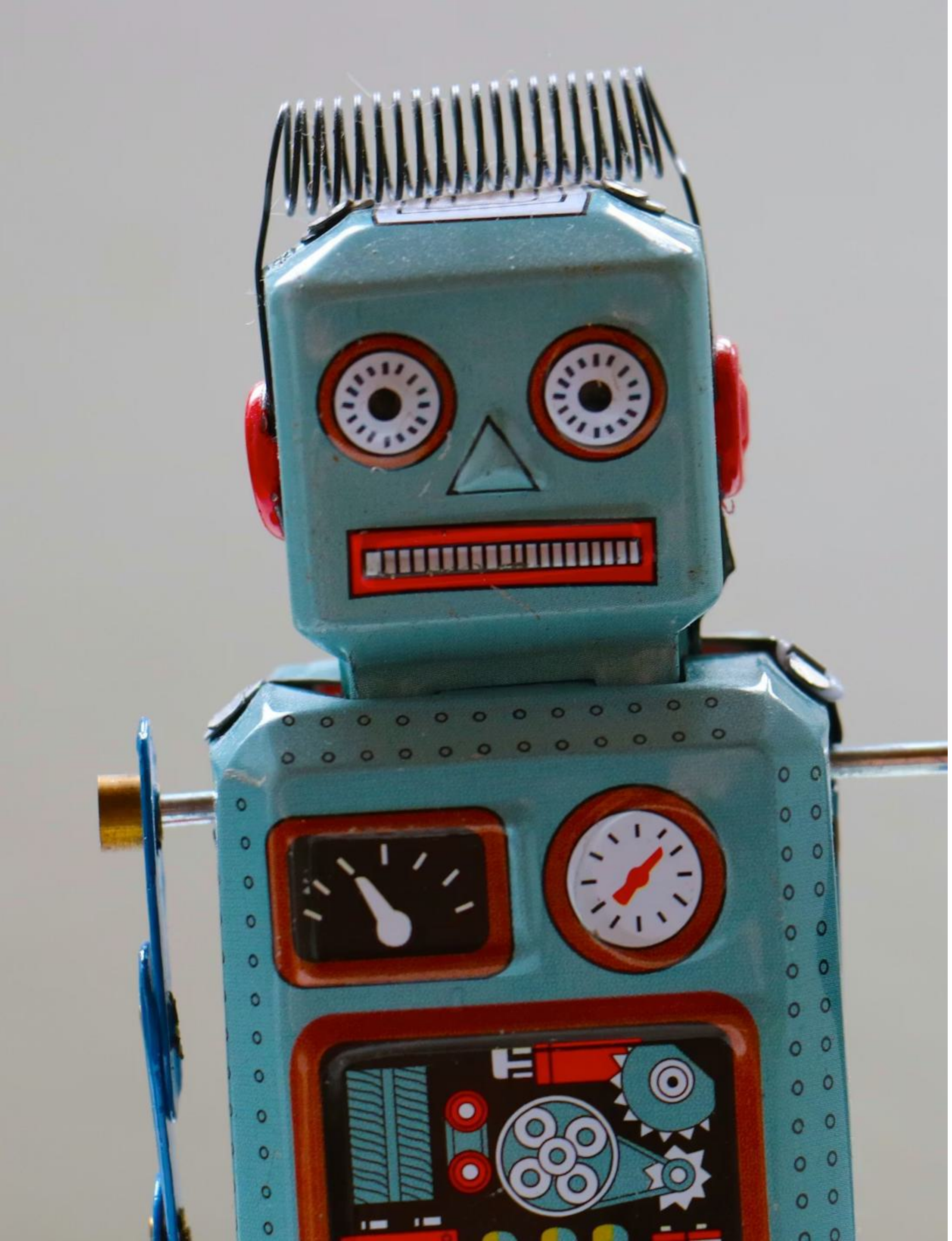
Contextualized pronunciation exercises

Repeated (say out loud)

Focus on form

Focus on meaning

Selected (click)



Conversational AI for L2 development

ConvAI for LL: recent changes

Access, studies, typology

An updated meta-analysis

Gaps, methods, data

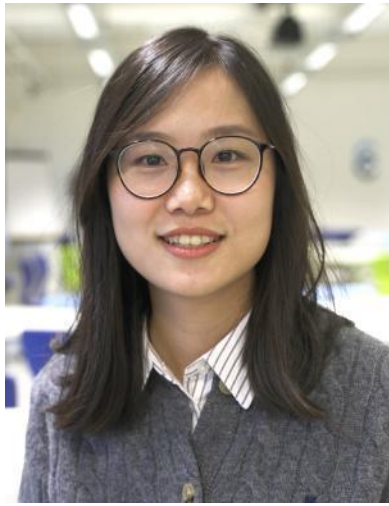
Effectiveness & moderators

Studies, systems, outcomes

Conclusions & insights

An updated meta-analysis of the effectiveness of Conversational AI for L2 development (Wang et al., *in prep.*)

- Researcher: Zhaori Wang (KU Leuven & UCLouvain)
- Inclusion criteria
  - 2010-2025
  - All conversational AI/DBCALL
    - + talking robots, + virtual reality
    - if ≥ repeated response (excluding point-and-click)
    - LLM vs non-LLM



Method: **Literature search**

- **6 major databases**: Web of Science, Scopus, Taylor & Francis, ProQuest, ERIC, CNKI
- Hand-search of CALL journals + educational technology journals
- Conference proceedings, reports, PhD dissertations
- 2010-2025 (August)
- K = **3 405 records** after deduplication

Method: **Inclusion criteria**

Dimension	**Include**	**Exclude**
Technology	Automated dialogue system	Wizard-of-Oz, CMC
Interaction	Meaningful L2 dialogue	Item-based exercises
Outcomes	Learning effects	Perception data only
Reporting	Sufficient statistics	Insufficient for ES

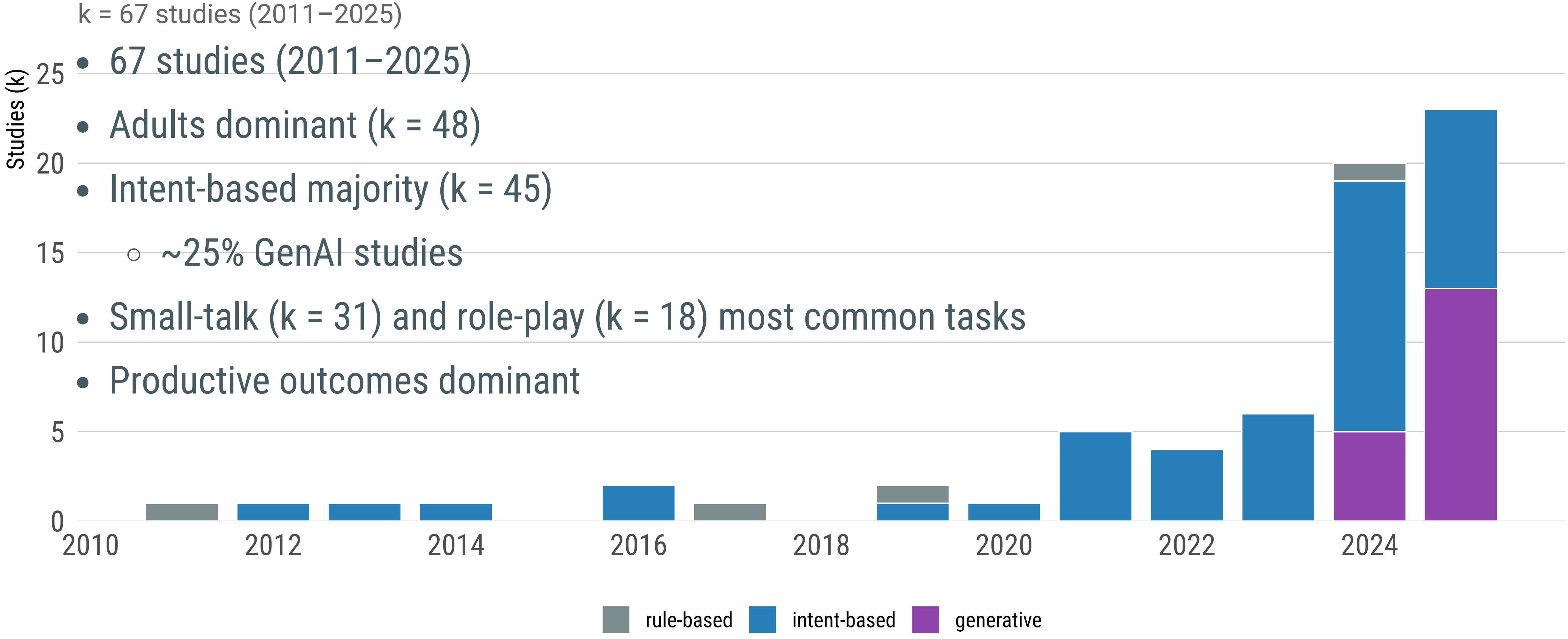
- Final dataset: **67 studies · 237 effects · 4 544 participants**
  - vs. 2022: 17 studies • 100 effects • 803 participants

Methods: **effect size calculation & statistical modeling**

- **Effect size**
  - **Hedges' *g*** (corrected Cohen's *d* for small samples)
  - Morris & DeShon (2002) formulas to align IG, IGPP, and PP designs
  - Benchmarks (Plonsky & Oswald, 2014): **0.4 = small · 0.7 = medium · 1.0 = large**

