

The AI Teacher Test

Measuring the Pedagogical Ability of Blender and GPT-3 in Educational Dialogues



Anaïs Tack

Stanford University

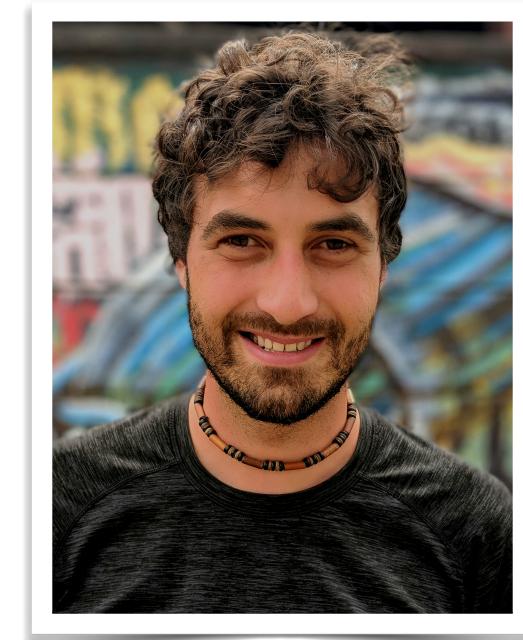
atack@cs.stanford.edu

<https://sig-edu.org>

Chris Piech

Stanford University

piech@cs.stanford.edu



Conversational AI for Education

- various roles (e.g., intelligent tutors, service-oriented assistants, ...)
- different objectives (e.g., improving student skills, boosting student motivation, ...)
- most prevalent type = the **AI teacher**
 - helping a student with skill improvement
 - providing more opportunities to practice

Are We There Yet? A Systematic Literature Review on Chatbots in Education (Wollny et al., Frontiers in Artificial Intelligence 2021)

Blender

(Roller et al., EACL 2021)

- Built on **transformers**
(Vaswani et al., 2017)
- Trained on **blended skill talk**
(Smith et al., 2020)
 - Being knowledgeable
 - Being empathetic
- Model sizes: **90M, 400M, 1B, 3B, 9B**

Human

Generative BST 9.4B

If you could have dinner with one person, dead or alive, who would that be?

That's a tough one. I think I'd have to go with Steve Jobs. I'd love to pick his brain.

Ah, interesting choice. What do you know about Steve Jobs?

He was one of the most influential people in the history of technology. He was a visionary.

What questions would you ask him?

I'd want to know how he came up with the idea for the iPod. It was so revolutionary at the time.

Totally...the iPod changed the way everyone listens to music.

It really did. It's crazy to think that it was released in 2001. That was a long time ago.

Omg wow, that makes me feel old!

I know, right? Time flies by so fast. I wish I could go back to those simpler times.

Me too, friend, me too. Speaking of music, what's your favorite album to listen to lately?

I've been listening to a lot of Led Zeppelin lately. They're my favorite band.

GPT-3

(Brown et al., NeurIPS 2020)

- 175B-parameter model
- Powerful generative model
- Few-shot in-context learning from prompt
- Various tasks, also **chatbot**

The screenshot shows a user interface for interacting with GPT-3. At the top, there's a purple icon with a white speech bubble containing three dots, followed by the word "Chat". Below this are two buttons: "Conversation" (which is highlighted) and "Generation". A horizontal line separates this header from the main content area.

