

Generating Teacher Responses in Educational Dialogues

The AI Teacher Test & BEA 2023 Shared Task

Anaïs Tack





International Alliance to Advance Learning in the Digital Era (IAALDE)

THE VISION FOR THE ALLIANCE INCLUDES BOTH OUTWARD-FACING AND INWARD-FACING GOALS.

Outward Goal:

To increase the visibility and impact of the research produced by members of the individual societies by providing a mechanism for the societies to collaborate and generate a cohesive picture of important findings related to learning in the digital era that could influence research, educational practices, and policy. A highly successful exemplar for achieving such a goal (at least within the USA) is the Computing Research Association (CRA) whose mission is "to enhance

Inward Goal:

To increase cross-society awareness both at a scientific level (around differing methods, theories, technologies, findings) and at an administrative level (especially around the organization of conference logistics).



International Alliance to Advance Learning and Development (IAALDE)

THE VISION FOR THE ALLIANCE INCLUDES BOTH OUTWARD AND INWARD GOALS.

Outward Goal:

To increase the visibility and impact of the research produced by members of the individual societies by providing a mechanism for the societies to collaborate and generate a cohesive picture of important findings related to learning in the digital era that could influence research, educational practices, and policy. A highly successful exemplar for achieving such a goal (at least within the USA) is the Computing Research Association (CRA) whose mission is "to enhance

THE SOCIETIES

INVOLVED IN THE EFFORT

ISLS

International Society of the Learning Sciences



AIED

International Artificial Intelligence in Education Society



ST&D

Society for Text and Discourse



IEDMS

International Educational Data Mining Society



EATEL

European Association of Technology-Enhanced Learning



SoLAR

Society for Learning Analytics Research



L@S

ACM's Learning at Scale



ATIEF

Association des Technologies de l'Information pour l'Education et la Formation



SIG EDU

ACL Special Interest Group on Building Educational Applications



APSCE

Asia Pacific Society for Computers in Education



ASLERD

Association of Smart Learning Ecosystems and Regional Development



GCSCE

Global Chinese Society for Computers in Education



logistics).



International Alliance to Advance Learning and Development (IAALDE)

THE VISION FOR THE ALLIANCE INCLUDES BOTH OUTWARD AND INWARD GOALS.

Outward Goal:

To increase the visibility and impact of the research produced by members of the individual societies by providing a mechanism for the societies to collaborate and generate a cohesive picture of important findings related to learning in the digital era that could influence research, educational practices, and policy. A highly successful exemplar for achieving such a goal (at least within the USA) is the Computing Research Association (CRA) whose mission is "to enhance

THE SOCIETIES

INVOLVED IN THE EFFORT

ISLS

International Society of the Learning Sciences



AIED

International Artificial Intelligence in Education Society



ST&D

Society for Text and Discourse



IEDMS

International Educational Data Mining Society



EATEL

European Association of Technology-Enhanced Learning



SoLAR

Society for Learning Analytics Research



L@S

ACM's Learning at Scale



ATIEF

Association des Technologies de l'Information pour l'Education et la Formation



SIG EDU

ACL Special Interest Group on Building Educational Applications



APSCE

Asia Pacific Society for Computers in Education



ASLERD

Association of Smart Learning Ecosystems and Regional Development



GCSCE

Global Chinese Society for Computers in Education





International Alliance to Advance Learning and Development in Education (IAALDE)

THE VISION FOR THE ALLIANCE INCLUDES BOTH OUTWARD AND INWARD GOALS.

Outward Goal:

To increase the visibility and impact of the research produced by members of the individual societies by providing a mechanism for the societies to collaborate and generate a cohesive picture of important findings related to learning in the digital era that could influence research, educational practices, and policy. A highly successful exemplar for achieving such a goal (at least within the USA) is the Computing Research Association (CRA) whose mission is "to enhance

THE SOCIETIES

INVOLVED IN THE EFFORT

ISLS

International Society of the Learning Sciences



AIED

International Artificial Intelligence in Education Society



ST&D

Society for Text and Discourse



EATEL

European Association of Technology-Enhanced Learning



SoLAR

Society for Learning Analytics Research



L@S

ACM's Learning at Scale



IEDMS

International Educational Data Mining Society



ATIEF

Association des Technologies de l'Information pour l'Education et la Formation



SIG EDU

ACL Special Interest Group on Building Educational Applications



APSCE

Asia Pacific Society for Computers in Education



ASLERD

Association of Smart Learning Ecosystems and Regional Development



GCSCE

Global Chinese Society for Computers in Education



EDM 2022

Intelligence in Education Week at Durham



The image shows a panoramic view of the Durham skyline. In the background, the historic Durham Cathedral and Palace are perched atop a green hill under a bright blue sky with scattered white clouds. In the foreground, the calm River Wear flows through the town, with a stone arch bridge spanning it. To the left, there are several traditional brick buildings. A pink navigation bar at the top of the image contains links: HOME (highlighted in pink), PROGRAMME ▾, VENUE ▾, ORGANISATION, CALLS ▾, SUBMISSION ▾, SPONSORS ▾, and REGISTRATION.

EDM 2022: The 15th International Conference on Educational Data Mining, 24 -27 July 2022

Durham University, Durham, UK

[CALL FOR PAPERS >](#)



The AI Teacher Test

EDM 2022 Best Short Paper

The AI Teacher Test: Measuring the Pedagogical Ability of Blender and GPT-3 in Educational Dialogues

Anaïs Tack
Stanford University
attack@cs.stanford.edu

Chris Piech
Stanford University
piech@cs.stanford.edu

ABSTRACT

How can we test whether state-of-the-art generative models, such as Blender and GPT-3, are good AI teachers, capable of replying to a student in an educational dialogue? Designing an AI teacher test is challenging: although evaluation methods are much-needed, there is no off-the-shelf solution to measuring pedagogical ability. This paper reports on a first attempt at an AI teacher test. We built a solution around the insight that you can run conversational agents in parallel to human teachers in real-world dialogues, simulate how different agents would respond to a student, and compare these counterpart responses in terms of three abilities: speak like a teacher, understand a student, help a student. Our method builds on the reliability of comparative judgments in education and uses a probabilistic model and Bayesian sampling to infer estimates of pedagogical ability. We find that, even though conversational agents (Blender in particular) perform well on conversational uptake, they are quantifiably worse than real teachers on several pedagogical dimensions, especially with regard to helpfulness (Blender: Δ ability = -0.75; GPT-3: Δ ability = -0.93).

Keywords

student-teacher dialogue, conversational agents, chatbots, Blender, GPT-3, evaluation methods, pairwise comparisons, Bayesian Bradley-Terry model

1. INTRODUCTION

Conversational agents (or chatbots) offer promising opportunities for education. They can fulfill various roles (such as intelligent tutors and service-oriented assistants) and pursue different objectives (e.g., improving student skills, boosting student motivation, and increasing instructional efficiency) [20]. Among all of these different vocations of an educational chatbot, the most prevalent one is the AI teacher helping a student with skill improvement and providing more opportunities to practice. Some recent meta-analyses have even reported a significant effect of chatbots on skill im-

provement, for example in language learning [1]. What is more, current advances in AI and natural language processing have led to the development of conversational agents that are founded on more powerful generative language models. Blender [17], for instance, is a state-of-the-art open-domain chatbot trained to blend skills such as being empathetic and knowledgeable [18], which are undeniably important characteristics of a good AI teacher. Furthermore, the current state-of-the-art in natural language generation is GPT-3 [4], a 175B-parameter model that is able to multitask different language generation skills (such as conversation). The astonishing power of GPT-3 is that it can perform these skills with few-shot in-context learning, merely from seeing a short prompt describing the task at hand (e.g., *The following is a conversation with an AI assistant.*). Emergent models such as GPT-3 have been described as *foundation models* since they serve as the “common basis from which many task-specific models are built via adaptation” [2 p.7].

Despite these promising opportunities, the use of powerful generative models as a foundation for downstream tasks also presents several crucial challenges. In the educational domain in particular, it is important to ascertain whether that foundation is solid or flimsy. Bommasani et al. [2] pp.67-72 stressed that, if we want to put these models into practice as AI teachers, it is imperative to determine whether they can (a) speak to students like a teacher, (b) understand students, and (c) help students improve their understanding. Consequently, there is a critical need to establish good evaluation methods of AI teachers. This is a hard problem because there is no off-the-shelf and universal solution to measuring teaching ability and effectiveness.

Therefore, we took on the challenge of designing an AI teacher test and conducted a pilot study. We ran Blender and GPT-3 in parallel to human teachers in language and mathematics educational dialogues, observed how they responded to a student, and compared these counterpart responses in terms of pedagogical ability. The major contributions of this work are as follows:

1. We pose the AI Teacher Test Challenge.
2. We implement a human-in-the-loop pairwise comparison test as a first attempt at an AI Teacher Test.
3. Our results show quantitatively how far conversational agents, particularly Blender and GPT-3, are behind human teachers in terms of pedagogical ability, despite them performing well on conversational uptake.

A. Tack and C. Piech. The AI teacher test: Measuring the pedagogical ability of blender and GPT-3 in educational dialogues. In A. Mitrovic and N. Bosch, editors, *Proceedings of the 15th International Conference on Educational Data Mining*, pages 522–529, Durham, United Kingdom, July 2022. International Educational Data Mining Society.

© 2022 Copyright is held by the author(s). This work is distributed under the Creative Commons Attribution NonCommercial NoDerivatives 4.0 International (CC BY-NC-ND 4.0) license.
<https://doi.org/10.5281/zanendo.6853187>

BEA 2023 Shared Task

The screenshot shows the CodaLab competition page for the BEA 2023 Shared Task. The page has a blue header with the CodaLab logo and navigation links for Search Competitions, My Competitions, Help, Sign Up, and Sign In. Below the header, the title "Competition" is displayed. On the left, there is a red "SIG EDU" logo. The main content area is titled "BEA 2023 Shared Task on Generating AI Teacher Responses in Educational Dialogues". It states that the competition is organized by anaistack and was last updated on May 30, 2023, at 9:28 a.m. UTC. The "First phase" is labeled "End" and the "Development" phase is labeled "Competition Ends". The development period is from March 24, 2023, at 11:59 a.m. UTC to May 6, 2023, at 11:59 a.m. UTC. Below this, there are tabs for Learn the Details, Phases, Participate, Results, and Forums. The "Overview" tab is selected. Other tabs include Evaluation and Terms and Conditions. The main text on the right describes the task as generating AI teacher responses in educational dialogues, organized during the BEA workshop at ACL 2023 in Toronto, Canada on July 13, 2023. The website is https://sig-edu.org/sharedtask/2023. A note in red says "Participation on CodaLab will only be accepted upon registration: https://forms.gle/JhDtAMJwadNmNgKv7". The text then discusses conversational agents in education, mentioning their potential roles and challenges, including the AI teacher test challenge. At the bottom, there are links to GitHub for contact & bug reports, About, Privacy and Terms, and version v1.6.