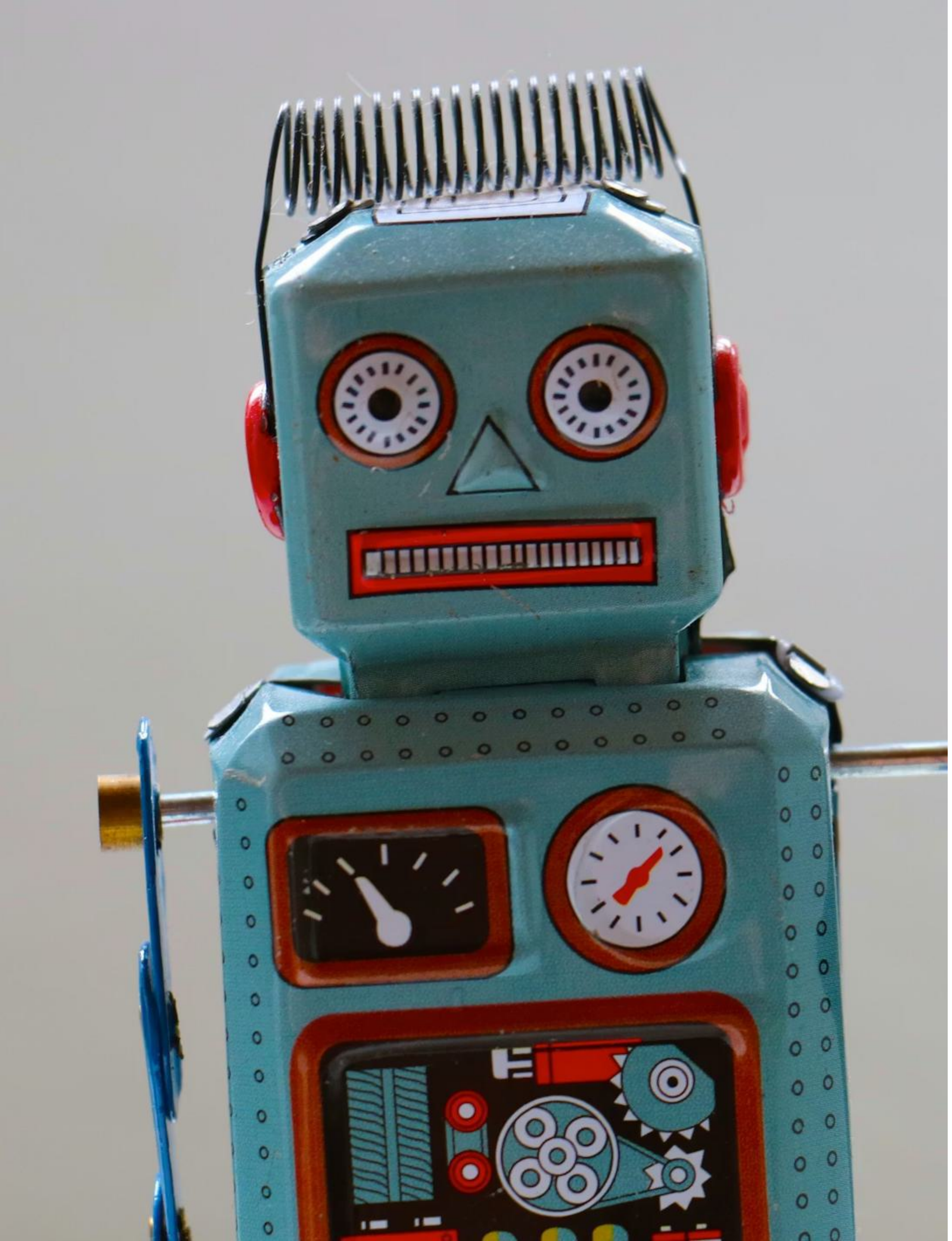
- **Statistical modelling**
  - **Three-level meta-analysis** (`metafor::rma.mv`) (Viechtbauer, 2010)
    - Level 1: sampling variance (per effect)
    - Level 2: within-study variance (across outcomes)
    - Level 3: between-study variance (across studies)

Dataset overview



k = 67 studies (2011–2025)

- 67 studies (2011–2025)
- Adults dominant (k = 48)
- Intent-based majority (k = 45)
  - ~25% GenAI studies
- Small-talk (k = 31) and role-play (k = 18) most common tasks
- Productive outcomes dominant



Conversational AI for L2 development

ConvAI for LL: recent changes

Access, studies, typology

An updated meta-analysis

Gaps, methods, data

Effectiveness & moderators

Studies, systems, outcomes

Conclusions & insights

Photo by Emilipothèse on Unsplash

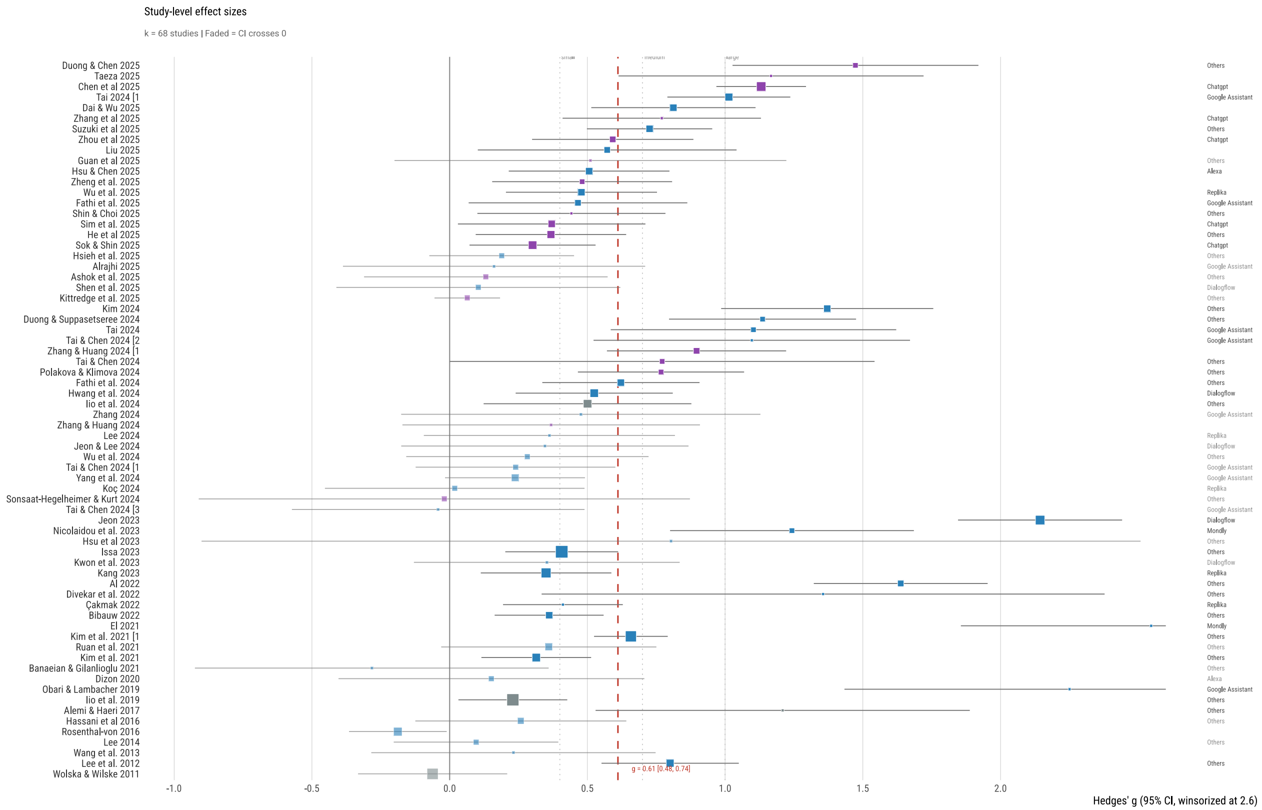
Results: **overall effect**

$$g = 0.61$$

$$95\%\ \mathrm{CI} = [0.48,\ 0.74]$$

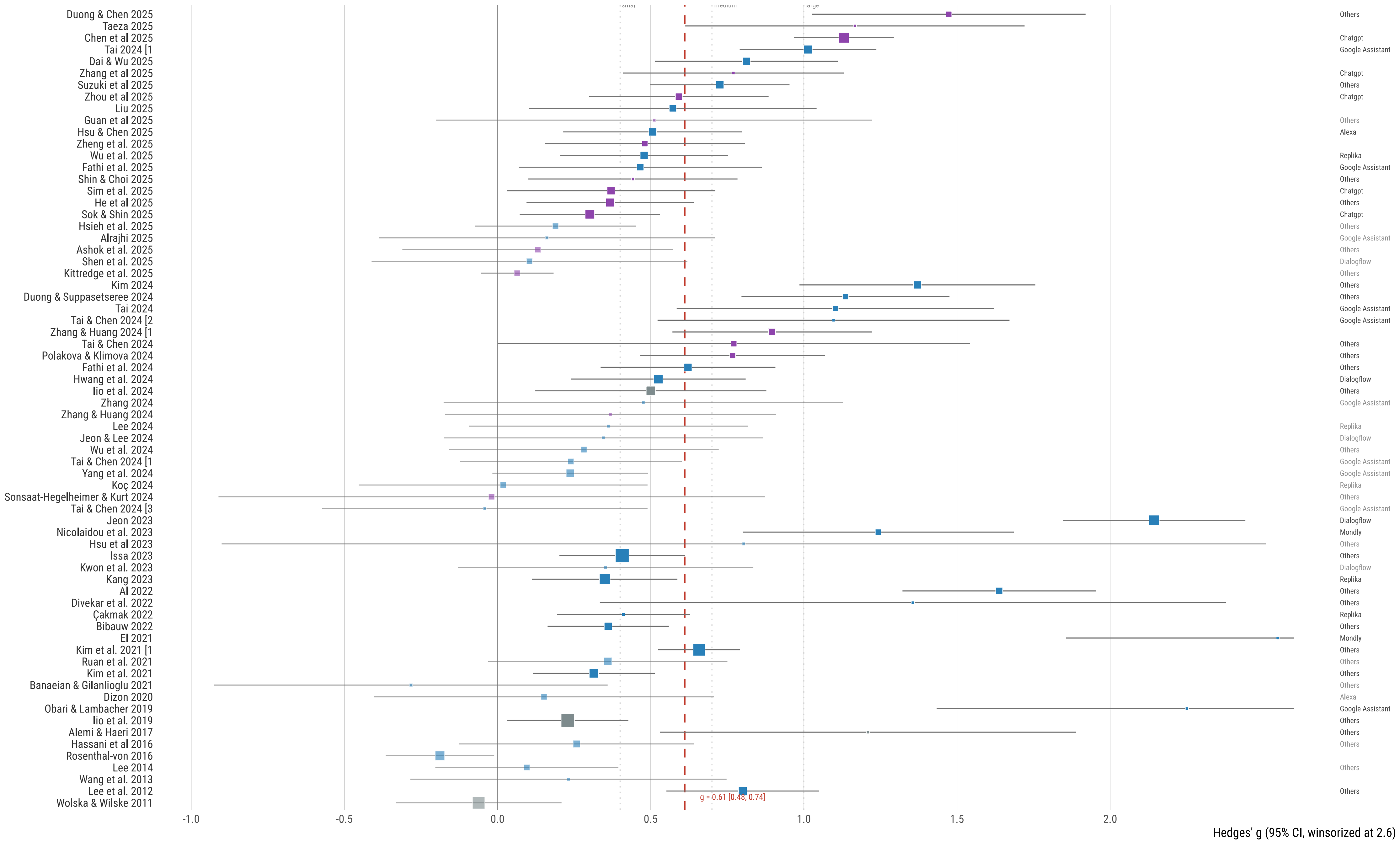
$$p < .001$$

- **Medium-sized positive effect**
- Dialogue-based CALL improves L2 performance overall
- Robust across design types