Prompt

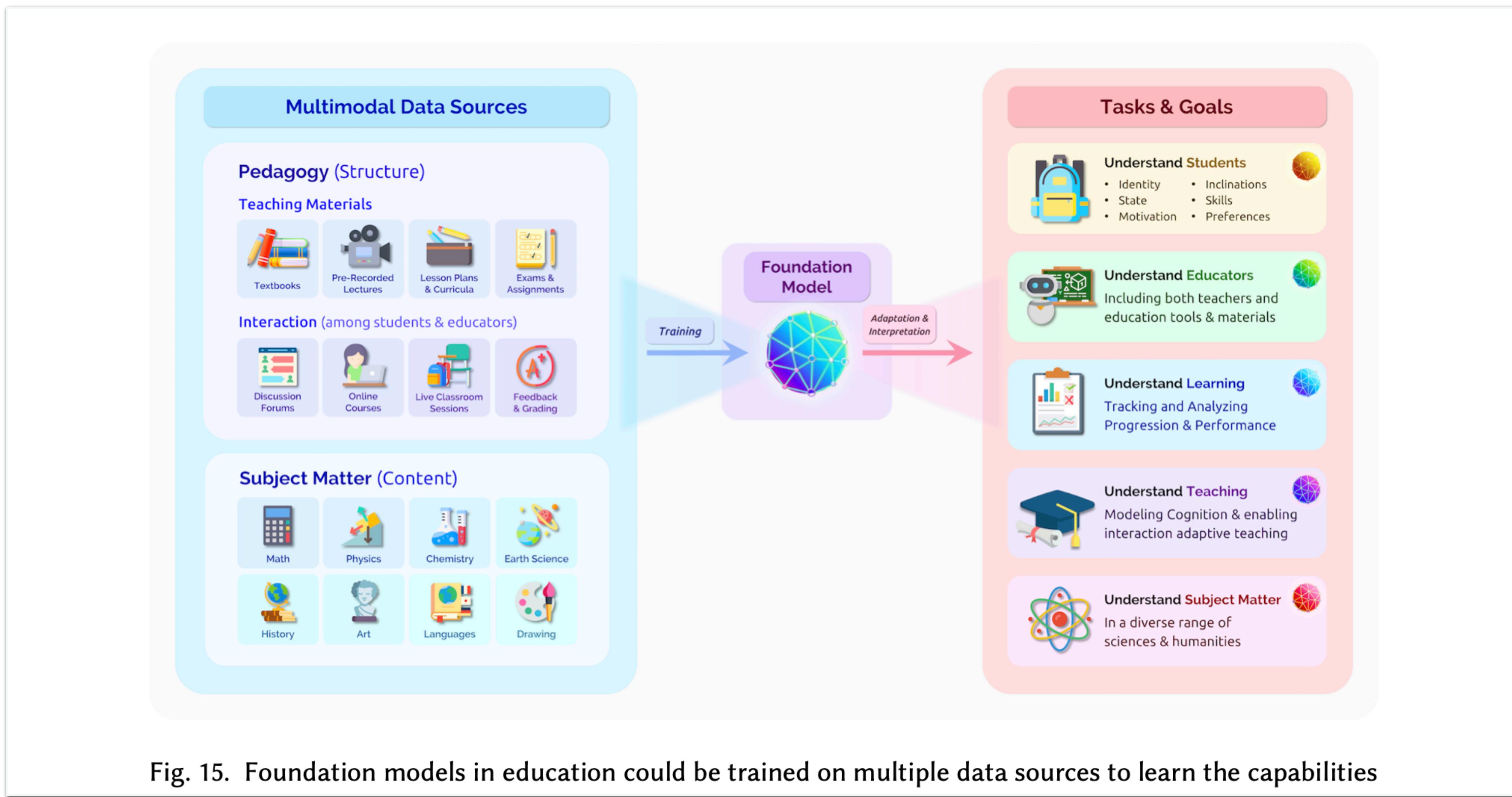
The following is a conversation with an AI assistant. The assistant is helpful, creative, clever, and very friendly.

Human: Hello, who are you?
AI: I am an AI created by OpenAI. How can I help you today?
Human: I'd like to cancel my subscription.
AI:

Sample response

Sure, I can help you cancel your subscription. Just let me know what type of subscription you have.

On the Opportunities and Risks of Foundation Models (Bommasani et al., arXiv 2021)



Foundation Models for Education

If we want to put foundation models into practice as AI teachers, we must determine whether they can:

- a) speak to students like a teacher
- b) understand students
- c) help students improve their understanding

On the Opportunities and Risks of Foundation Models (Bommasani et al., arXiv 2021)

Contributions

- A. AI Teacher Test Challenge
- B. A human-in-the-loop pairwise comparison test
 - 1) Real-world student-teacher dialogues
 - 2) Simulate agent responses (Blender, GPT-3)
 - 3) Measure pedagogical ability (3 dimensions) with Bayesian sampling (HMC)