Search Competitions My Competitions Help Sign Up Sign In

Competition

SIG
EDU

BEA 2023 Shared Task on Generating AI Teacher Responses in Educational Dialogues

Organized by anaistack - Current server time: May 30, 2023, 9:28 a.m. UTC

First phase End

Development Competition Ends

March 24, 2023, 11:59 a.m. UTC May 6, 2023, 11:59 a.m. UTC

Learn the Details Phases Participate Results Forums

Overview

Evaluation

Terms and Conditions

BEA 2023 Shared Task

Generating AI Teacher Responses in Educational Dialogues

Shared task organized during the BEA workshop at ACL 2023.
Toronto, Canada
July 13, 2023

Website: <https://sig-edu.org/sharedtask/2023>

Participation on CodaLab will only be accepted upon registration: <https://forms.gle/JhDtAMJwadNmNgKv7>

Conversational agents offer promising opportunities for education. They can fulfill various roles (intelligent tutors and service-oriented assistants) and pursue different objectives (e.g., improving student skills and increasing instructional efficiency) (Wollny et al. 2021). Among all of these different vocations of an educational chatbot, the most prevalent one is the AI teacher helping a student with skill improvement and providing more opportunities to practice. Some recent meta-analyses have even reported a significant effect of chatbots on skill improvement, for example in language learning (Bibauw et al. 2022). What is more, current advances in AI and natural language processing have led to the development of conversational agents that are founded on more powerful generative language models.

Despite these promising opportunities, the use of powerful generative models as a foundation for downstream tasks also presents several crucial challenges. In the educational domain in particular, it is important to ascertain whether that foundation is solid or flimsy. Bommasani et al. (2021: pp. 67–72) stressed that, if we want to put these models into practice as AI teachers, it is imperative to determine whether they can (a) speak to students like a teacher, (b) understand students, and (c) help students improve their understanding. Therefore, Tack and Piech (2022) formulated the AI teacher test challenge: How can we test whether state-of-the-art generative models are good AI teachers, capable of replying to a student in an educational dialogue?

Join us on GitHub for contact & bug reports About Privacy and Terms v1.6

The AI Teacher Test

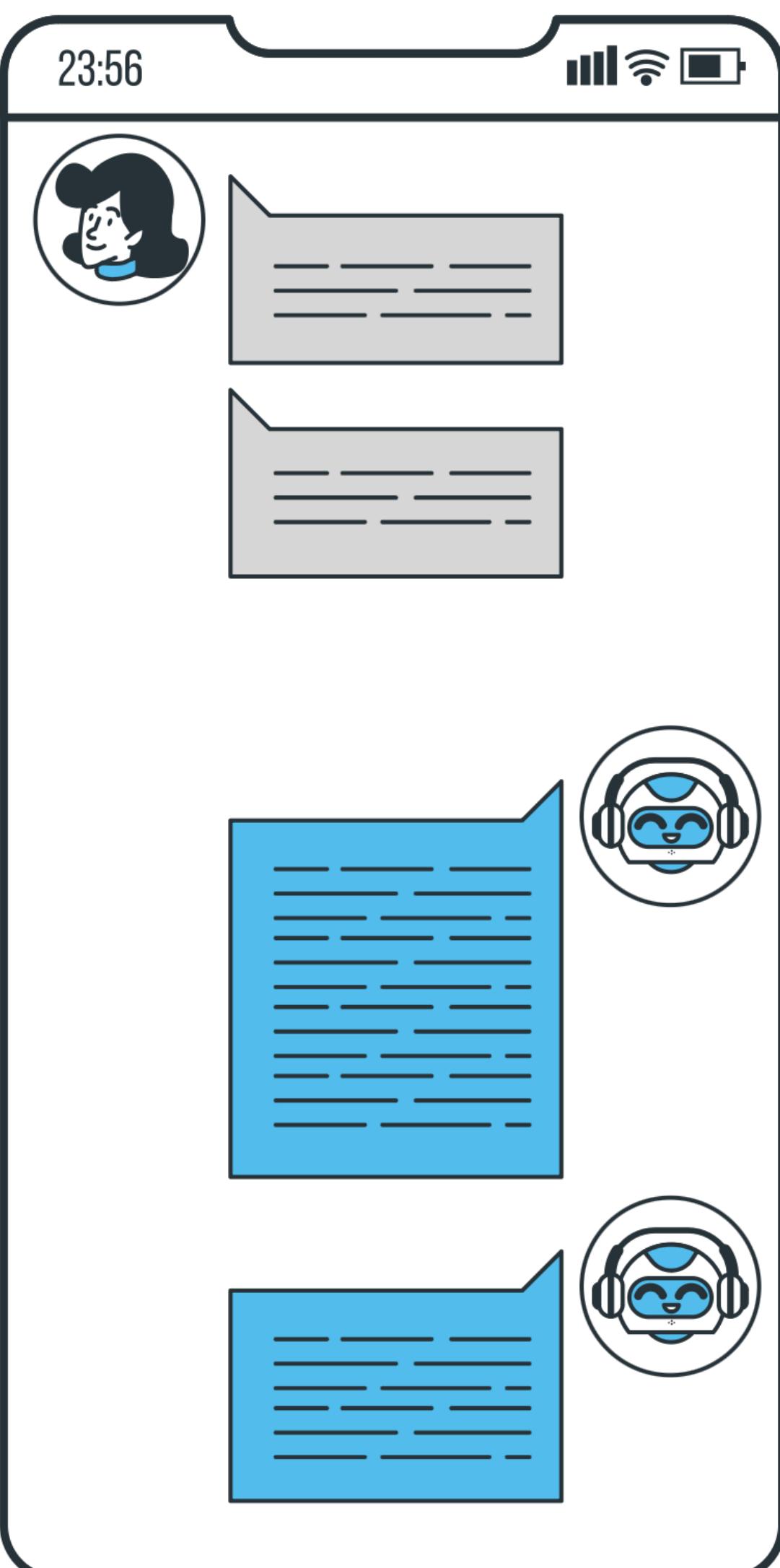
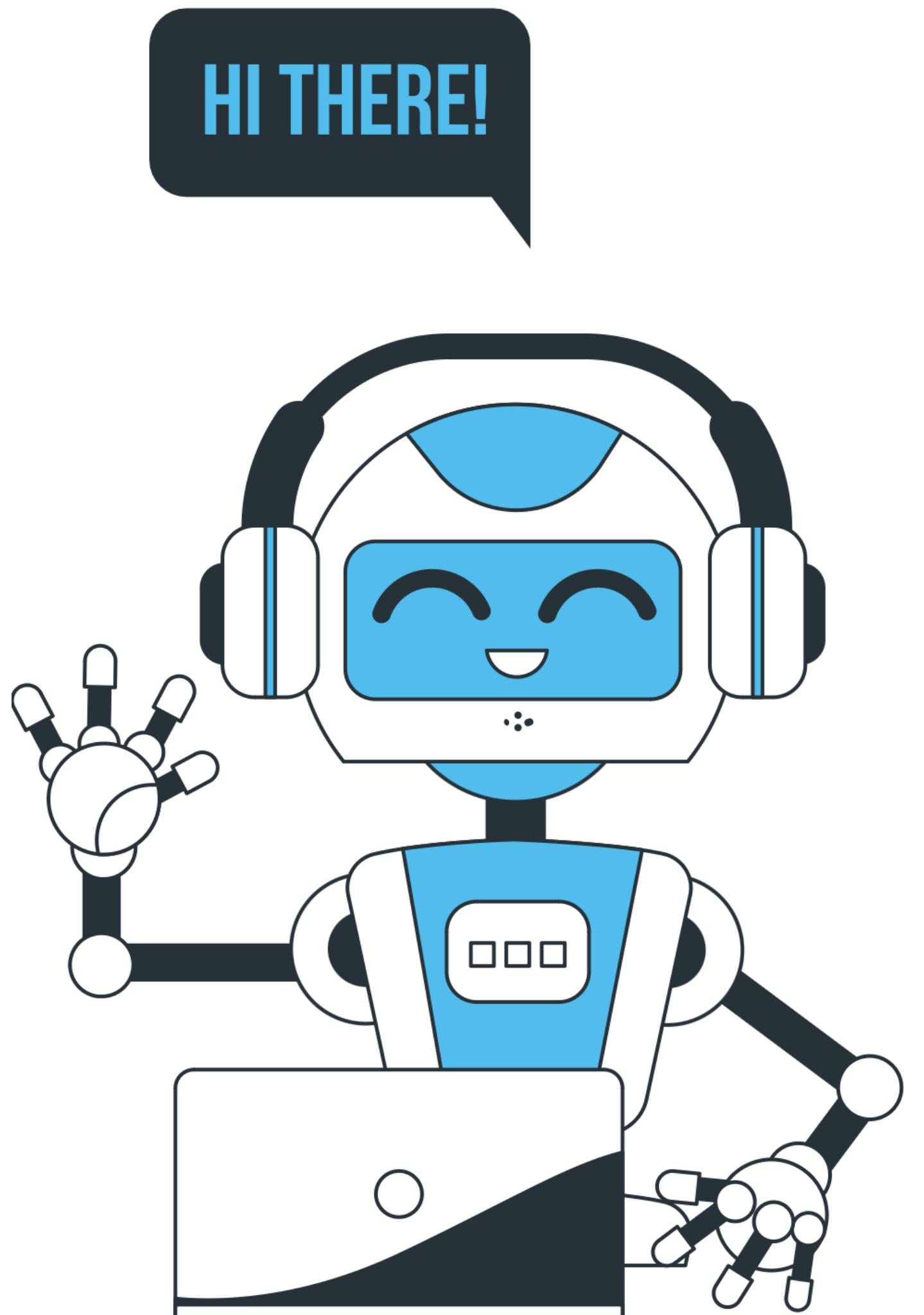
Measuring the Pedagogical Ability of Blender and GPT-3 in
Educational Dialogues

Anaïs Tack • Chris Piech



Conversational AI for Education

- intelligent tutors, service-oriented assistants, ...
- improving student skills, boosting student motivation, ...
- most prevalent = the **AI teacher**
 - helping a student with skill improvement
 - providing more opportunities to practice