Study-level effect sizes

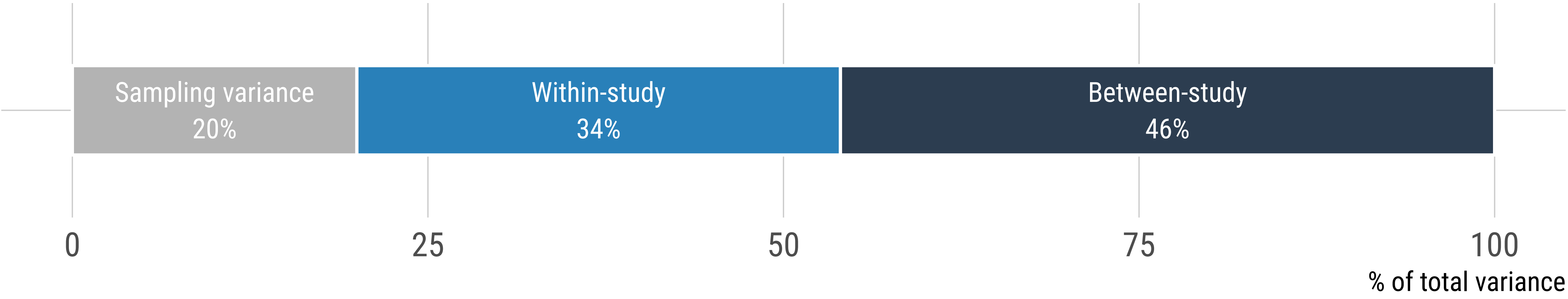
k = 68 studies | Faded = CI crosses 0



Results: **heterogeneity**

Variance decomposition

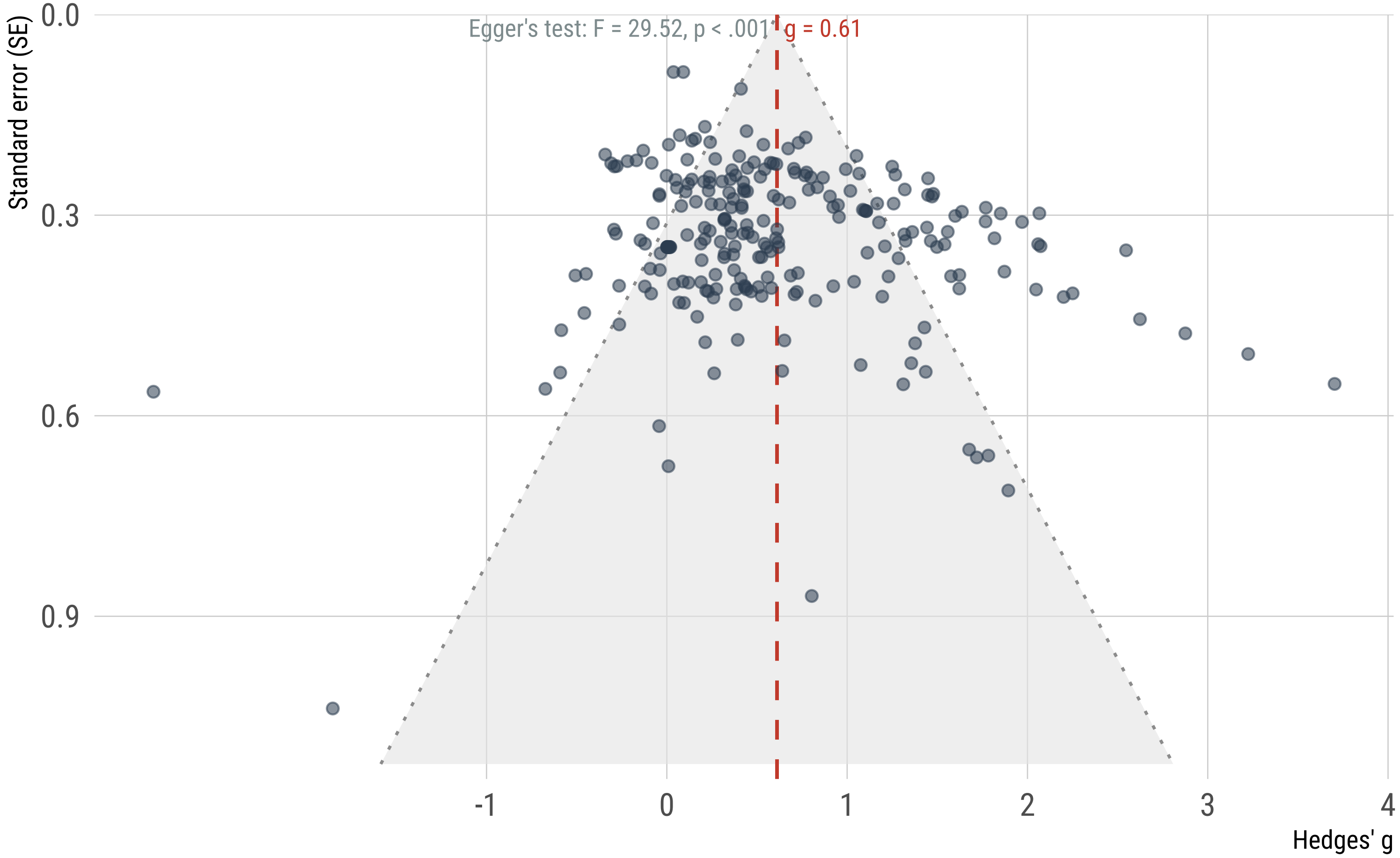
$Q(236) = 1057.27, p < .001$



- → Substantial variance at both levels
- → Moderator analyses justified

Results: **publication bias**

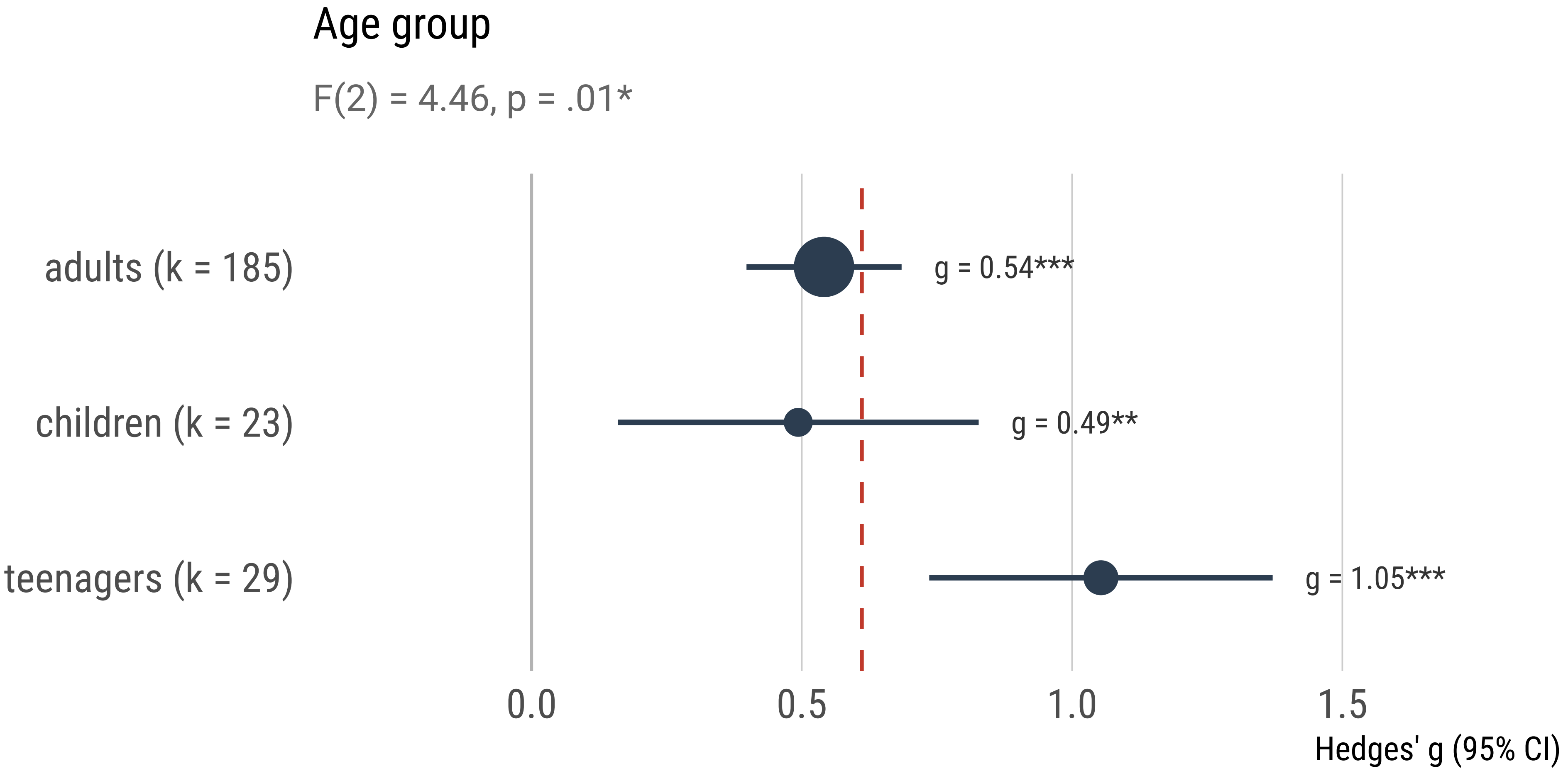
- Slight right-side asymmetry
- Egger's regression test:
  F = 29.52, *p* < .001
- Grey literature included;
  k = 67 reduces risk
- Interpret with caution: effect sizes may be slightly inflated