<https://github.com/anaistack/ai-teacher-test>

The AI Teacher Test Challenge

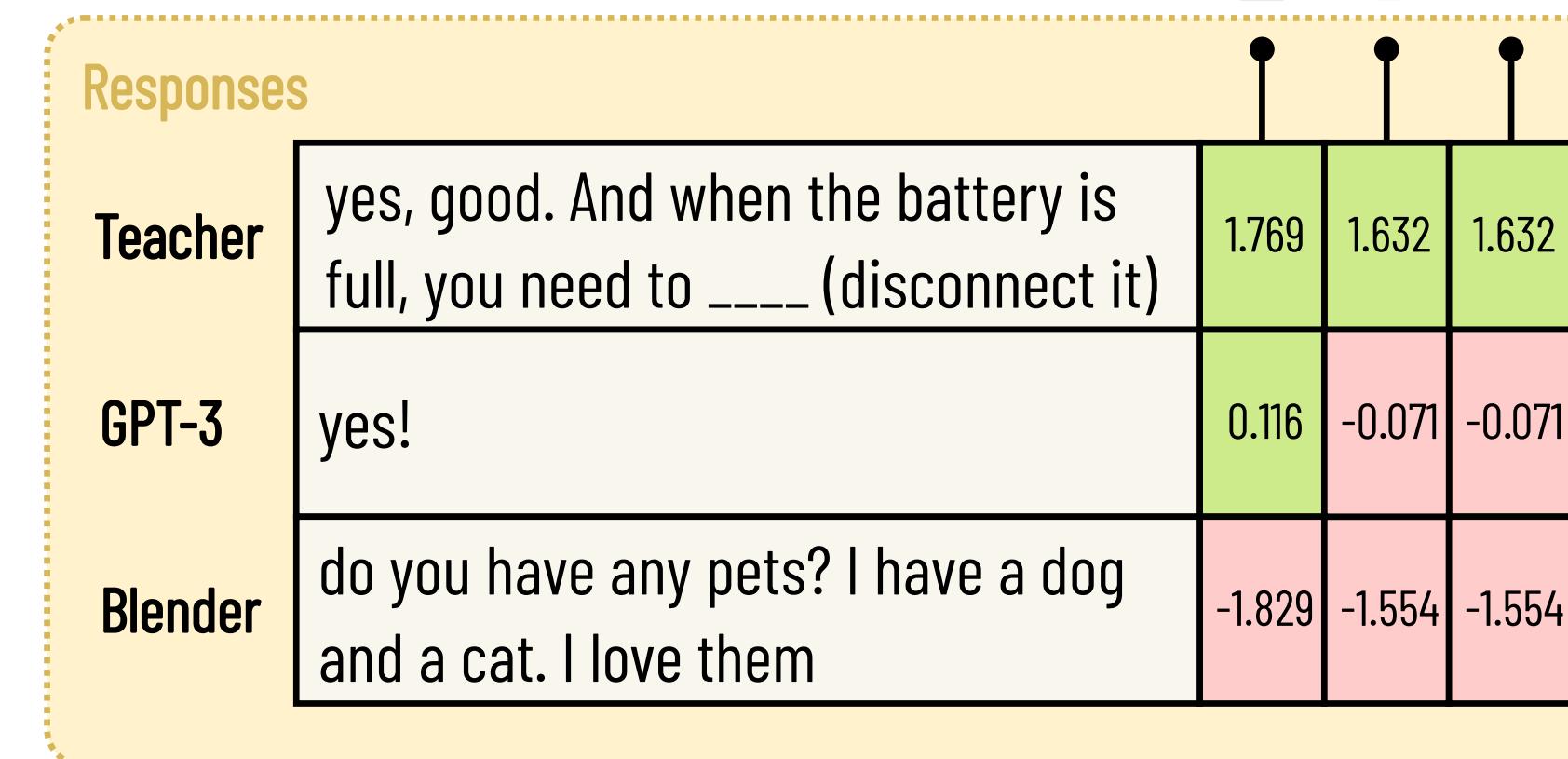
Challenge: Given a space of possible responses (either human or artificially generated), evaluate a reply in terms of pedagogical ability and estimate this score relative to other replies.

Desiderata:

1. evaluate a teacher agent's response in context
2. score the agent's response with respect to several pedagogical abilities
3. consider other possibilities (better or worse) and rank the teacher's response in comparison to these

The following is a conversation between a student and a teacher working on a language exercise.