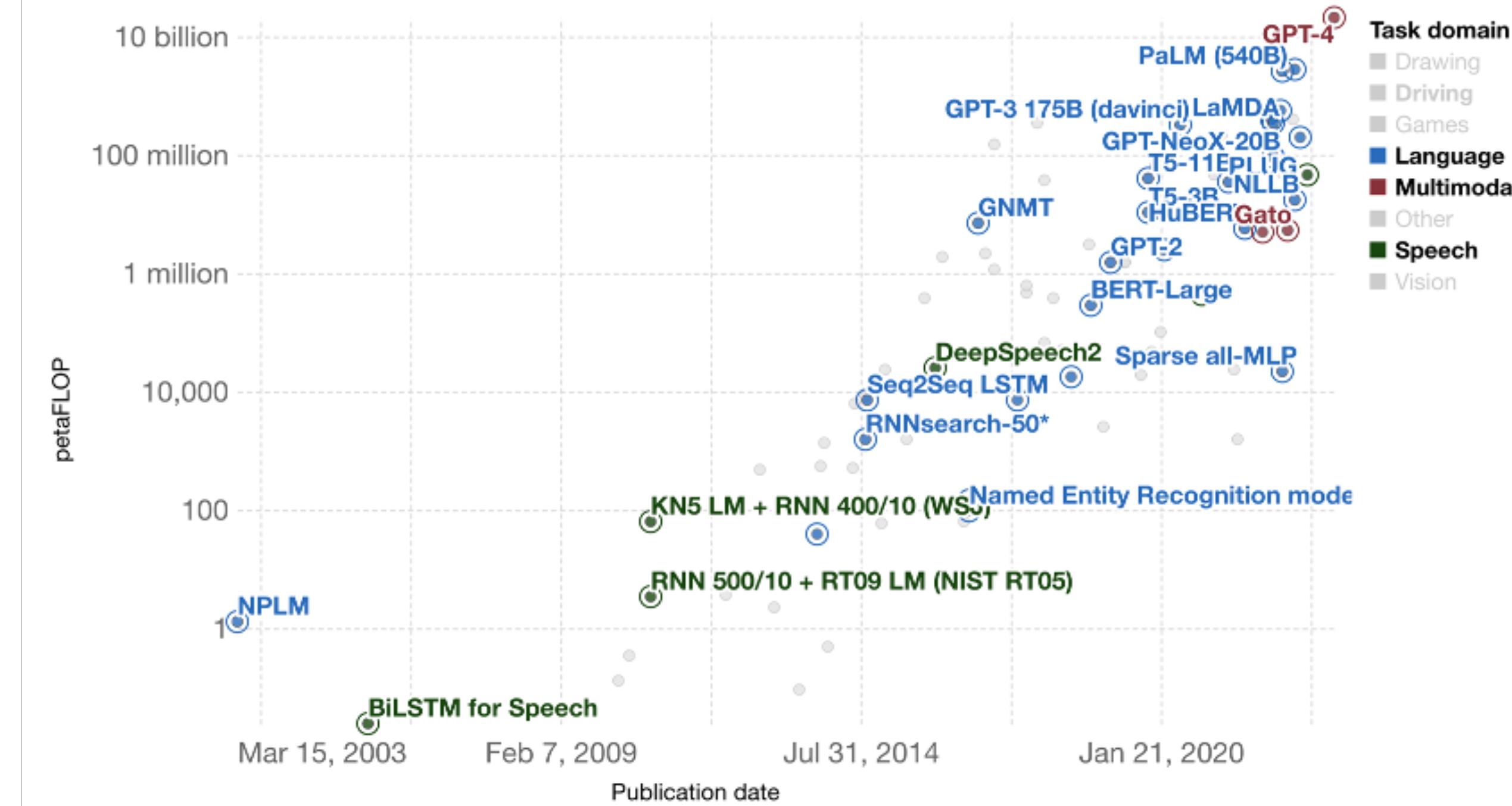
LLMs as Foundation Models

- **foundation** upon which task-specific applications are built via **transfer learning**
- « **scale** is what makes foundation models powerful »
 - big data
 - computer hardware (GPU)
 - transformer architecture

Computation used to train notable artificial intelligence systems

Computation is measured in total petaFLOP, which is 10^{15} floating-point operations¹.

Our World
in Data



Source: Sevilla et al. (2023)

Note: Computation is estimated based on published results in the AI literature and comes with some uncertainty. The authors expect the estimate to be correct within a factor of 2.

OurWorldInData.org/artificial-intelligence • CC BY

1. **Floating-point operation:** A floating-point operation (FLOP) is a type of computer operation. One FLOP is equivalent to one addition, subtraction, multiplication, or division of two decimal numbers.

Foundation Models in Education

On the Opportunities and Risks of Foundation Models

3.3 Education

Authors: Ali Malik, Dorottya Demszky, Pang Wei Koh, Moussa Doumbouya, Drew A. Hudson, Allen Nie, Hamed Nilforoshan, Alex Tamkin, Emma Brunskill, Noah Goodman, Chris Piech

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

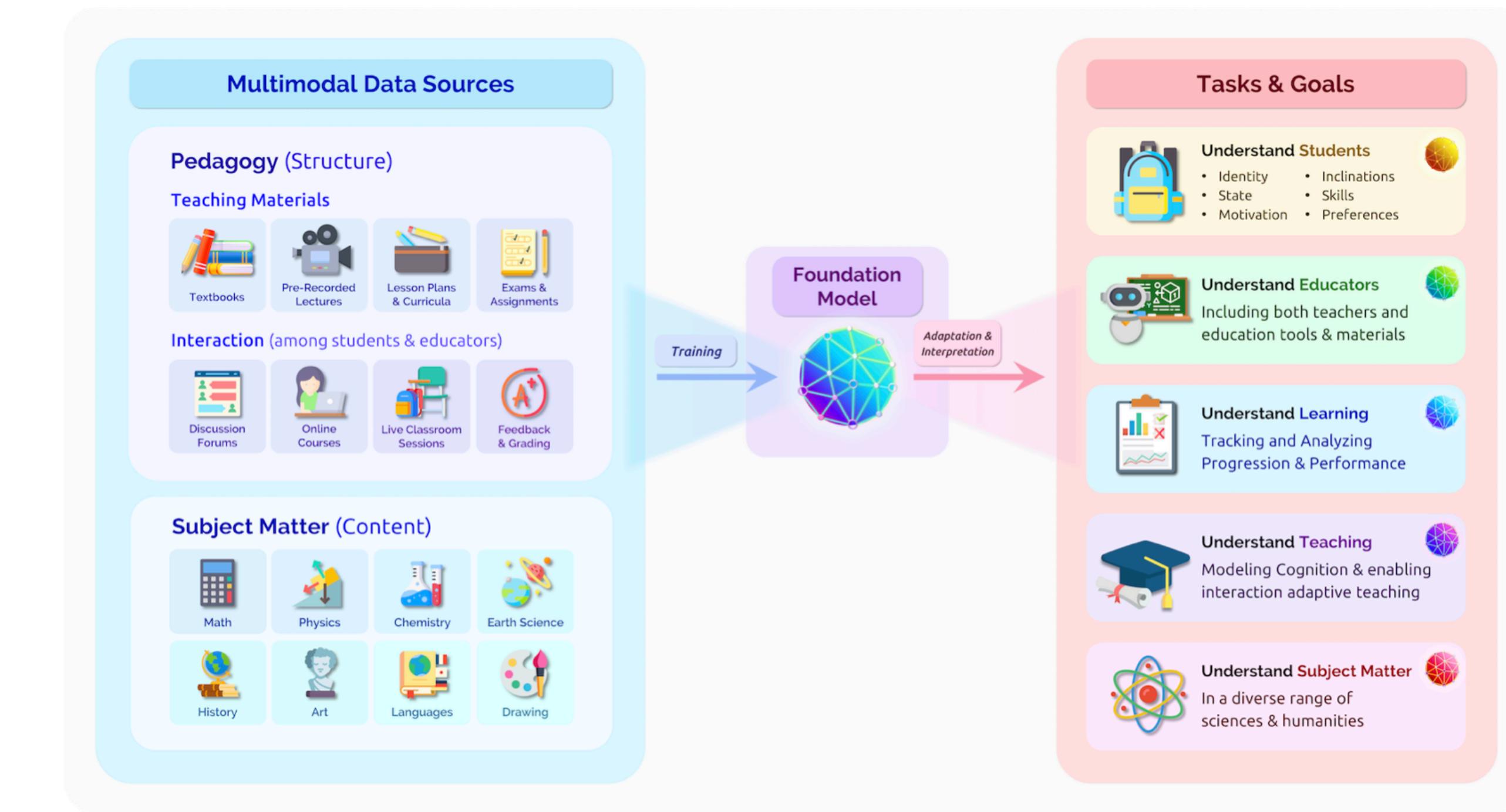


Fig. 15. Foundation models in education could be trained on multiple data sources to learn the capabilities necessary for education: an understanding of various subject matter and different pedagogical techniques. These foundation models can be applied in a general-purpose way across a range of tasks and goals such as understanding students, assisting teachers, and generating educational content.

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

The AI Teacher Test 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 **in context**
2. score with respect to several **pedagogical abilities**
3. compare to other possible replies (better or worse) = **ranking**

| | |
|---------|--|
| 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 |

likely said by a teacher
understanding the student
helping the student

| Responses | | | |
|-----------|---|--------|--------|
| Teacher | yes, good. And when the battery is full, you need to ____ (disconnect it) | 1.769 | 1.632 |
| GPT-3 | yes! | 0.116 | -0.071 |
| Blender | do you have any pets? I have a dog and a cat. I love them | -1.829 | -1.554 |

Real-World Student-Teacher Dialogues

Teacher-Student Chatroom Corpus

(Caines et al., NLP4CALL 2020)

- Cambridge Assessment
- language education (ESL)
- one-on-one tutoring setting
- 102 dialogues (chatrooms)
- annotated utterances
- speaker and chat metadata

Chatroom

yes, good! And to charge it up, you
need to __ it __

...

connect to the source of electricity

i understand

plug it __?