Funnel plot

k = 237 effects; slight right-side asymmetry

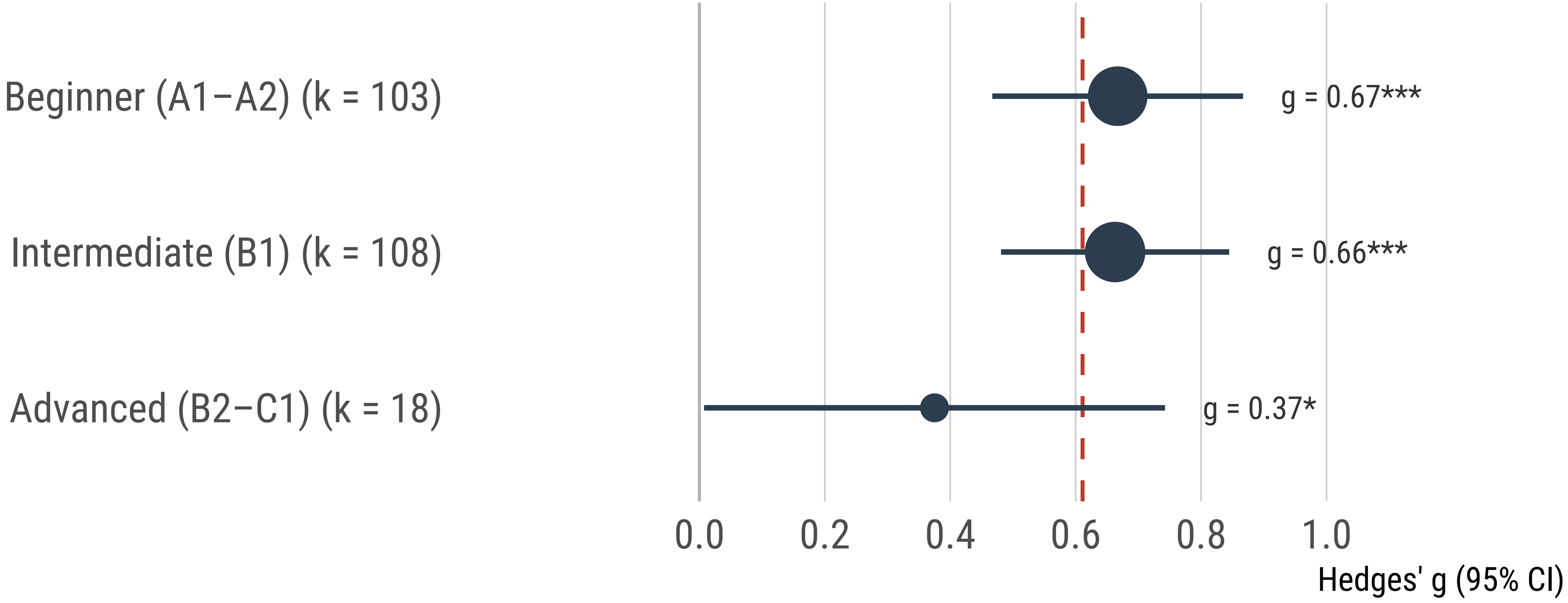
Moderators: **study context**

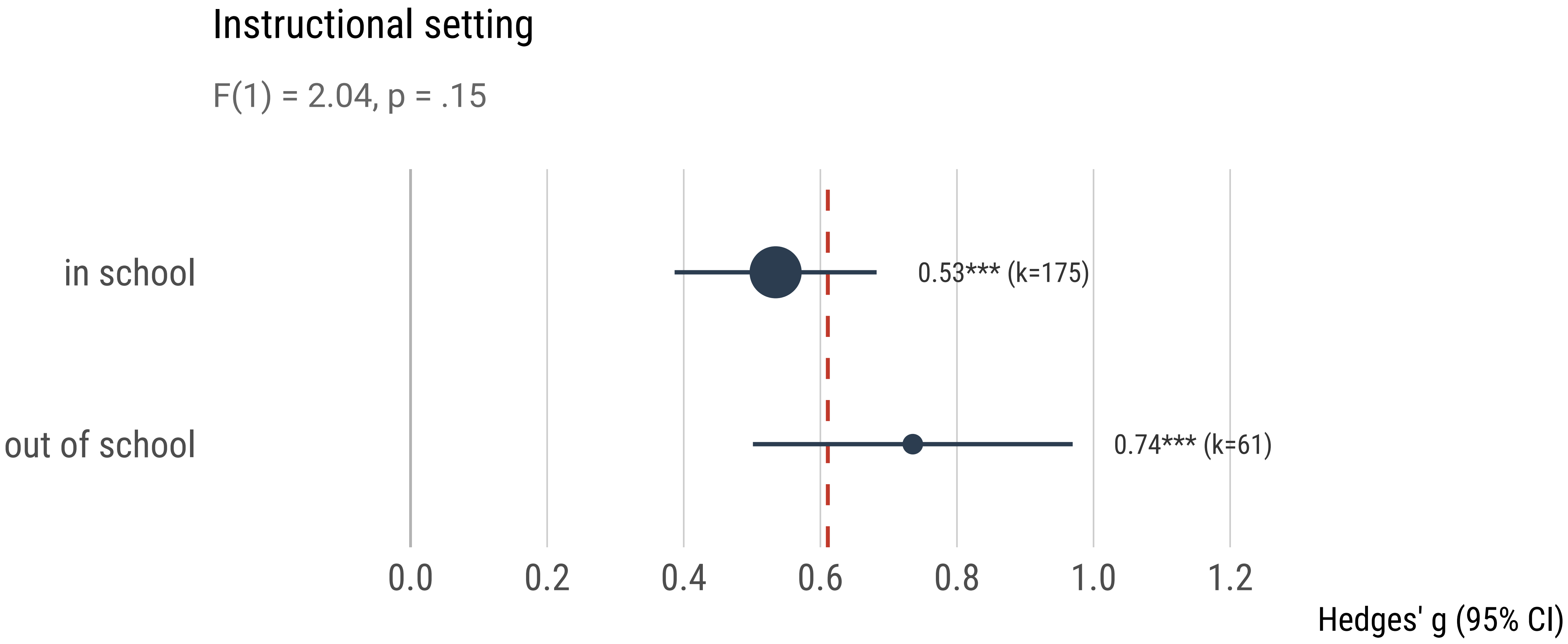


Moderators: **study context: L2 proficiency level**

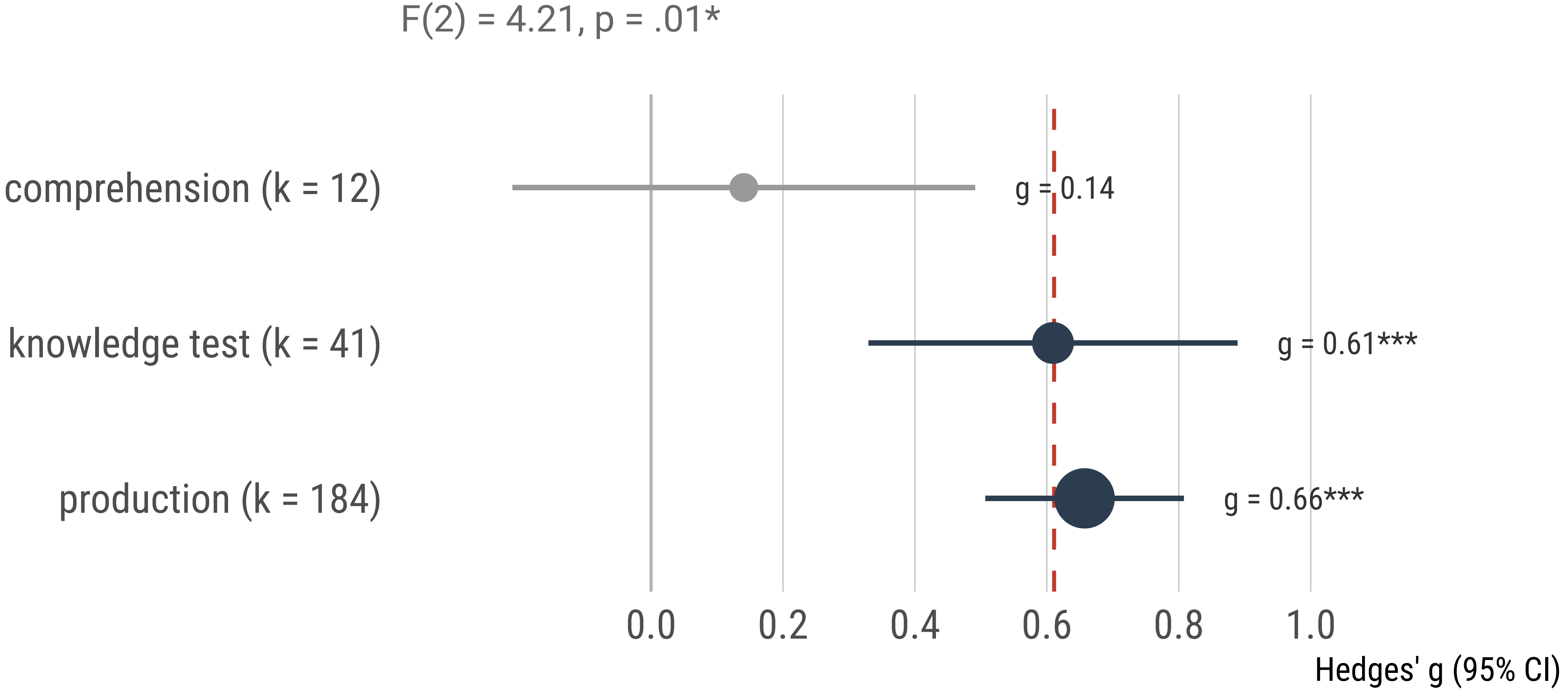
F(2) = 1.61, p = .26 (n.s.)