Teacher	Yes, good! And to charge it up, you need to __ it __
Student	...
Teacher	connect to the source of electricity
Student	i understand
Teacher	plug it __?
Student	in



Real-World Student-Teacher Dialogues

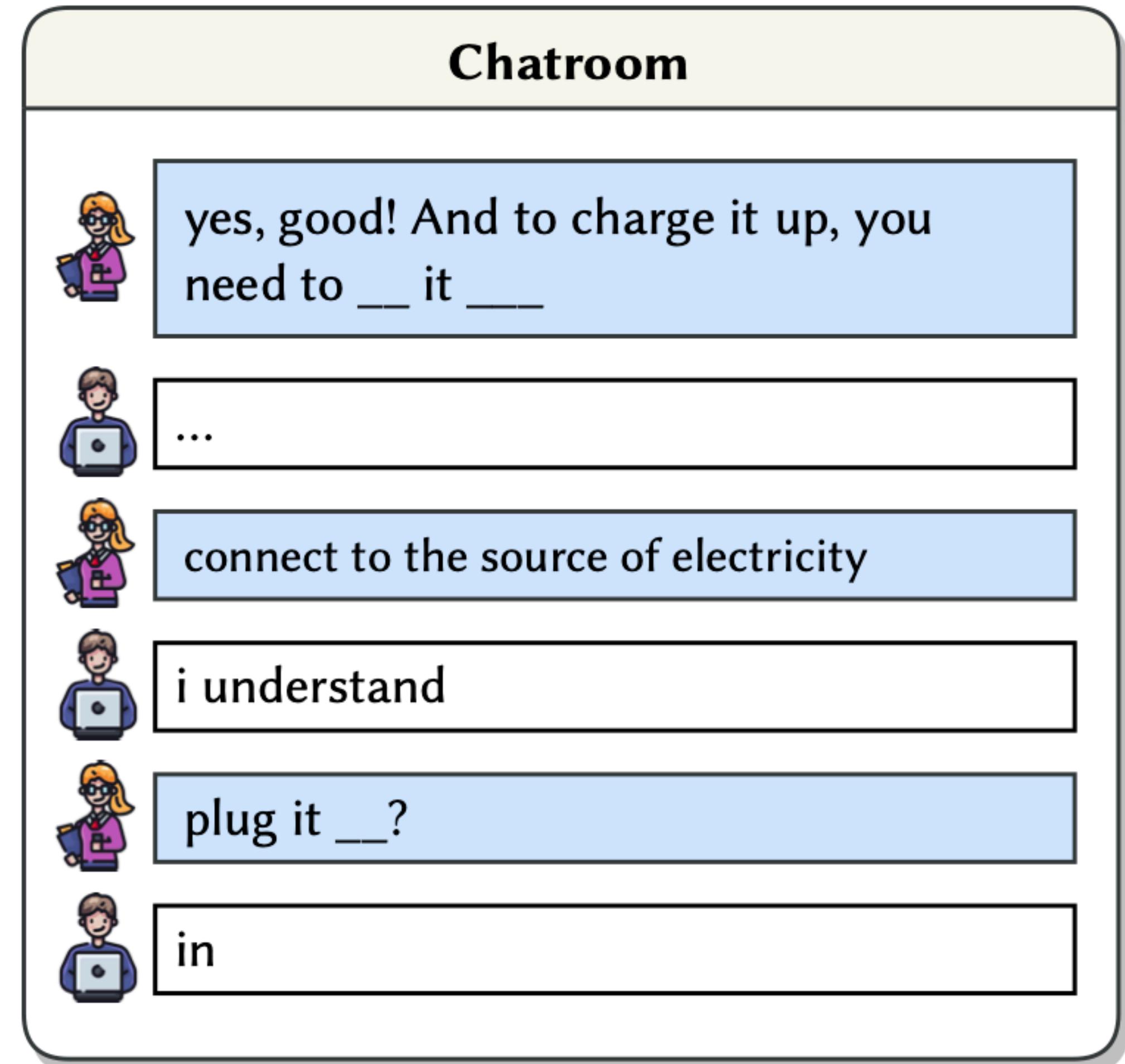
1. ***Teacher-Student Chatroom Corpus***
(Caines et al., NLP4CALL 2020)
2. ***Educational Uptake Dataset***
(Demszky et al., ACL 2021)