in

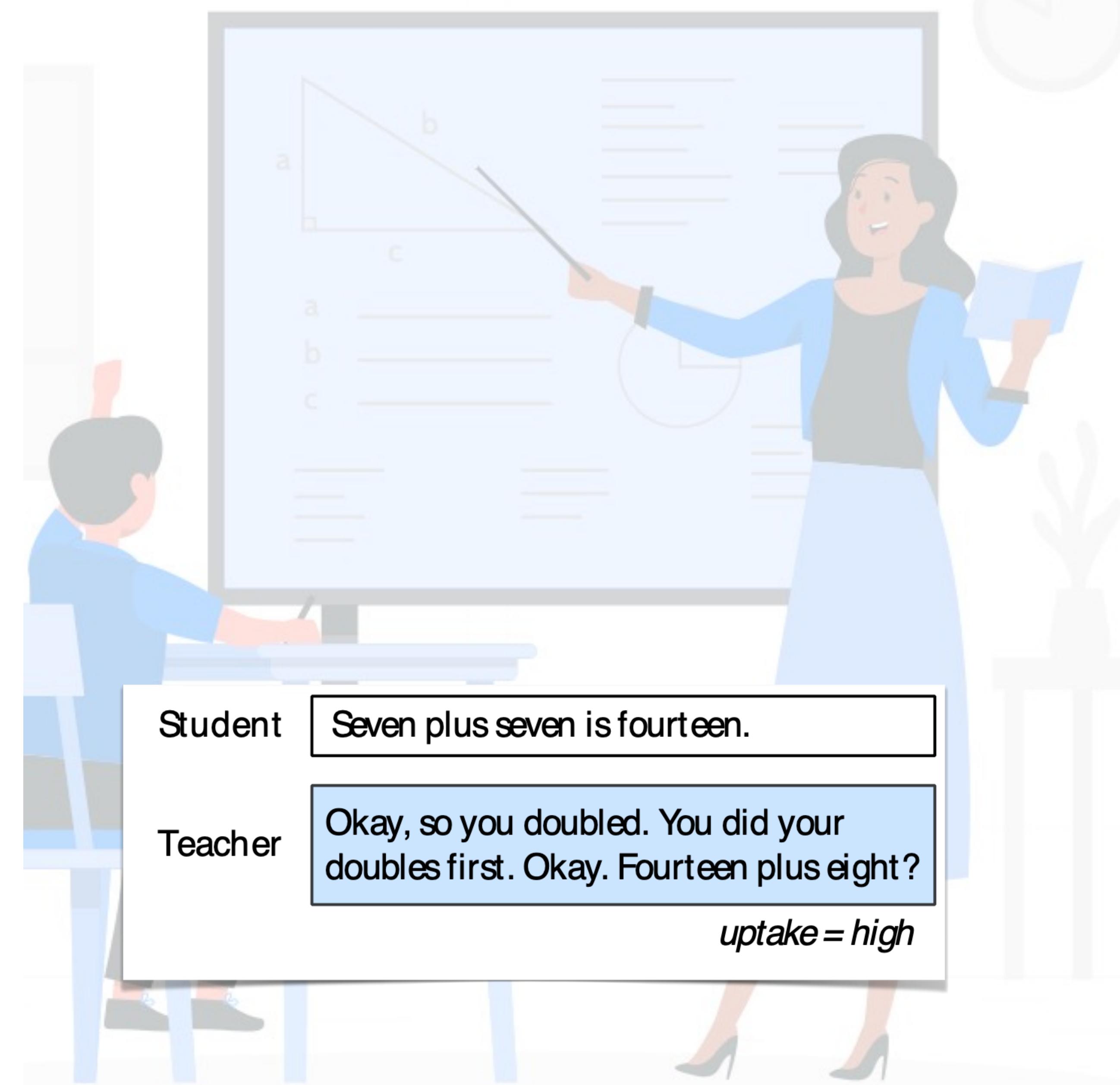
Educational Uptake Dataset

(Demszky et al., ACL 2021)

- National Center for Teacher Effectiveness Main Study (NCTE)
- mathematics education
- classroom setting
- 2,246 dialogic pairs (student-teacher)

Conversational Uptake Measure

- next utterance classification
- fine-tuned BERT language model



Generative Language Models

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

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!

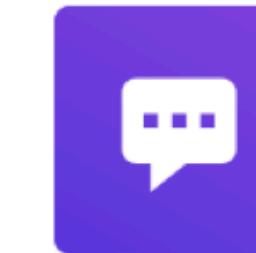
I know, right? Time flies by so fast. I wish I could go back to those simpler times.

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)



Chat

Conversation

Generation

Open ended conversation with an AI assistant.

- 175B-parameter model
- powerful generative model
- few-shot in-context learning from prompt
- various tasks, also **chatbot**

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.

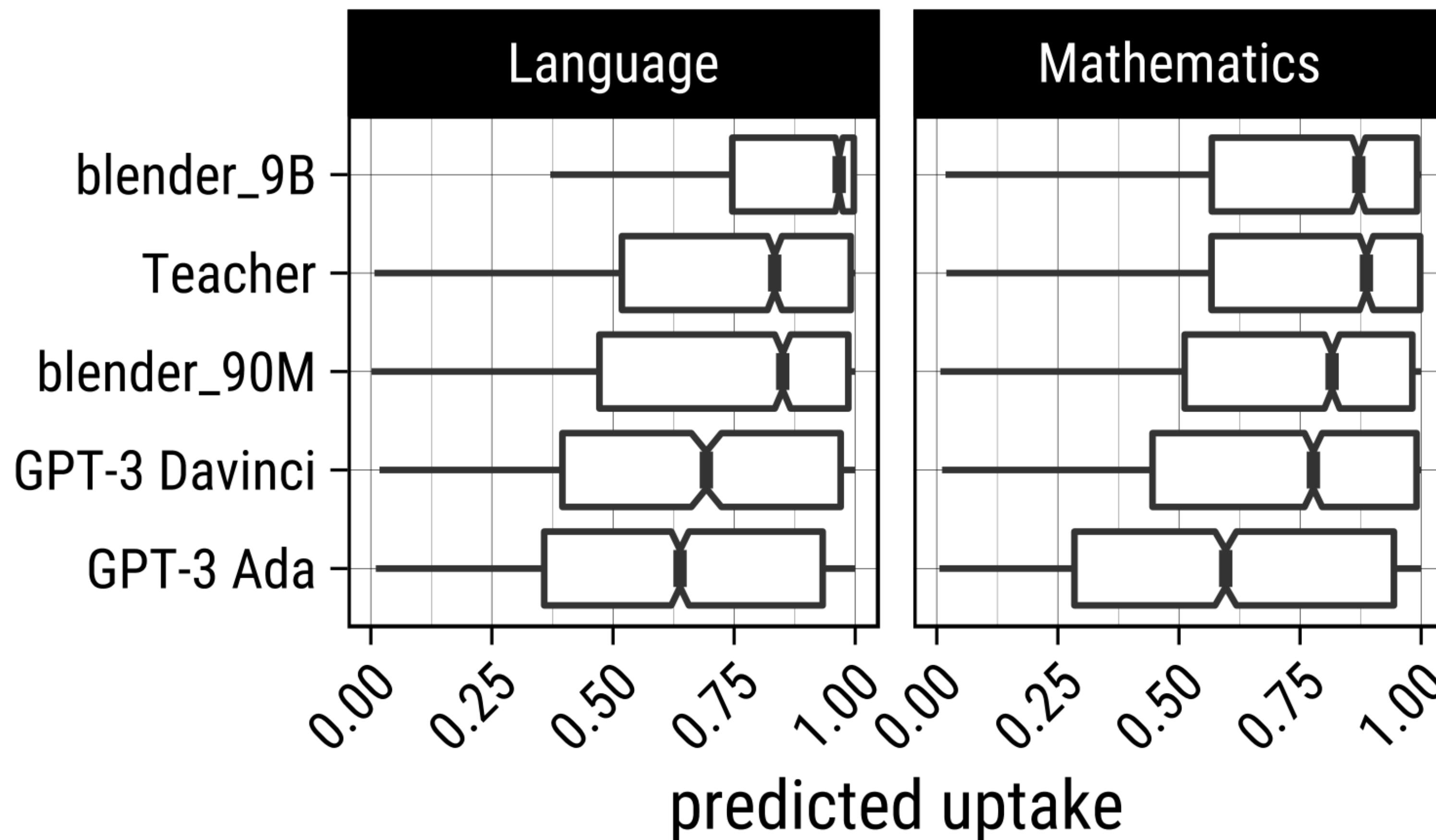
AI:

Sample response

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

Baseline: Conversational Uptake

(Demszky et al., ACL 2021)



Comparative Judgments

- survey: 50 items; 120 raters (Prolific)
- select 1 out of 3 **comparisons**
 - a) Teacher ↔ Blender
 - b) Teacher ↔ GPT-3
 - c) Blender ↔ GPT-3
- three **abilities**:
 1. likely said by a teacher
 2. understanding the student
 3. helping the student

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 B I cannot tell

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

 A B I cannot tell

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

 A B I cannot tell

Bradley-Terry Model

(Bradley and Terry, 1952)

$$p_{ijkl} = \sigma(\alpha_{0kl} + \alpha_{ikl} - \alpha_{jkl})$$

For each:

- **item l** (prior dialogue context)
- **ability k** (speak; understand; help)

Estimate: **probability** that **response i** is **preferred** over **response j**

- latent **ability** parameters : α_{ikl} α_{jkl}
- intercept, home advantage : α_{0kl}

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 |

likely said by a teacher
understanding the student
helping the student

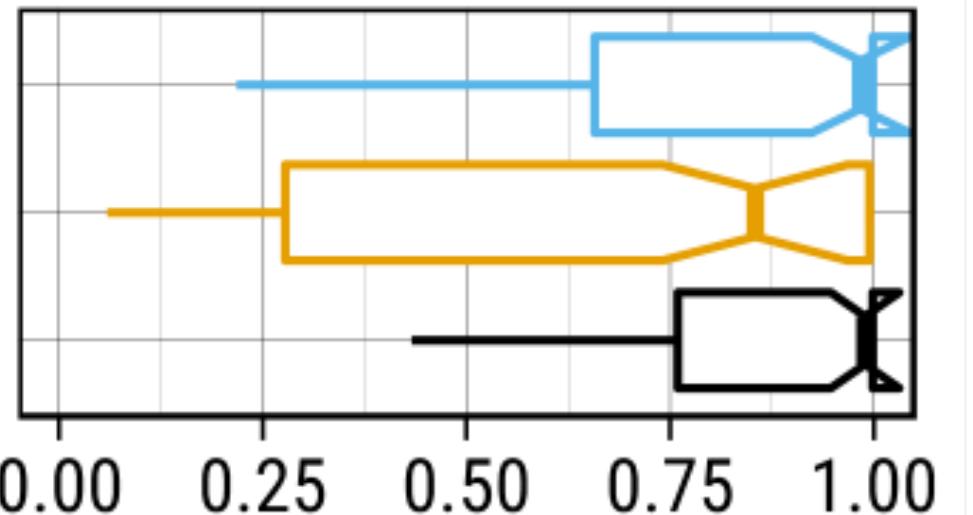
| Responses | | | | |
|-----------|---|--------|--------|--------|
| | | 1.769 | 1.632 | 1.632 |
| Teacher | yes, good. And when the battery is full, you need to ____ (disconnect it) | | | |
| GPT-3 | yes! | 0.116 | -0.071 | -0.071 |
| Blender | do you have any pets? I have a dog and a cat. I love them | -1.829 | -1.554 | -1.554 |

Results

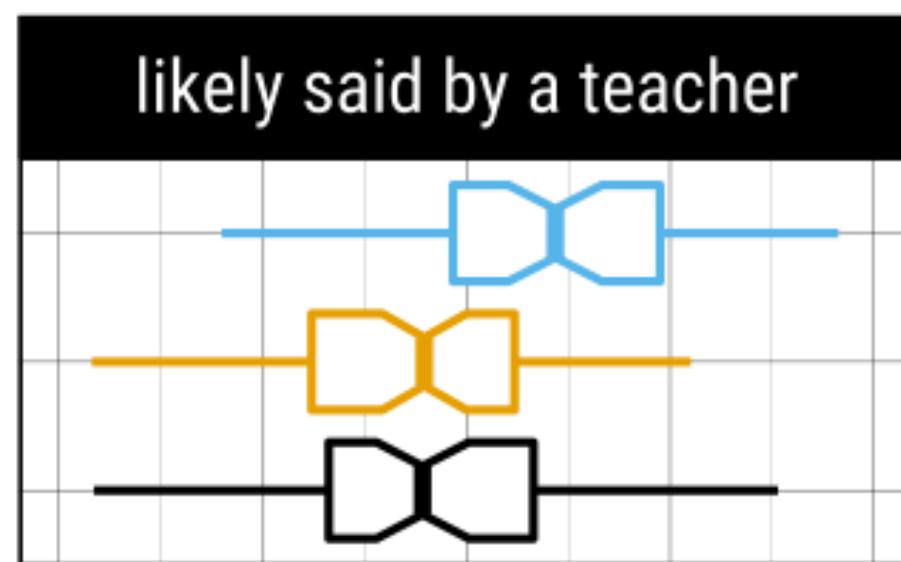
a statistically significant difference between human teachers and AI teachers