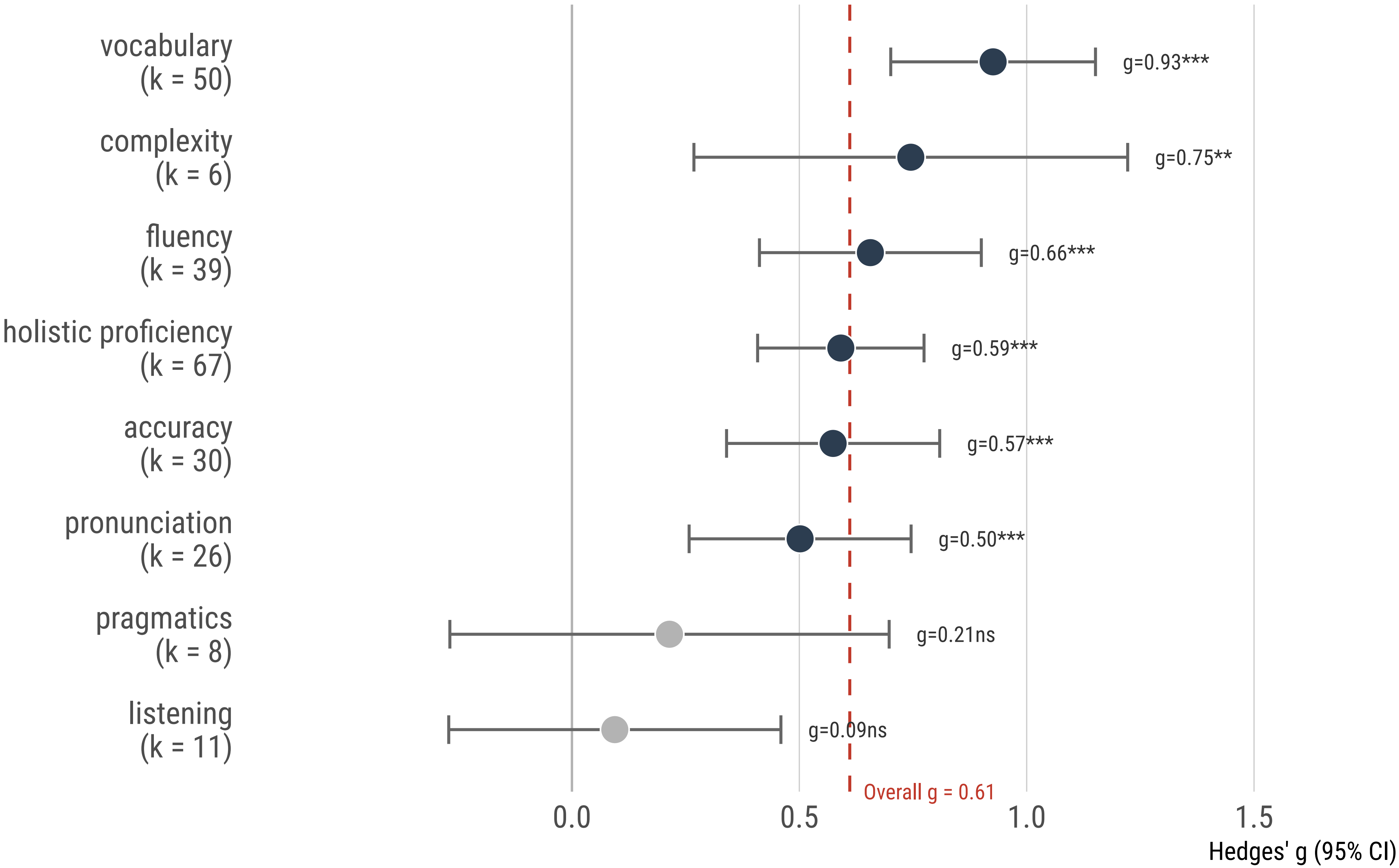
- → Chatbot practice works across proficiency levels, but **evidence thins out at advanced**




Instructional setting
F(1) = 2.04, p = .15
in school
0.53*** (k=175)
out of school
0.74*** (k=61)
0.0
0.2
0.4
0.6
0.8
1.0
1.2
Hedges' g (95% CI)