Teacher-Student Chatroom Corpus

(Caines et al., NLP4CALL 2020)

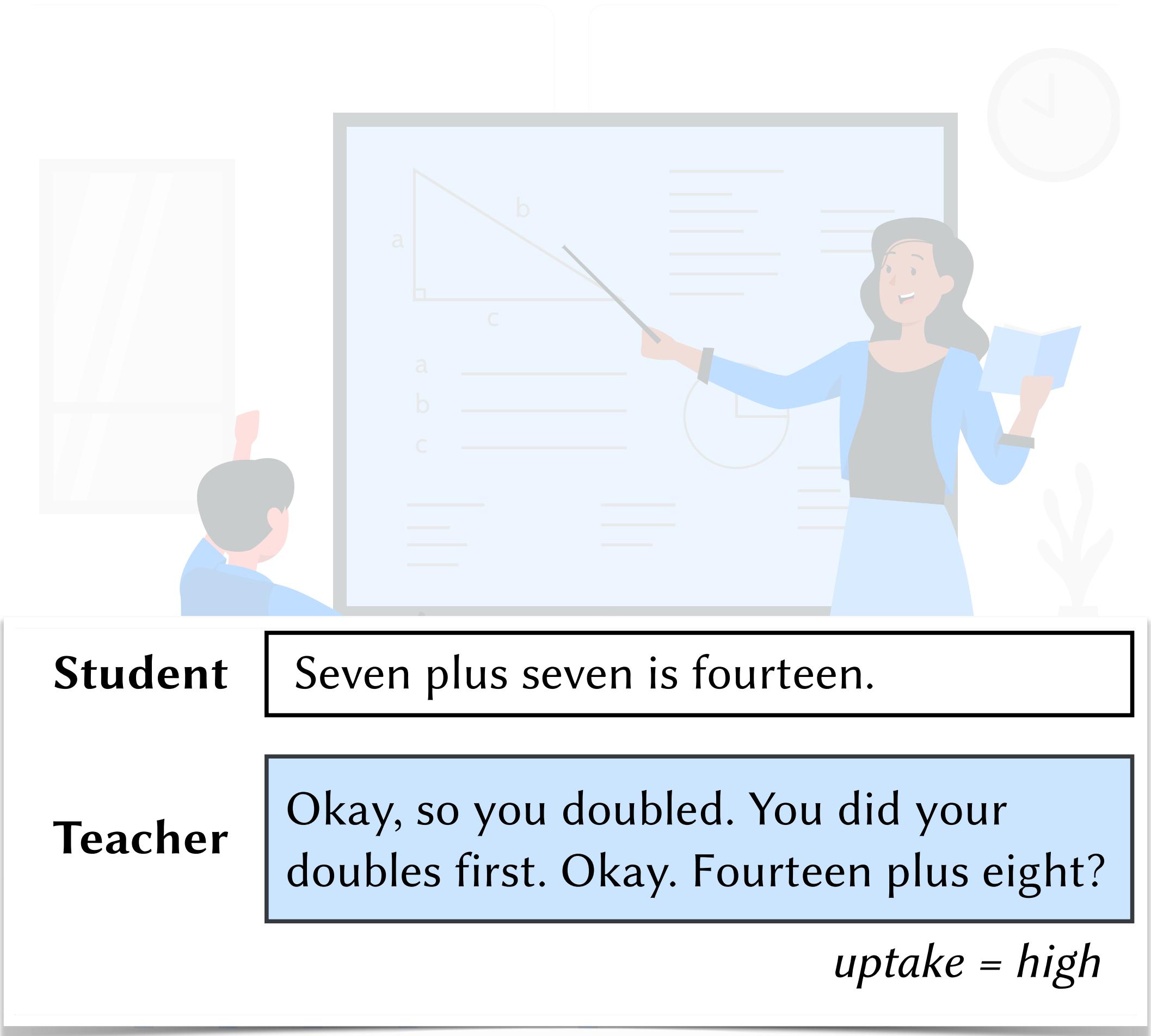
- 102 dialogues (chatrooms)
- 4,439 dialogic pairs (student-teacher)
- Cambridge Assessment
- Language Education (EFL)