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

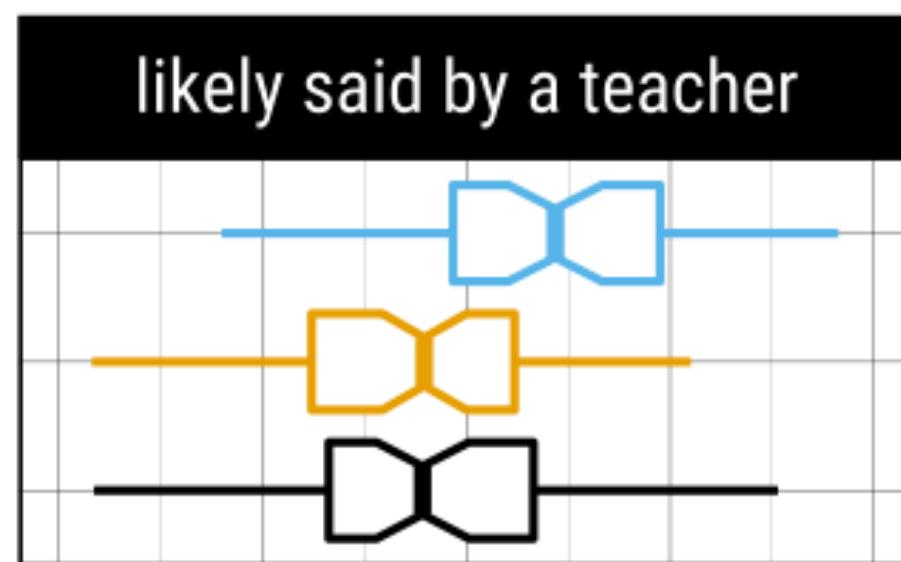
uptake



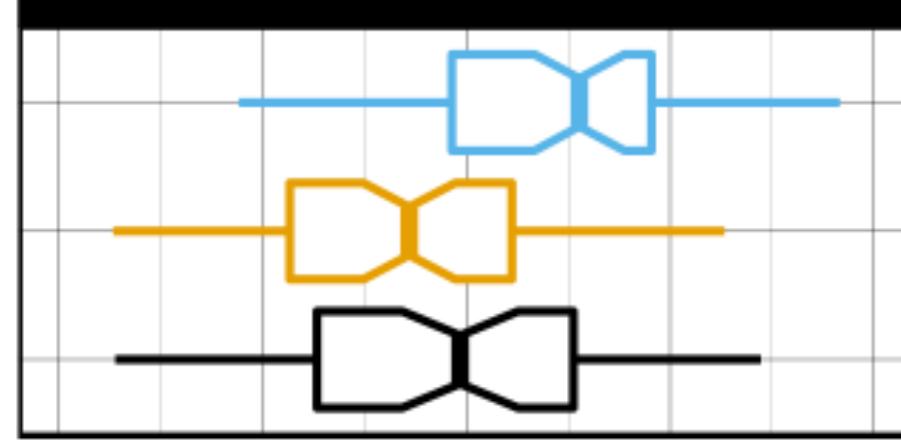
scores



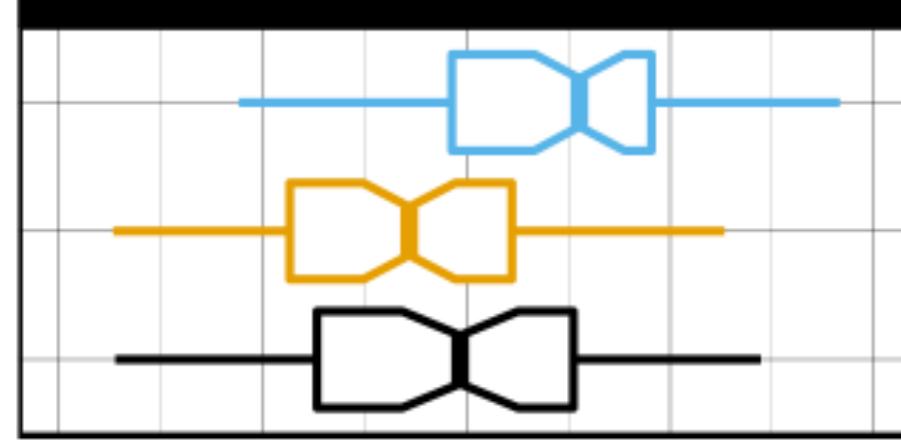
ranking



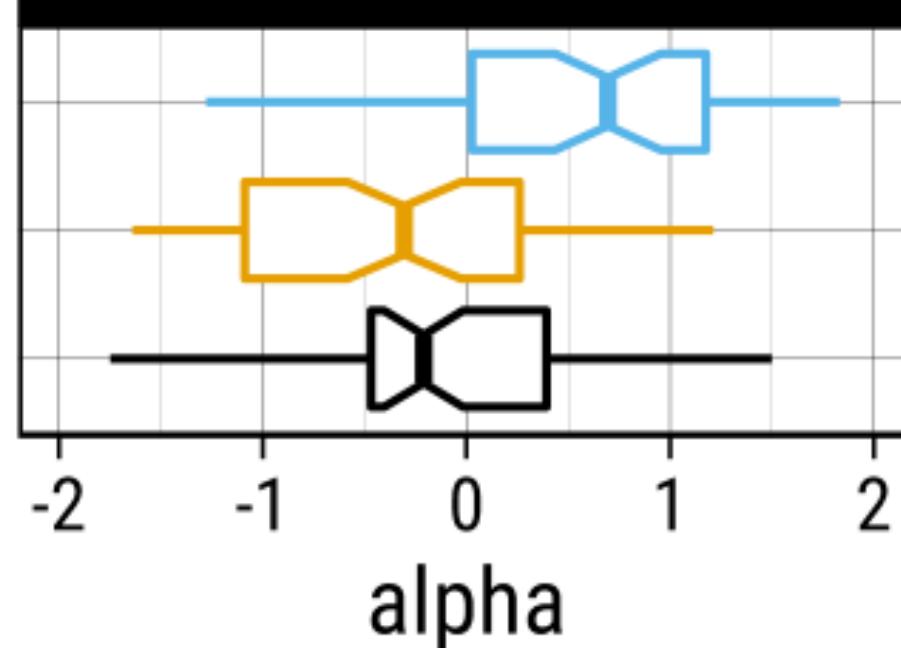
understanding the student



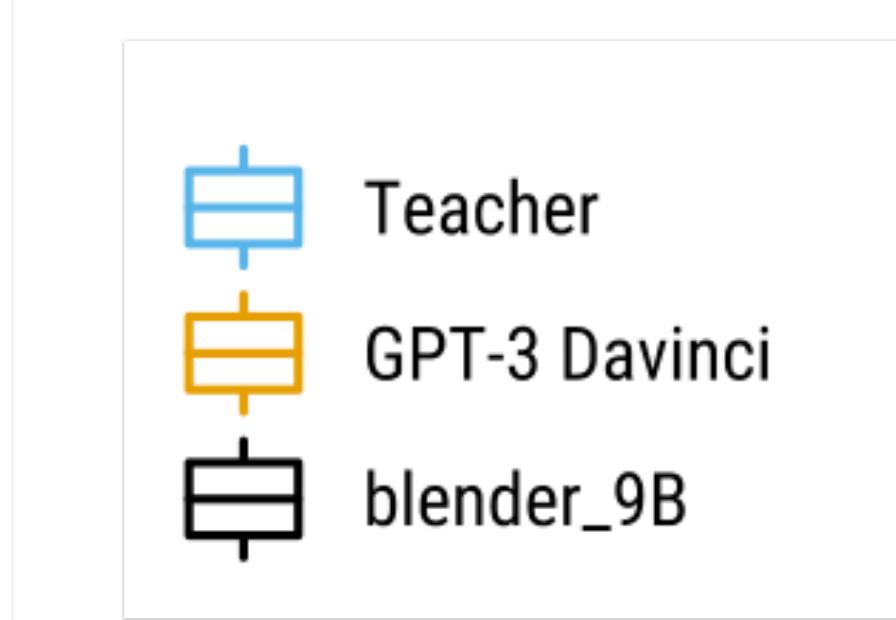
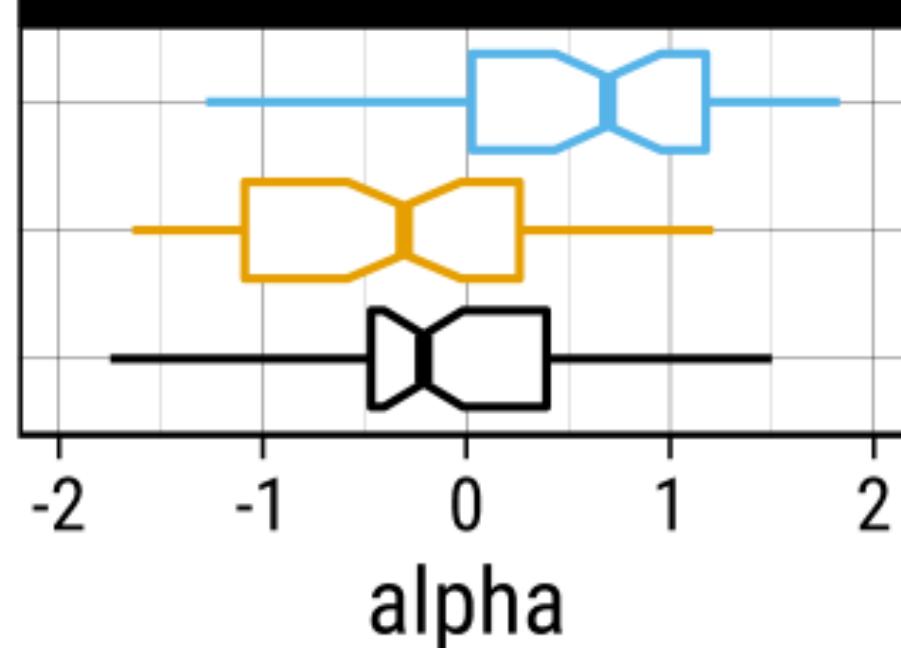
understanding the student