Moderators: **type of outcome**

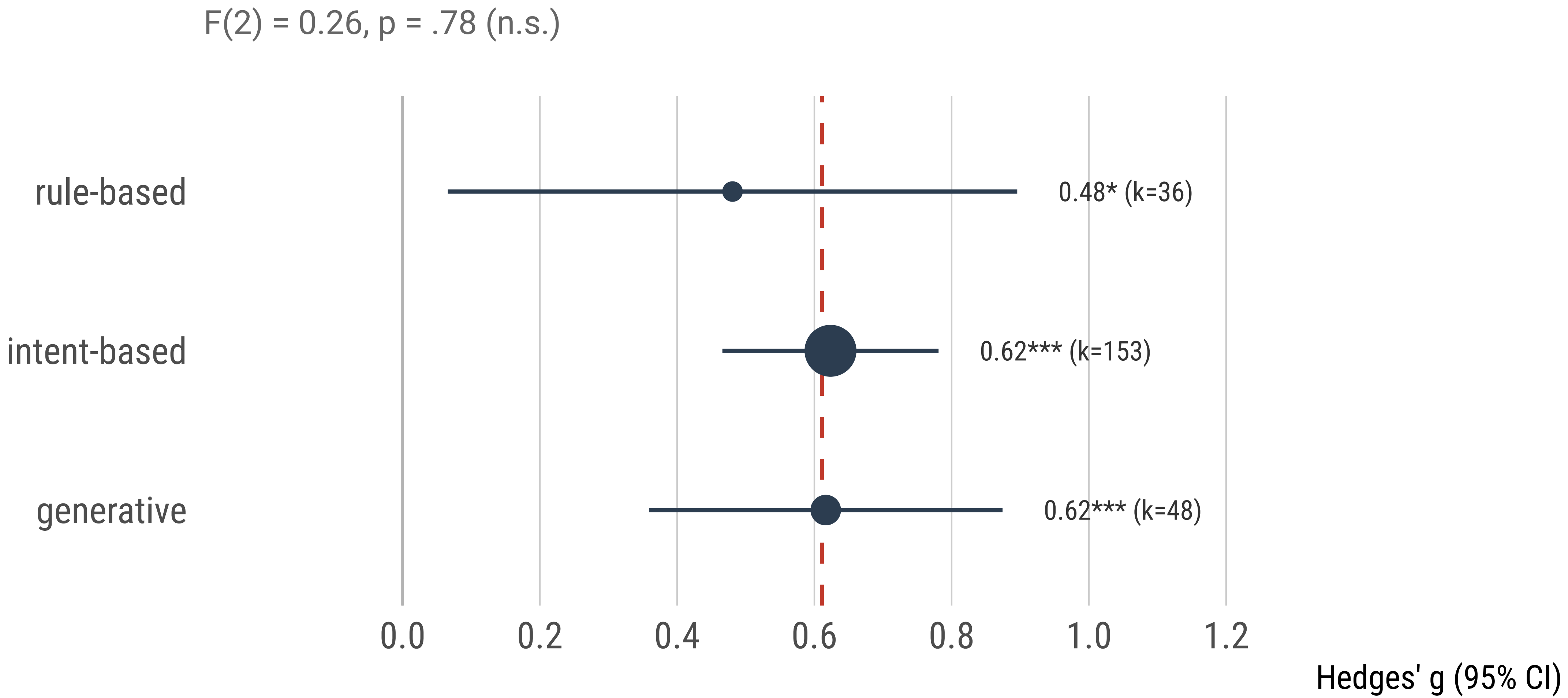




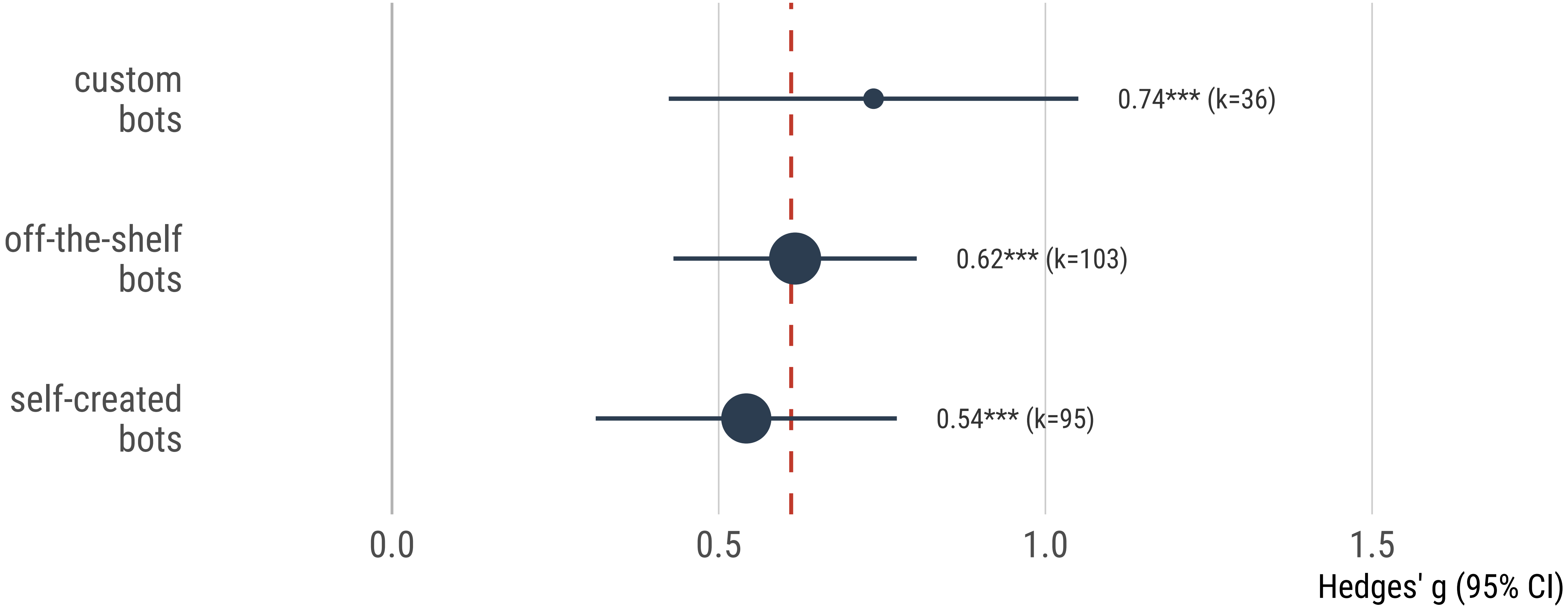
Effect size by outcome variable

Filled = p < .05 | Sorted by effect size

Moderators: **technology**



- → Technology type matters (much) less than **pedagogical implementation**

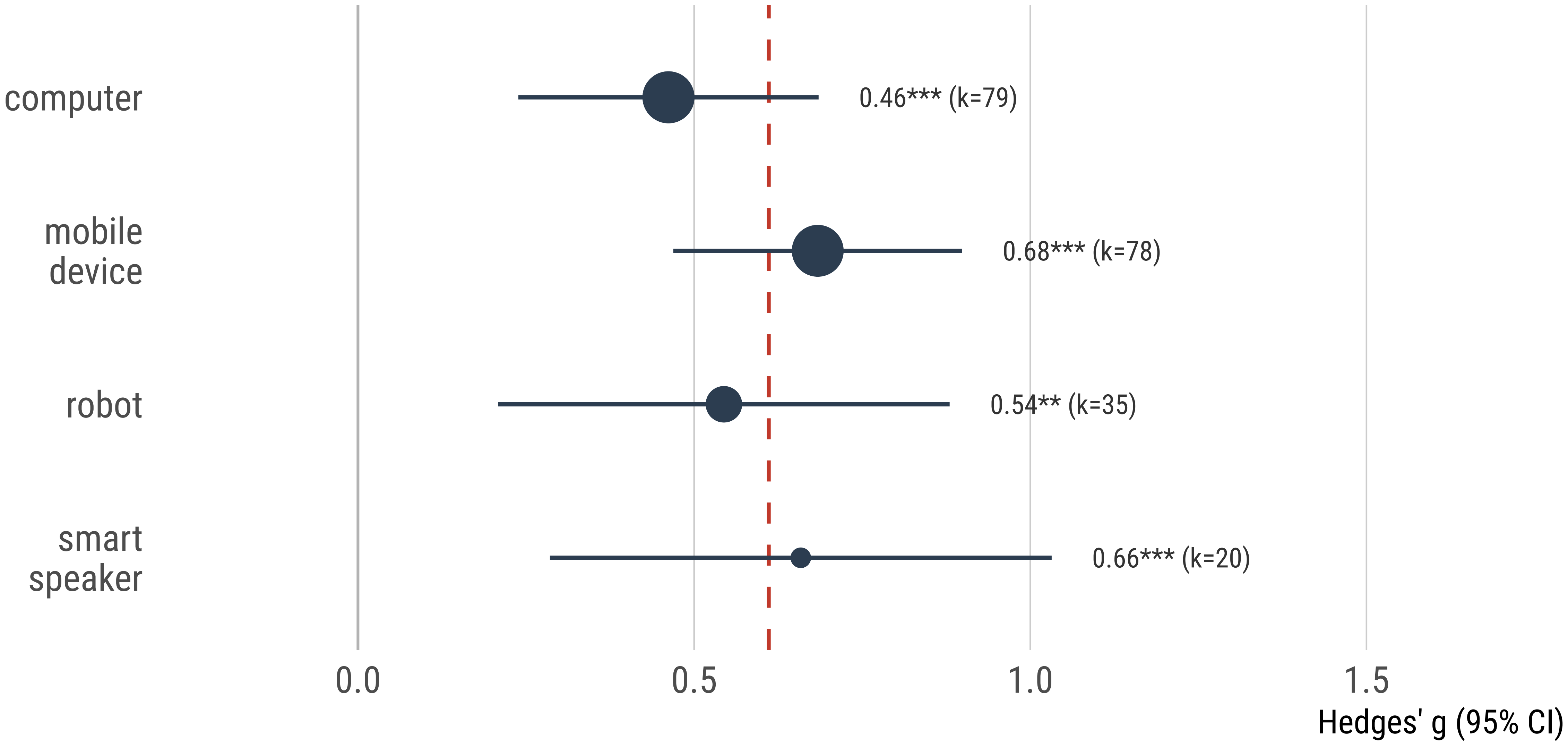


Design approach

F(2) = 0.48, p = .61 (n.s.)

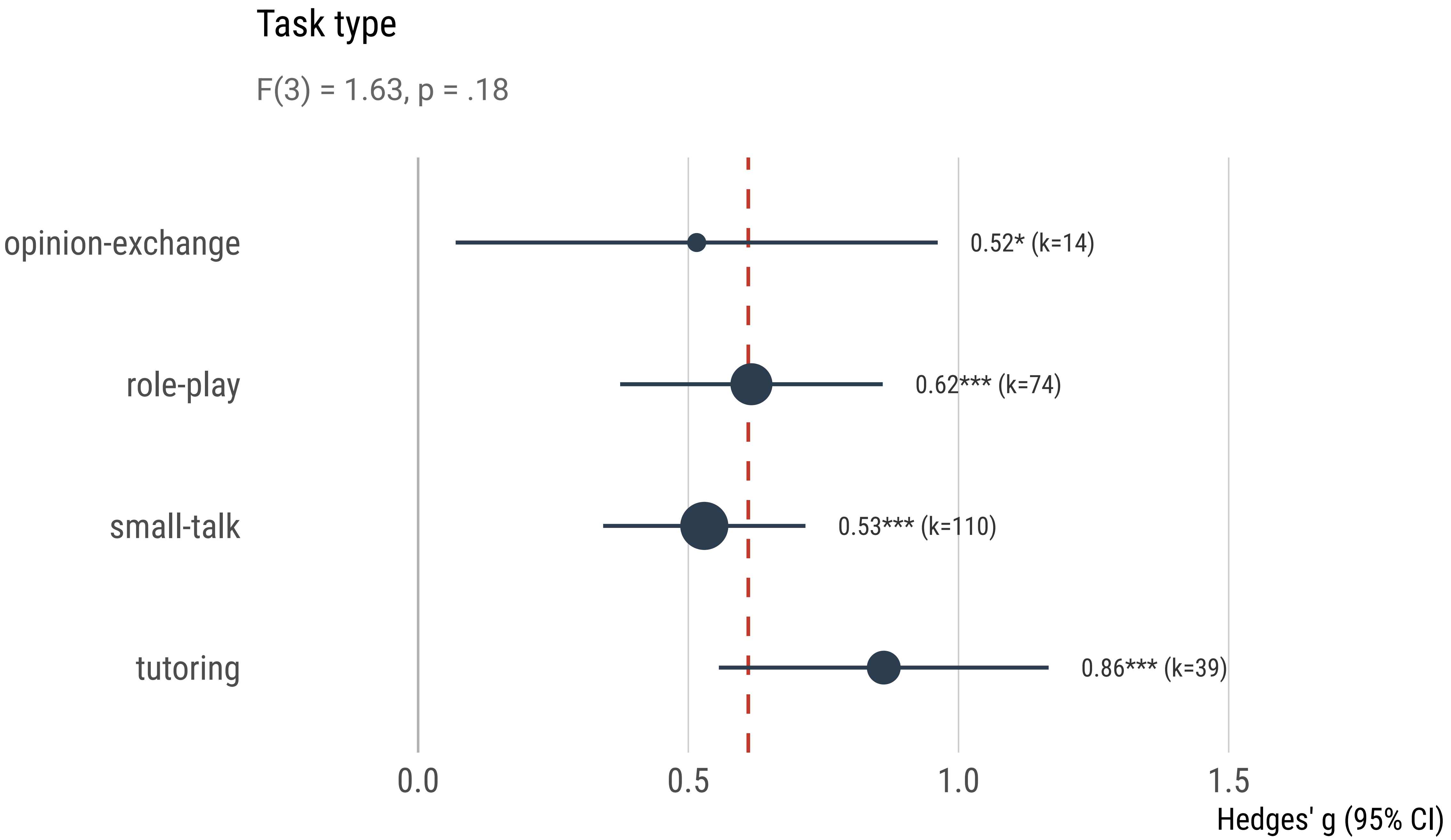
Device

F(3) = 0.72, p = .53 (n.s.)