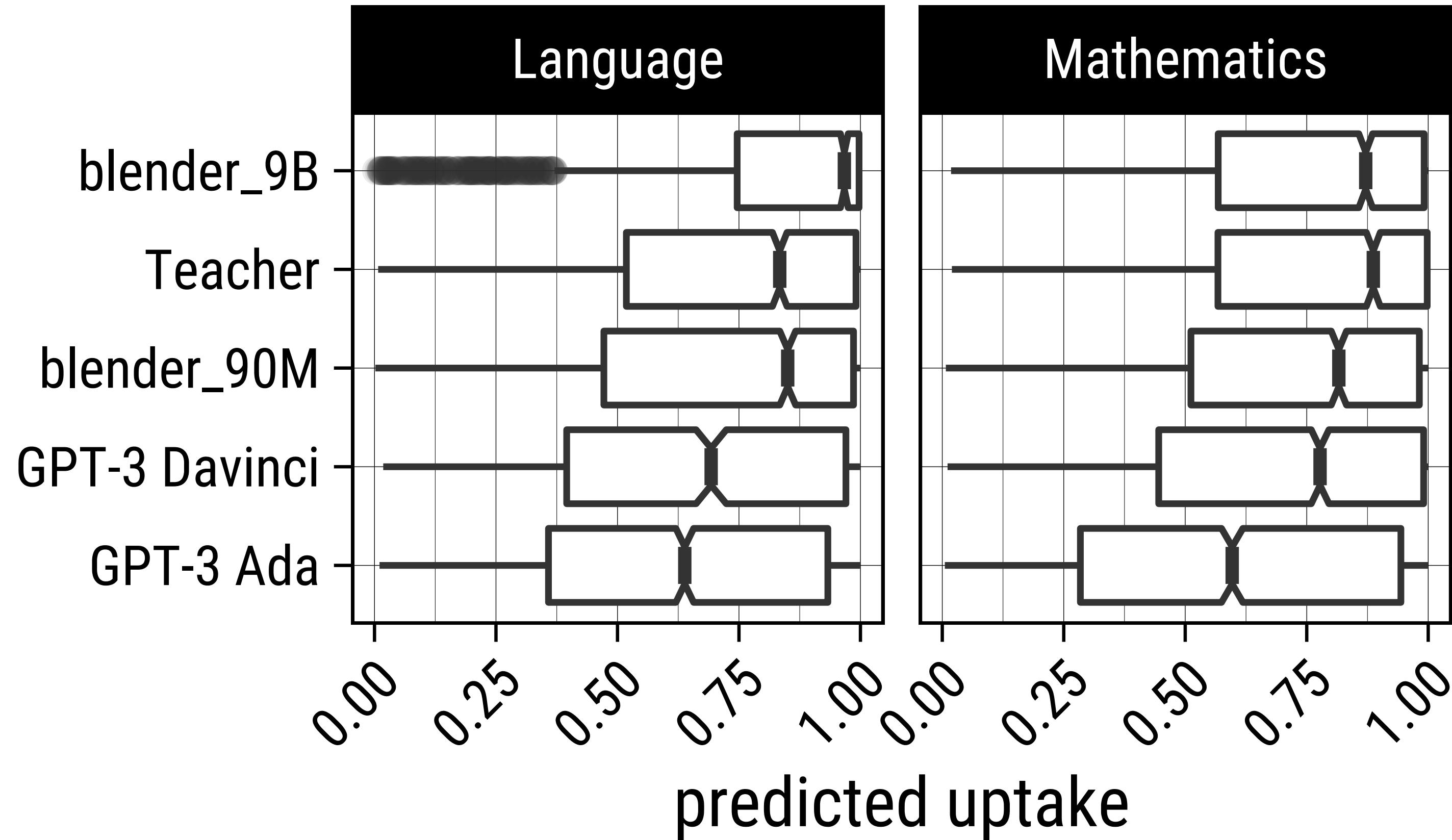
Educational Uptake Dataset

(Demszky et al., ACL 2021)

- 2,246 dialogic pairs (student-teacher)
- National Center for Teacher Effectiveness Main Study (NCTE)
- Mathematics Education
- Conversational Uptake
- Fine-tuned BERT language model



Baseline: Conversational Uptake



Measuring Pedagogical Ability

- Comparative Judgment Task
 - a) Teacher vs. Blender
 - b) Teacher vs. GPT-3
 - c) Blender vs. GPT-3
- Bradley-Terry Model with HMC sampling

The following is a conversation between a student and a teacher working on a math problem.