helping the student



helping the student



The BEA 2023 Shared Task

Generating AI Teacher Responses in Educational Dialogues

<https://sig-edu.org/sharedtask/2023>

Anaïs Tack • Ekaterina Kochmar • Zheng Yuan • Serge Bibauw • Chris Piech

Data Samples

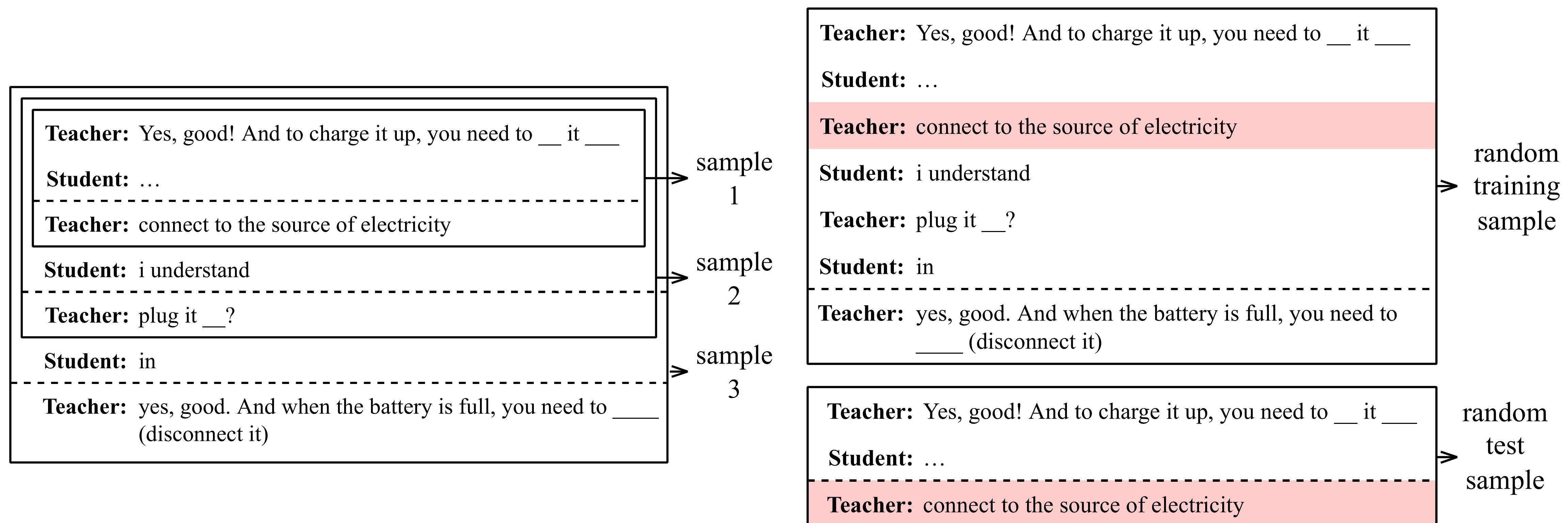
- **TSCC** data
(Caines et al., 2020)
- **excerpts** extracted with sliding windows
- at most **100 tokens** (copyright)
 - dialogue context
 - teacher reference
- **selected** samples with **eliciting** and/or **scaffolding** in reference

| SPEAKER | UTTERANCE | |
|----------|--|--|
| 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 | |
| Teacher: | yes, good. And when the battery is full, you need to __ (disconnect it) | |

DIALOGUE CONTEXT
= REFERENCE RESPONSE

Figure 1: An Example of a Sample Taken from the *Teacher-Student Chatroom Corpus*

Conflicts Between Data Samples



CodaLab Competition

- Development Phase

March 24 → April 30

- **Training**

2,747 samples

- **Held-out**

305 samples (random)

- Evaluation Phase

May 1 → May 5

- **Test**

273 samples (no conflicts)

The screenshot shows the CodaLab competition interface for the BEA 2023 Shared Task. At the top, there's a header with the CodaLab logo, search and account links, and browser tabs. Below the header, the title "Competition" is displayed. A "SIG EDU" logo is present. The main content area is titled "BEA 2023 Shared Task on Generating AI Teacher Responses in Educational Dialogues". It shows the competition is organized by anaistack and currently active. The "First phase" is labeled "Development" and ends on March 24, 2023, at 11:59 a.m. UTC. The "Competition Ends" is on May 6, 2023, at 11:59 a.m. UTC. Navigation tabs include "Learn the Details" (selected), "Phases", "Participate", "Results", and "Forums". Sub-sections under "Learn the Details" are "Overview" (selected), "Evaluation", and "Terms and Conditions". To the right, the task title "BEA 2023 Shared Task" and subtitle "Generating AI Teacher Responses in Educational Dialogues" are shown. It notes the task was organized during the BEA workshop at ACL 2023 in Toronto, Canada, on July 13, 2023. The website is listed as <https://sig-edu.org/sharedtask/2023>. A red note states "Participation on CodaLab will only be accepted upon registration: <https://forms.gle/JhDtAMijwadNmgKv7>". A detailed description follows, mentioning conversational agents' potential in education, their roles, and challenges like determining if they're good AI teachers. The footer includes links to GitHub, About, Privacy and Terms, and version v1.6.

Teams and Systems

Teams and Systems

Teams and Systems

Teams and Systems

Evaluation Metrics

BERTScore

(Zhang* et al., 2020)

Compare: *referenced metric*

- reference (teacher) response
- submitted (generated) response

Scores: precision, recall, **F1-score**

DialogRPT

(Gao et al., 2020)

Compare: *reference-free metric*

- preceding dialogue context
- submitted (generated) response

Scores:

- updown (up vs. down votes), human vs. random, human vs. machine
- **final = ensemble of all scores**

Automated Evaluation

| Team | System | BERTScore | | | DialogRPT | | | | Rank |
|----------------------|-----------------------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|------|
| | | P | R | F1 | U | HvR | HvM | Final | |
| NAIST | NAISTeacher | 0.71 (9) | 0.71 (1) | 0.71 (1) | 0.48 (2) | 0.98 (1) | 1.00 (1) | 0.46 (2) | 1.5 |
| NBU | ADAIO | 0.72 (4) | 0.69 (3) | 0.71 (3) | 0.40 (5) | 0.97 (2) | 0.98 (5) | 0.37 (3) | 3.0 |
| Cornell | GPT-4 ^(TP) | 0.71 (7) | 0.69 (2) | 0.70 (5) | 0.52 (1) | 0.86 (8) | 0.98 (2) | 0.47 (1) | 3.0 |
| aiitis | S-ICL | 0.72 (3) | 0.69 (5) | 0.70 (4) | 0.40 (4) | 0.92 (5) | 0.98 (4) | 0.36 (5) | 4.5 |
| RETUYT-InCo | OPT-2.7B | 0.74 (1) | 0.68 (6) | 0.71 (2) | 0.38 (7) | 0.90 (7) | 0.96 (9) | 0.35 (7) | 4.5 |
| Cornell | GPT-4 | 0.72 (5) | 0.69 (4) | 0.70 (6) | 0.40 (6) | 0.93 (4) | 0.98 (3) | 0.36 (6) | 6.0 |
| Data Science-NLP-HSG | Untrained | 0.72 (6) | 0.63 (8) | 0.67 (8) | 0.41 (3) | 0.93 (3) | 0.95 (10) | 0.37 (4) | 6.0 |
| RETUYT-InCo | Alpaca | 0.72 (2) | 0.68 (7) | 0.70 (7) | 0.37 (8) | 0.91 (6) | 0.96 (7) | 0.34 (8) | 7.5 |
| DT | DistilGPT2 | 0.67 (10) | 0.62 (9) | 0.64 (10) | 0.36 (9) | 0.75 (10) | 0.96 (6) | 0.29 (9) | 9.5 |
| TanTanLabs | zero-shot-with-filler | 0.71 (8) | 0.60 (10) | 0.65 (9) | 0.32 (10) | 0.85 (9) | 0.96 (8) | 0.29 (10) | 9.5 |
| TEACHER | REFERENCE | 1.00 | 1.00 | 1.00 | 0.37 | 0.86 | 0.99 | 0.32 | |

Table 2: Leaderboard for the evaluation phase with scores and ranks for BERTScore (P = precision, R = recall) and DialogRPT (U = updown, HvR = human vs. rand, HvM = human vs. machine)

Automated Evaluation

- best system = **NAISTeacher**
GPT-3.5 Turbo with a **prompt and rerank** approach

| Team | System | BERTScore | | | DialogRPT | | | Rank | |
|----------------------|-----------------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|-----|
| | | P | R | F1 | U | HvR | HvM | | |
| NAIST | NAISTeacher | 0.71 (9) | 0.71 (1) | 0.71 (1) | 0.48 (2) | 0.98 (1) | 1.00 (1) | 0.46 (2) | 1.5 |
| NBU | ADAIO | 0.72 (4) | 0.69 (3) | 0.71 (3) | 0.40 (5) | 0.97 (2) | 0.98 (5) | 0.37 (3) | 3.0 |
| Cornell | GPT-4 ^(TP) | 0.71 (7) | 0.69 (2) | 0.70 (5) | 0.52 (1) | 0.86 (8) | 0.98 (2) | 0.47 (1) | 3.0 |
| aiitis | S-ICL | 0.72 (3) | 0.69 (5) | 0.70 (4) | 0.40 (4) | 0.92 (5) | 0.98 (4) | 0.36 (5) | 4.5 |
| RETUYT-InCo | OPT-2.7B | 0.74 (1) | 0.68 (6) | 0.71 (2) | 0.38 (7) | 0.90 (7) | 0.96 (9) | 0.35 (7) | 4.5 |
| Cornell | GPT-4 | 0.72 (5) | 0.69 (4) | 0.70 (6) | 0.40 (6) | 0.93 (4) | 0.98 (3) | 0.36 (6) | 6.0 |
| Data Science-NLP-HSG | Untrained | 0.72 (6) | 0.63 (8) | 0.67 (8) | 0.41 (3) | 0.93 (3) | 0.95 (10) | 0.37 (4) | 6.0 |
| RETUYT-InCo | Alpaca | 0.72 (2) | 0.68 (7) | 0.70 (7) | 0.37 (8) | 0.91 (6) | 0.96 (7) | 0.34 (8) | 7.5 |
| DT | DistilGPT2 | 0.67 (10) | 0.62 (9) | 0.64 (10) | 0.36 (9) | 0.75 (10) | 0.96 (6) | 0.29 (9) | 9.5 |
| TanTanLabs | zero-shot-with-filler | 0.71 (8) | 0.60 (10) | 0.65 (9) | 0.32 (10) | 0.85 (9) | 0.96 (8) | 0.29 (10) | 9.5 |
| TEACHER | REFERENCE | 1.00 | 1.00 | 1.00 | 0.37 | 0.86 | 0.99 | 0.32 | |

Table 2: Leaderboard for the evaluation phase with scores and ranks for BERTScore (P = precision, R = recall) and DialogRPT (U = updown, HvR = human vs. rand, HvM = human vs. machine)