Moderators: **pedagogical implementation**

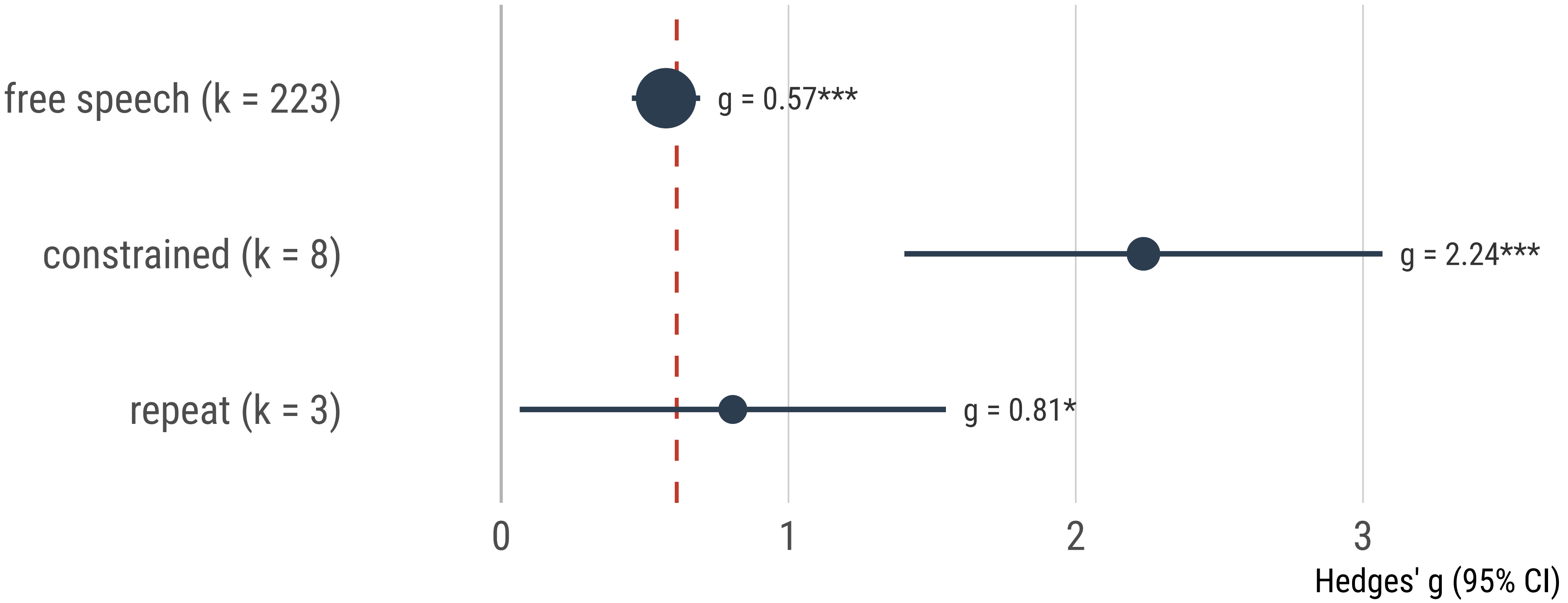
- Scaffolding: n.s. (yes g = 0.68 vs no g = 0.58)
- Corrective feedback: n.s.
- Collaborative learning: g = 1.25 (k = 5, preliminary)

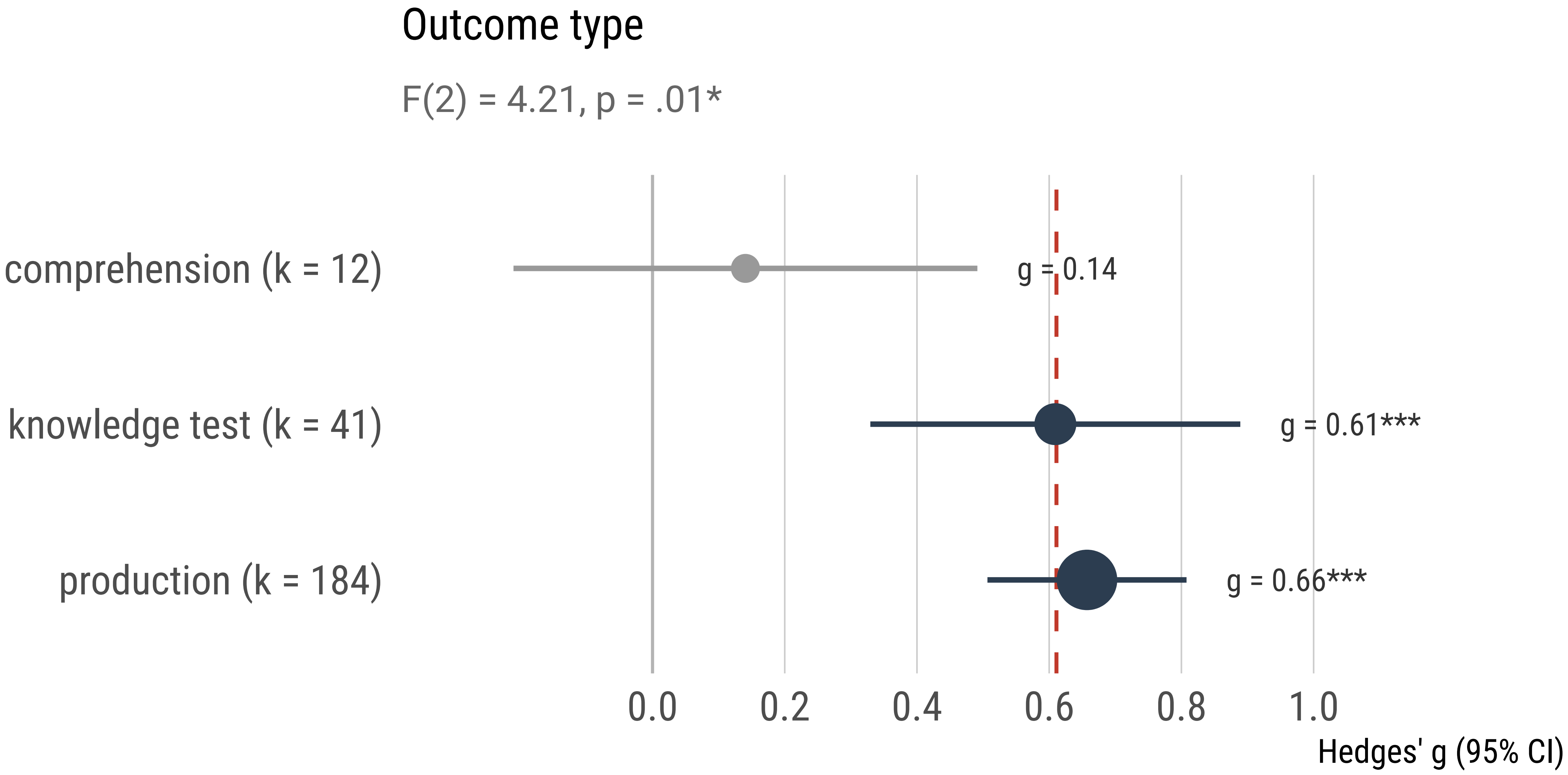

Task type
F(3) = 1.63, p = .18
opinion-exchange
role-play
small-talk
tutoring
0.52* (k=14)
0.62*** (k=74)
0.53*** (k=110)
0.86*** (k=39)
0.0
0.5
1.0
1.5
Hedges' g (95% CI)

Moderators: **outcome variables**

Constraints on learner output

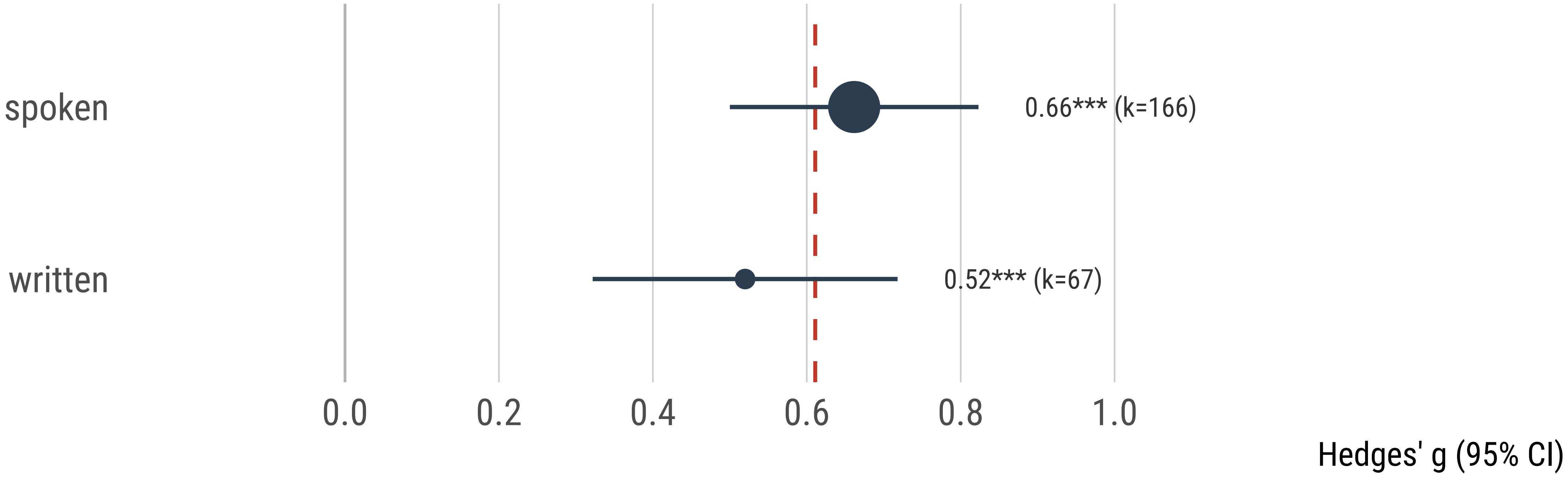
F(2) = 7.72, p < .001***




Outcome type
F(2) = 4.21, p = .01*
comprehension (k = 12)
knowledge test (k = 41)
production (k = 184)
g = 0.14
g = 0.61***
g = 0.66***
0.0
0.2
0.4
0.6
0.8
1.0
Hedges' g (95% CI)