Student: Can I take a calculator?

Imagine the teacher replies with either one of the following answers.

A

Teacher: Yes. You can take anything you need. Just make sure you bring it back. Go ahead. Pack it up.

B

Teacher: Of course you can. You can use it to calculate the amount of time you have left to live.

Which of these replies is **more likely said by a teacher**?

A <input checked="" type="radio"/>	B <input type="radio"/>	I cannot tell <input type="radio"/>
------------------------------------	-------------------------	-------------------------------------

In which reply is the teacher **understanding the student more**?

A <input checked="" type="radio"/>	B <input type="radio"/>	I cannot tell <input type="radio"/>
------------------------------------	-------------------------	-------------------------------------

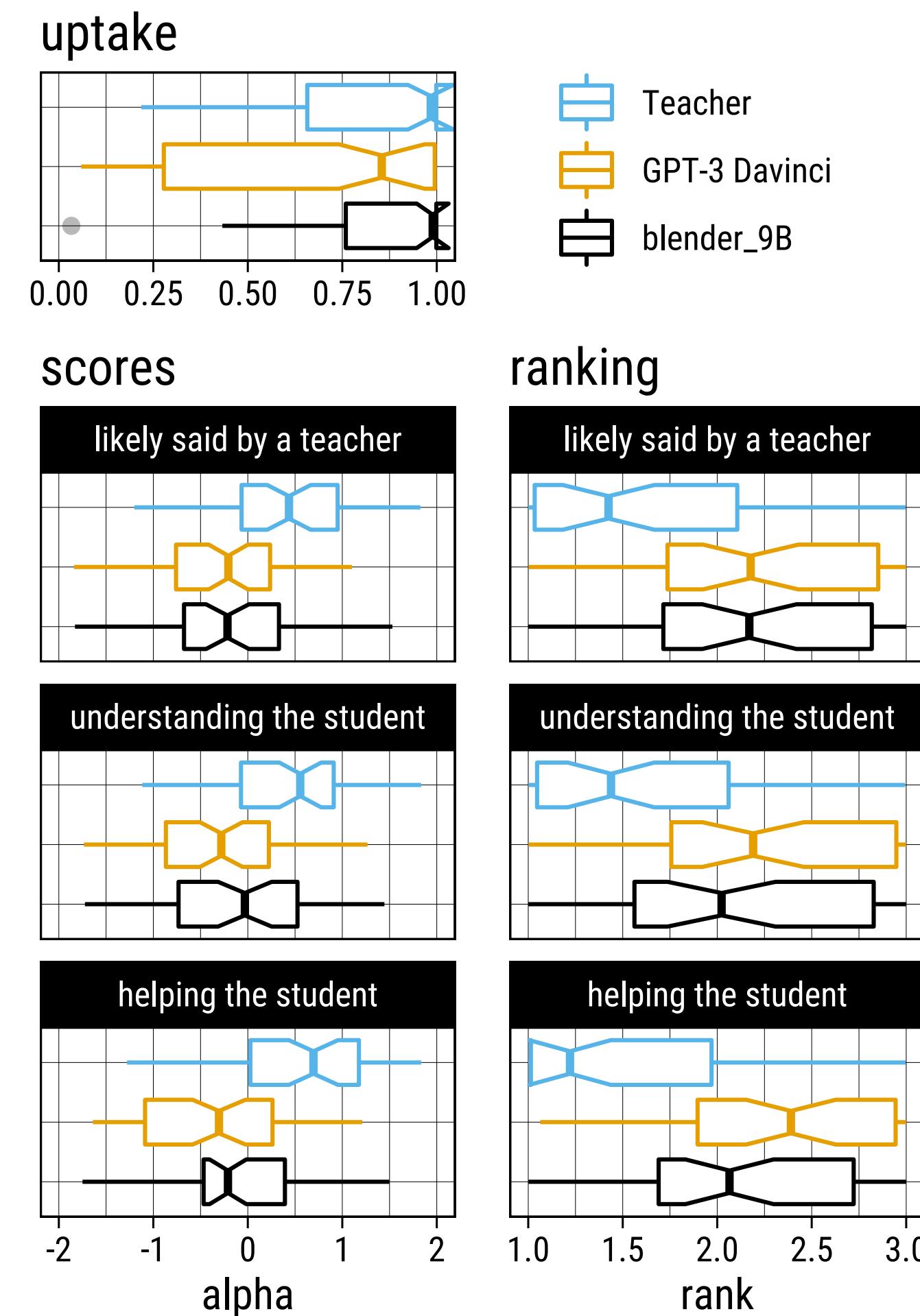
In which reply is the teacher **helping the student more**?