Automated Evaluation

- best system = **NAISTeacher**
GPT-3.5 Turbo with a **prompt and rerank** approach
- **sensitivity of DialogRPT** towards presence/absence of **prefix**

teacher: <response>



| Team | System | BERTScore | | | DialogRPT | | | | Rank |
|----------------------|-----------------------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|------|
| | | P | R | F1 | U | HvR | HvM | Final | |
| NAIST | NAISTeacher | 0.71 (9) | 0.71 (1) | 0.71 (1) | 0.48 (2) | 0.98 (1) | 1.00 (1) | 0.46 (2) | 1.5 |
| NBU | ADAIO | 0.72 (4) | 0.69 (3) | 0.71 (3) | 0.40 (5) | 0.97 (2) | 0.98 (5) | 0.37 (3) | 3.0 |
| Cornell | GPT-4 ^(TP) | 0.71 (7) | 0.69 (2) | 0.70 (5) | 0.52 (1) | 0.86 (8) | 0.98 (2) | 0.47 (1) | 3.0 |
| aiitis | S-ICL | 0.72 (3) | 0.69 (5) | 0.70 (4) | 0.40 (4) | 0.92 (5) | 0.98 (4) | 0.36 (5) | 4.5 |
| RETUYT-InCo | OPT-2.7B | 0.74 (1) | 0.68 (6) | 0.71 (2) | 0.38 (7) | 0.90 (7) | 0.96 (9) | 0.35 (7) | 4.5 |
| Cornell | GPT-4 | 0.72 (5) | 0.69 (4) | 0.70 (6) | 0.40 (6) | 0.93 (4) | 0.98 (3) | 0.36 (6) | 6.0 |
| Data Science-NLP-HSG | Untrained | 0.72 (6) | 0.63 (8) | 0.67 (8) | 0.41 (3) | 0.93 (3) | 0.95 (10) | 0.37 (4) | 6.0 |
| RETUYT-InCo | Alpaca | 0.72 (2) | 0.68 (7) | 0.70 (7) | 0.37 (8) | 0.91 (6) | 0.96 (7) | 0.34 (8) | 7.5 |
| DT | DistilGPT2 | 0.67 (10) | 0.62 (9) | 0.64 (10) | 0.36 (9) | 0.75 (10) | 0.96 (6) | 0.29 (9) | 9.5 |
| TanTanLabs | zero-shot-with-filler | 0.71 (8) | 0.60 (10) | 0.65 (9) | 0.32 (10) | 0.85 (9) | 0.96 (8) | 0.29 (10) | 9.5 |
| TEACHER | REFERENCE | 1.00 | 1.00 | 1.00 | 0.37 | 0.86 | 0.99 | 0.32 | |

Table 2: Leaderboard for the evaluation phase with scores and ranks for BERTScore (P = precision, R = recall) and DialogRPT (U = updown, HvR = human vs. rand, HvM = human vs. machine)

Automated Evaluation

- best system = **NAISTeacher**
GPT-3.5 Turbo with a **prompt and rerank** approach
- **sensitivity of DialogRPT** towards presence/absence of **prefix**

teacher: <response>



| Team | System | BERTScore | | | DialogRPT | | | | Rank |
|----------------------|-----------------------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|------|
| | | P | R | F1 | U | HvR | HvM | Final | |
| NAIST | NAISTeacher | 0.71 (9) | 0.71 (1) | 0.71 (1) | 0.48 (2) | 0.98 (1) | 1.00 (1) | 0.46 (2) | 1.5 |
| NBU | ADAIO | 0.72 (4) | 0.69 (3) | 0.71 (3) | 0.40 (5) | 0.97 (2) | 0.98 (5) | 0.37 (3) | 3.0 |
| Cornell | GPT-4 ^(TP) | 0.71 (7) | 0.69 (2) | 0.70 (5) | 0.52 (1) | 0.86 (8) | 0.98 (2) | 0.47 (1) | 3.0 |
| aiitis | S-ICL | 0.72 (3) | 0.69 (5) | 0.70 (4) | 0.40 (4) | 0.92 (5) | 0.98 (4) | 0.36 (4) | 4.5 |
| RETUYT-InCo | OPT-2.7B | 0.74 (1) | 0.68 (6) | 0.71 (2) | 0.38 (7) | 0.90 (7) | 0.98 (9) | 0.34 (7) | 4.5 |
| Cornell | GPT-4 | 0.72 (5) | 0.69 (4) | 0.70 (6) | 0.40 (6) | 0.93 (4) | 0.98 (3) | 0.36 (6) | 6.0 |
| Data Science-NLP-HSG | Untrained | 0.72 (6) | 0.63 (8) | 0.67 (8) | 0.41 (3) | 0.93 (3) | 0.95 (10) | 0.37 (4) | 6.0 |
| RETUYT-InCo | Alpaca | 0.72 (2) | 0.68 (7) | 0.70 (7) | 0.37 (8) | 0.91 (6) | 0.96 (7) | 0.34 (8) | 7.5 |
| DT | DistilGPT2 | 0.67 (10) | 0.62 (9) | 0.64 (10) | 0.36 (9) | 0.75 (10) | 0.96 (6) | 0.29 (9) | 9.5 |
| TanTanLabs | zero-shot-with-filler | 0.71 (8) | 0.60 (10) | 0.65 (9) | 0.32 (10) | 0.85 (9) | 0.96 (8) | 0.29 (10) | 9.5 |
| TEACHER | REFERENCE | 1.00 | 1.00 | 1.00 | 0.37 | 0.86 | 0.99 | 0.32 | |

Table 2: Leaderboard for the evaluation phase with scores and ranks for BERTScore (P = precision, R = recall) and DialogRPT (U = updown, HvR = human vs. rand, HvM = human vs. machine)

Automated Evaluation

- best system = **NAISTeacher**
GPT-3.5 Turbo with a **prompt and rerank** approach
- **sensitivity of DialogRPT** towards presence/absence of **prefix**

teacher: <response>



| Team | System | BERTScore | | | DialogRPT | | | | Rank |
|----------------------|-----------------------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|------|
| | | P | R | F1 | U | HvR | HvM | Final | |
| NAIST | NAISTeacher | 0.71 (9) | 0.71 (1) | 0.71 (1) | 0.48 (2) | 0.98 (1) | 1.00 (1) | 0.46 (2) | 1.5 |
| NBU | ADAIO | 0.72 (4) | 0.69 (3) | 0.71 (3) | 0.40 (5) | 0.97 (2) | 0.98 (5) | 0.37 (3) | 3.0 |
| Cornell | GPT-4 ^(TP) | 0.71 (7) | 0.69 (2) | 0.70 (5) | 0.52 (1) | 0.86 (8) | 0.98 (2) | 0.47 (1) | 3.0 |
| aiitis | S-ICL | 0.72 (3) | 0.69 (5) | 0.70 (4) | 0.40 (4) | 0.92 (5) | 0.98 (4) | 0.36 (4) | 4.5 |
| RETUYT-InCo | OPT-2.7B | 0.74 (1) | 0.68 (6) | 0.71 (2) | 0.38 (7) | 0.90 (7) | 0.98 (9) | 0.34 (7) | 4.5 |
| Cornell | GPT-4 | 0.72 (5) | 0.69 (4) | 0.70 (6) | 0.40 (6) | 0.93 (4) | 0.98 (3) | 0.36 (6) | 6.0 |
| Data Science-NLP-HSG | Untrained | 0.72 (6) | 0.63 (8) | 0.67 (8) | 0.41 (3) | 0.93 (3) | 0.95 (10) | 0.37 (4) | 6.0 |
| RETUYT-InCo | Alpaca | 0.72 (2) | 0.68 (7) | 0.70 (7) | 0.37 (8) | 0.91 (6) | 0.96 (7) | 0.34 (8) | 7.5 |
| DT | DistilGPT2 | 0.67 (10) | 0.62 (9) | 0.64 (10) | 0.36 (9) | 0.75 (10) | 0.96 (6) | 0.29 (9) | 9.5 |
| TanTanLabs | zero-shot-with-filler | 0.71 (8) | 0.60 (10) | 0.65 (9) | 0.32 (10) | 0.85 (9) | 0.96 (8) | 0.29 (10) | 9.5 |
| TEACHER | REFERENCE | 1.00 | 1.00 | 1.00 | 0.37 | 0.86 | 0.99 | 0.32 | |

Table 2: Leaderboard for the evaluation phase with scores and ranks for BERTScore (P = precision, R = recall) and DialogRPT (U = updown, HvR = human vs. rand, HvM = human vs. machine)

Automated Evaluation

- best system = **NAISTeacher**
GPT-3.5 Turbo with a **prompt and rerank** approach
- **sensitivity of DialogRPT** towards presence/absence of **prefix**
teacher : <response>
- **well-engineered prompts** lead to a high rank on BERTScore and DialogRPT