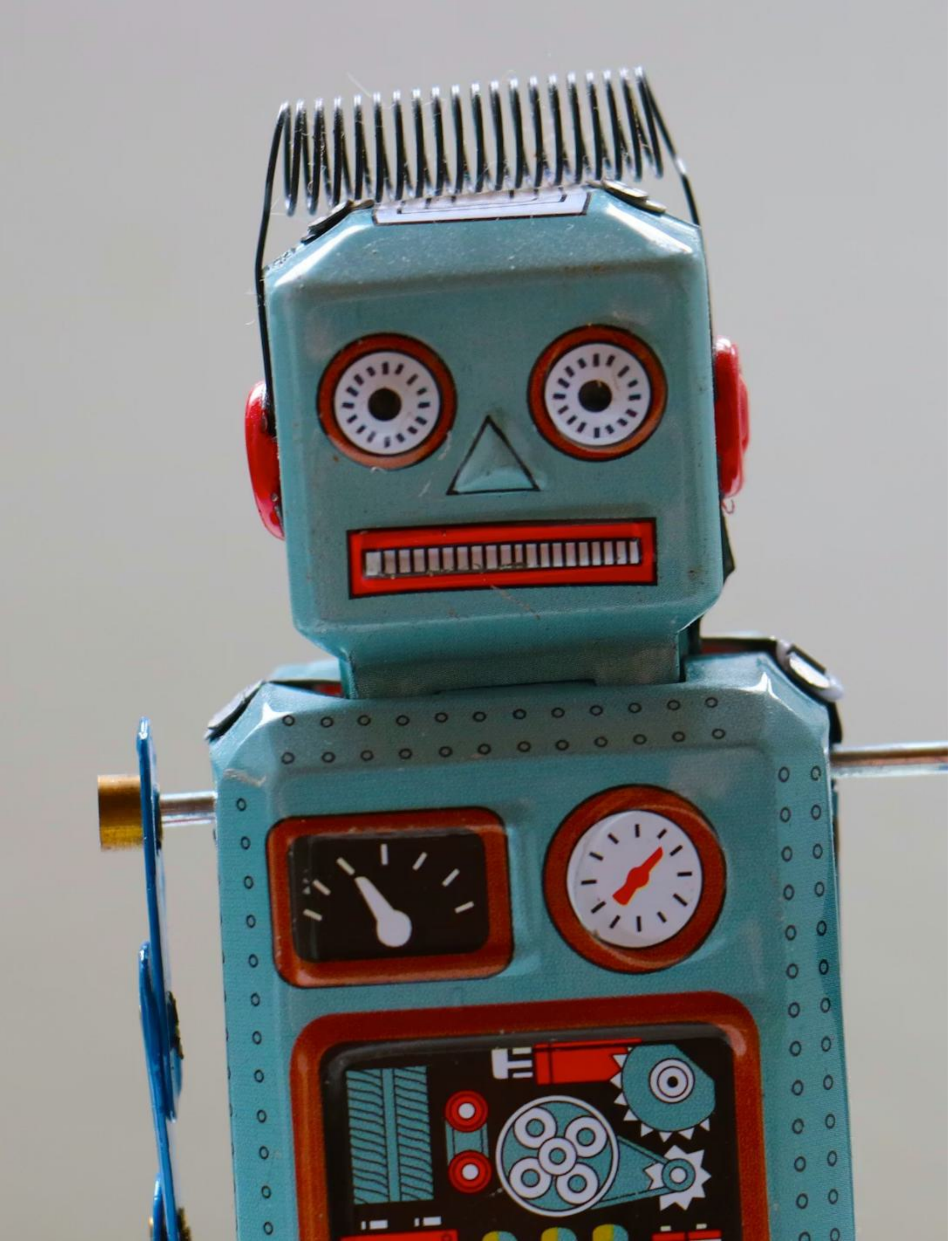
Outcome modality

F(1) = 1.33, p = .26 (n.s.)



Moderators: **outcome/assessment modality**

- **Modality matching** *(F(1) = 10.36, p = .002)*
  - Matched: g = 0.65
  - Mismatched: g = 0.26
  - → **2.5× larger effect when practice and test modality align**
- **Transfer direction**
  - Written → spoken: g = 0.51* (significant)
  - Spoken → written: n.s.



Conversational AI for L2 development

ConvAI for LL: recent changes

Access, studies, typology

An updated meta-analysis

Gaps, methods, data

Effectiveness & moderators

Studies, systems, outcomes

Conclusions & insights

Implications for automated speaking assessment

Chatbot as standardised interlocutor:

- Eliminates interviewer variability → fairer, more replicable elicitation
- Same prompts, pacing and patience for every learner
- LLM-powered chatbots can sustain near-human conversation quality at scale
- Caveat: current evidence is from *practice* studies; chatbot-as-examiner validity (construct, consequential) remains to be established

Discussion: **key takeaways**

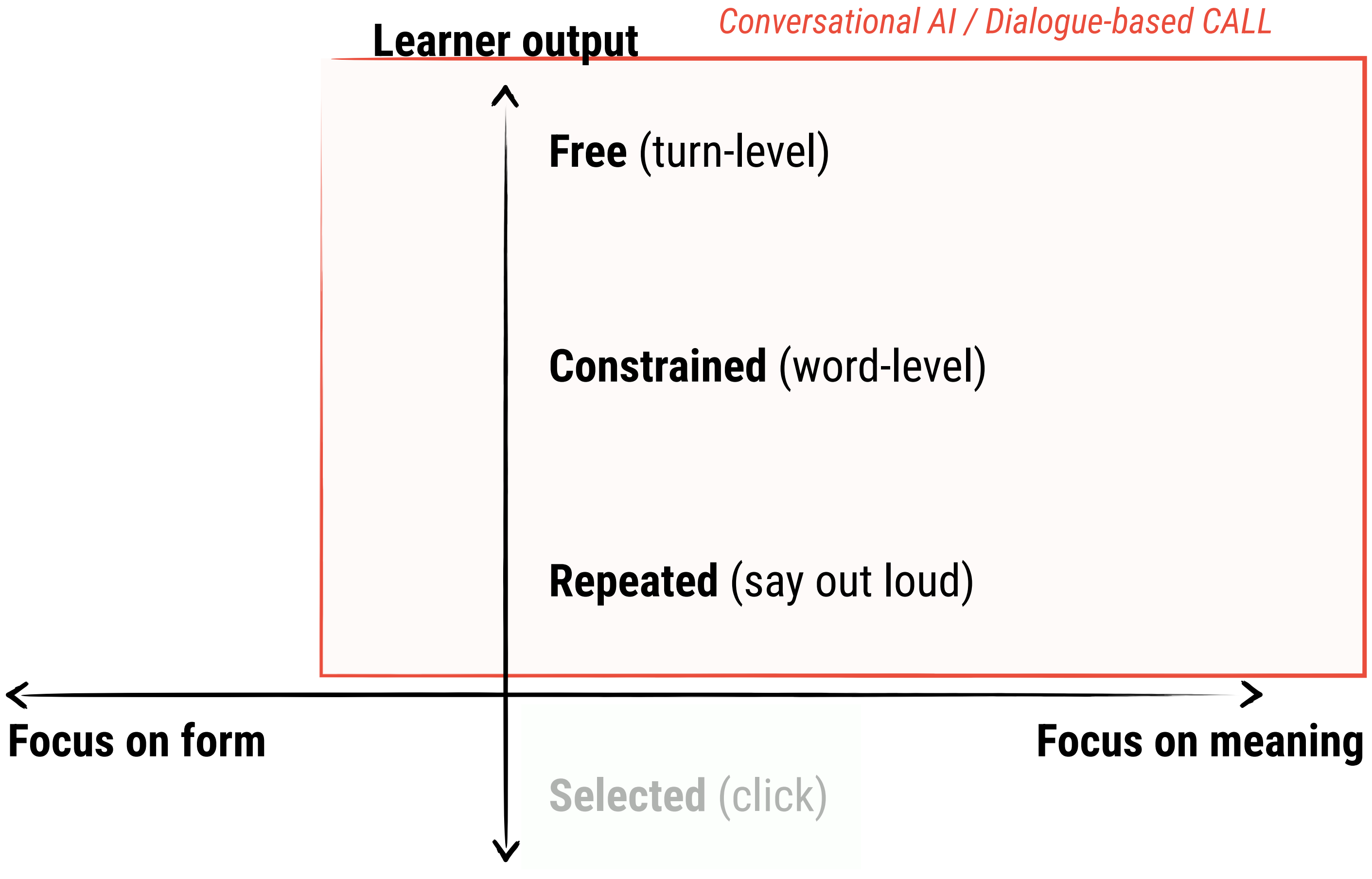
1. **Robust medium effect** across 14 years of research
2. **Vocabulary, fluency and syntactic complexity** are most sensitive outcomes
3. **GenAI ≠ bigger effects**: pedagogical design matters more than technology but LLMs open up access and possibilities for research.
4. **Collaborative use** and **delayed measurement** associated with larger effects
5. **Matching practice and test modality** doubles observed effect size
6. Too many studies with "opportunities" outcome variables
    - Mismatch between rationale for ConvAI and outcome variables.
7. It's all about the **implementation**! Technical, instructional, classroom integration

Thank you !
Questions ? Comments ?
Disagreements and
suggestions welcome.

Serge Bibauw
serge.bibauw@uclouvain.be
https://serge.bibauw.be

Zhaori Wang
zhaori.wang@kuleuven.be

UCLouvain



Images: Unsplash & authors from papers

References

Bibauw, S., François, T., & Desmet, P. (2019). Discussing with a computer to practice a foreign language: Research synthesis and conceptual framework of dialogue-based CALL. Computer Assisted Language Learning, 32(8), 827–877. https://doi.org/10.1080/09588221.2018.1535508

Bibauw, S., Van den Noortgate, W., François, T., & Desmet, P. (2022). Dialogue systems for language learning: A meta-analysis. Language Learning & Technology, 26(1), 1–24. https://doi.org/10125/73488

Hou, Z., & Min, S. (2025). Dialogue-based computer-assisted language learning systems for second language speaking development: A three-level meta-analysis. *ReCALL*. https://doi.org/10.1017/S0958344025100268

Howatt, A. P. R. (1984). *A history of English language teaching*. Oxford University Press.

Long, M. H. (1996). The role of the linguistic environment in second language acquisition. In W. C. Ritchie & T. K. Bhatia (Eds), *Handbook of second language acquisition* (pp. 413–468). Academic Press.

Morris, S. B., & DeShon, R. P. (2002). Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. *Psychological Methods, 7*(1), 105–125. https://doi.org/10.1037/1082-989X.7.1.105

Norris, J. M., & Ortega, L. (2000). Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis. *Language Learning, 50*(3), 417–528. https://doi.org/10.1111/0023-8333.00136

Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning, 64*(4), 878–912. https://doi.org/10.1111/lang.12079

Spada, N., & Tomita, Y. (2010). Interactions between type of instruction and type of language feature: A meta-analysis. *Language Learning, 60*(2), 263–308. https://doi.org/10.1111/j.1467-9922.2010.00562.x

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. Journal of Statistical Software, 36(3), 1–48. https://doi.org/10.18637/jss.v036.i03

Wang, F., Cheung, A. C. K., Neitzel, A. J., & Chai, C. S. (2024). Does chatting with chatbots improve language learning performance? A meta-analysis of chatbot-assisted language learning. *Review of Educational Research*, 00346543241255621. https://doi.org/10.3102/00346543241255621

Wang, Z., Bibauw, S., Metwali, S., Noreillie, A.-S., & Desmet, P. (*in prep*). Hype or hope? A meta-analysis of conversational chatbots on L2 learning.