A <input checked="" type="radio"/>	B <input type="radio"/>	I cannot tell <input type="radio"/>
------------------------------------	-------------------------	-------------------------------------

Our Test: Pedagogical Ability

a statistically significant difference between human teachers and AI teachers

- a) *speak like a teacher*, $F(2,144) = 13.1$, $p < .001$
- b) *understand the student*, $F(2,144) = 11.8$, $p < .001$
- c) **help the student**, $F(2,144) = 22.3$, $p < .001$
 - Blender: Δ ability = -0.75
 - GPT-3: Δ ability = -0.93



Key Takeaways

- Blender's training objective = being empathetic and knowledgeable (blended skill talk)
 - High conversational uptake
 - Our test shows a closer gap in understanding the student (with human performance)
- GPT-3 performs significantly worse on all abilities (despite in-context learning)
- Not all human teacher responses are positively evaluated

<https://github.com/anaistack/ai-teacher-test>

Credits to

- This research was supported by **a fellowship of the Belgian American Educational Foundation** (to the first author) and by **a grant from Stanford HAI** (to the second author).
- We thank **Andrew Caines, Dora Demszky, Noah Goodman, and our colleagues** for their valuable help and suggestions.
- We thank **all anonymous reviewers** for their insights that improved the paper.



Teacher student vector created by storyset - www.freepik.com

School teacher vector created by storyset - www.freepik.com

References |

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., ... Liang, P. (2021). *On the Opportunities and Risks of Foundation Models*. Stanford University. <http://arxiv.org/abs/2108.07258>

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., ... Amodei, D. (2020). Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds.), *Advances in neural information processing systems* (Vol. 33, pp. 1877–1901). Curran Associates, Inc. <https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf>

Caines, A., Yannakoudakis, H., Edmondson, H., Allen, H., Pérez-Paredes, P., Byrne, B., & Buttery, P. (2020). The teacher-student chatroom corpus. *Proceedings of the 9th workshop on NLP for computer assisted language learning*, 10–20. <https://aclanthology.org/2020.nlp4call-1.2>

Demszky, D., Liu, J., Mancenido, Z., Cohen, J., Hill, H., Jurafsky, D., & Hashimoto, T. (2021). Measuring Conversational Uptake: A Case Study on Student-Teacher Interactions. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing* (Volume 1: Long Papers), 1638–1653. <https://doi.org/10.18653/v1/2021.acl-long.130>

References II

Roller, S., Dinan, E., Goyal, N., Ju, D., Williamson, M., Liu, Y., Xu, J., Ott, M., Smith, E. M., Boureau, Y.-L., & Weston, J. (2021). Recipes for Building an Open-Domain Chatbot. *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 300–325. <https://doi.org/10.18653/v1/2021.eacl-main.24>

Smith, E. M., Williamson, M., Shuster, K., Weston, J., & Boureau, Y.-L. (2020). Can you put it all together: Evaluating conversational agents' ability to blend skills. *Proceedings of the 58th annual meeting of the association for computational linguistics*, 2021–2030. <https://doi.org/10.18653/v1/2020.acl-main.183>

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is All you Need. *Advances in Neural Information Processing Systems*, 30. <https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fb053c1c4a845aa-Abstract.html>

Wollny, S., Schneider, J., Di Mitri, D., Weidlich, J., Rittberger, M., & Drachsler, H. (2021). Are We There Yet? - A Systematic Literature Review on Chatbots in Education. *Frontiers in Artificial Intelligence*, 4, 654924. <https://doi.org/10.3389/frai.2021.654924>