#1 : good teaching examples

#2 : teaching approaches and goals

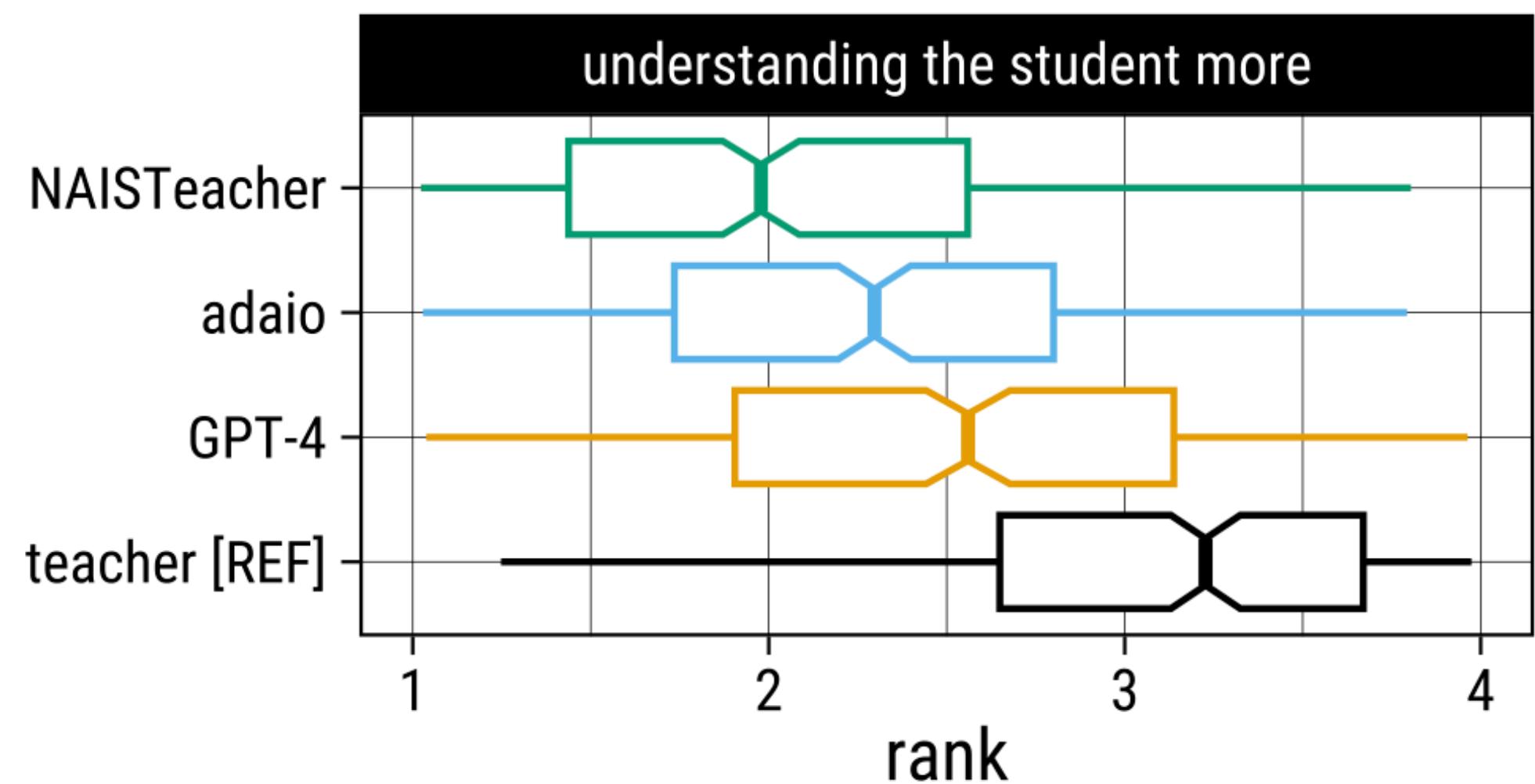
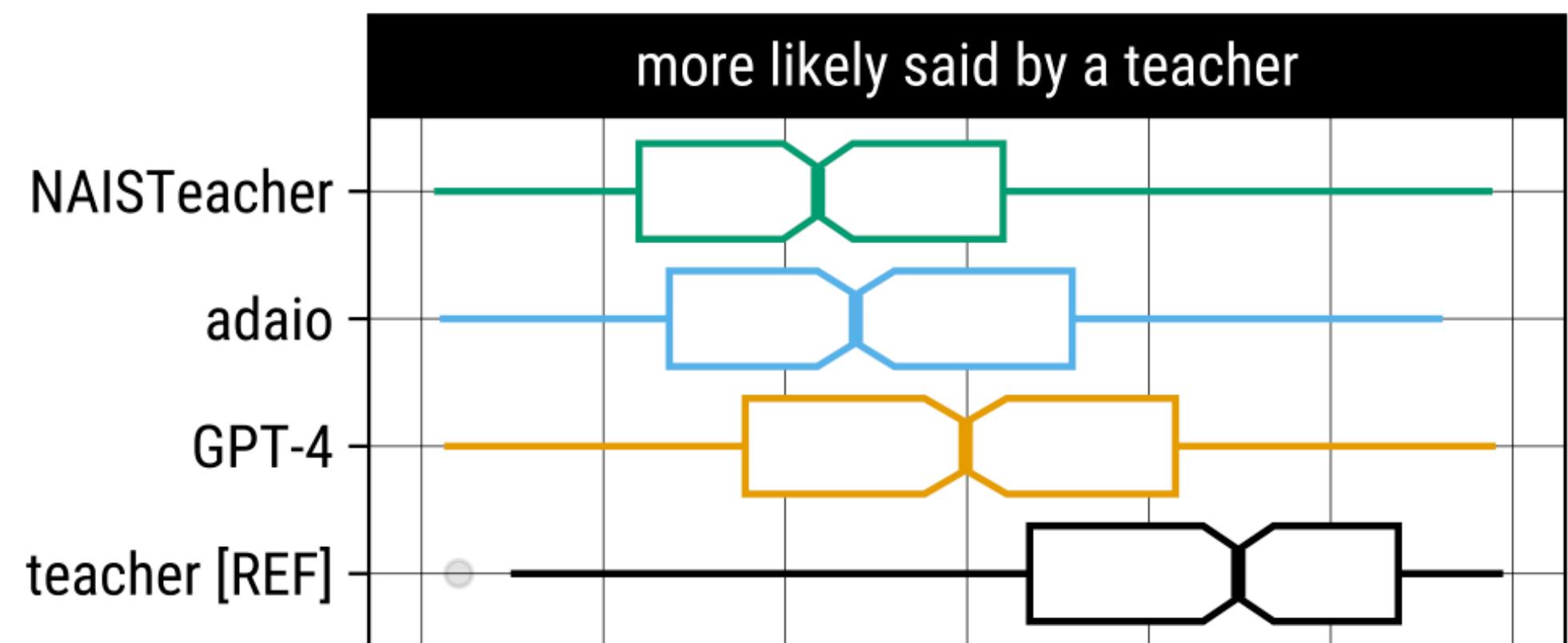
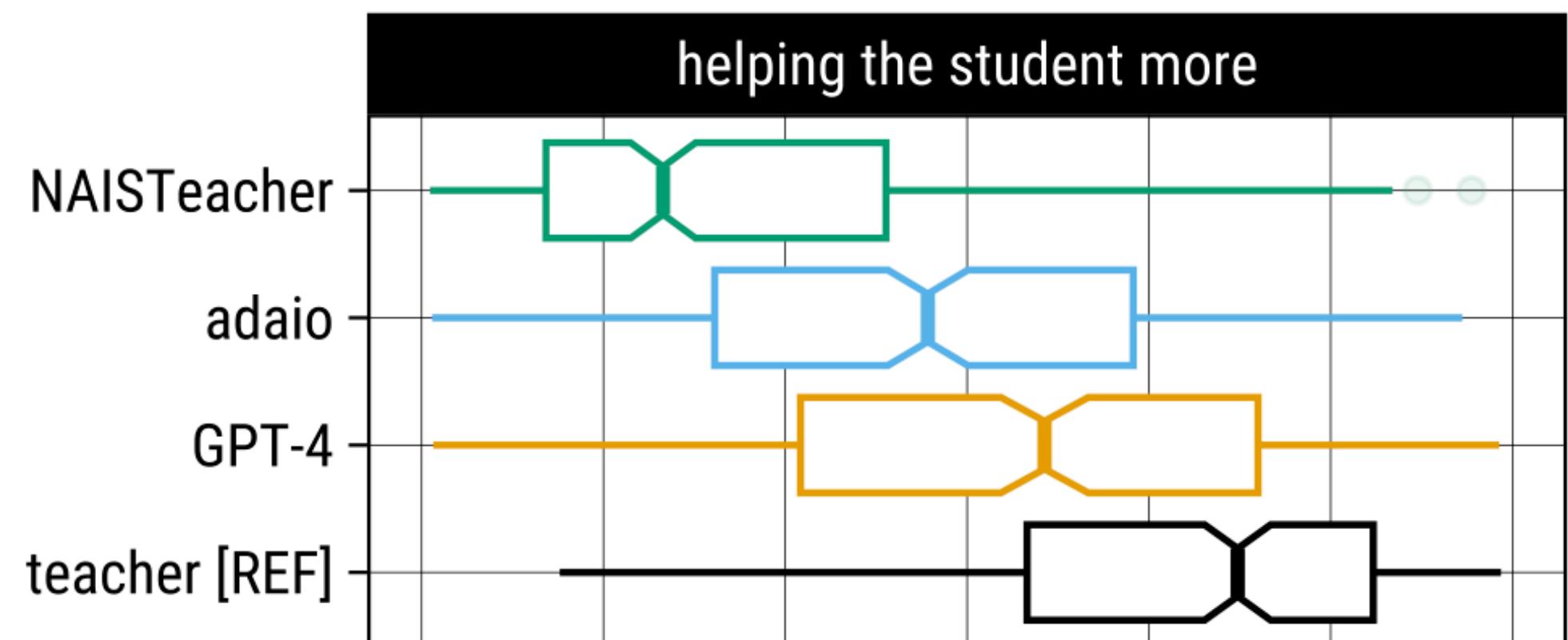
| Team | System | BERTScore | | | DialogRPT | | | Rank | |
|----------------------|-----------------------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|-----|
| | | P | R | F1 | U | HvR | HvM | | |
| NAIST | NAISTeacher | 0.71 (9) | 0.71 (1) | 0.71 (1) | 0.48 (2) | 0.98 (1) | 1.00 (1) | 0.46 (2) | 1.5 |
| NBU | ADAIO | 0.72 (4) | 0.69 (3) | 0.71 (3) | 0.40 (5) | 0.97 (2) | 0.98 (5) | 0.37 (3) | 3.0 |
| Cornell | GPT-4 ^(TP) | 0.71 (7) | 0.69 (2) | 0.70 (5) | 0.52 (1) | 0.86 (8) | 0.98 (2) | 0.47 (1) | 3.0 |
| aiitis | S-ICL | 0.72 (3) | 0.69 (5) | 0.70 (4) | 0.40 (4) | 0.92 (5) | 0.98 (4) | 0.36 (5) | 4.5 |
| RETUYT-InCo | OPT-2.7B | 0.74 (1) | 0.68 (6) | 0.71 (2) | 0.38 (7) | 0.90 (7) | 0.96 (9) | 0.35 (7) | 4.5 |
| Cornell | GPT-4 | 0.72 (5) | 0.69 (4) | 0.70 (6) | 0.40 (6) | 0.93 (4) | 0.98 (3) | 0.36 (6) | 6.0 |
| Data Science-NLP-HSG | Untrained | 0.72 (6) | 0.63 (8) | 0.67 (8) | 0.41 (3) | 0.93 (3) | 0.95 (10) | 0.37 (4) | 6.0 |
| RETUYT-InCo | Alpaca | 0.72 (2) | 0.68 (7) | 0.70 (7) | 0.37 (8) | 0.91 (6) | 0.96 (7) | 0.34 (8) | 7.5 |
| DT | DistilGPT2 | 0.67 (10) | 0.62 (9) | 0.64 (10) | 0.36 (9) | 0.75 (10) | 0.96 (6) | 0.29 (9) | 9.5 |
| TanTanLabs | zero-shot-with-filler | 0.71 (8) | 0.60 (10) | 0.65 (9) | 0.32 (10) | 0.85 (9) | 0.96 (8) | 0.29 (10) | 9.5 |
| TEACHER | REFERENCE | 1.00 | 1.00 | 1.00 | 0.37 | 0.86 | 0.99 | 0.32 | |

Table 2: Leaderboard for the evaluation phase with scores and ranks for BERTScore (P = precision, R = recall) and DialogRPT (U = updown, HvR = human vs. rand, HvM = human vs. machine)

Human Evaluation

- 298 raters recruited from Prolific
 - 1. English-speaking country
 - 2. English L1
 - 3. working in **education & training** sector
- feedback from raters:
teacher chatroom language is “unprofessional”

ranking



NAISTeacher

A Prompt and Rerank Approach

- built on **GPT-3.5-Turbo**
 - zero-shot prompts
 - few-shot prompts
 - handcrafted examples
 - generated examples
(chain-of-thought prompting)
 - iterative examples
- **reranking** with DialogRPT

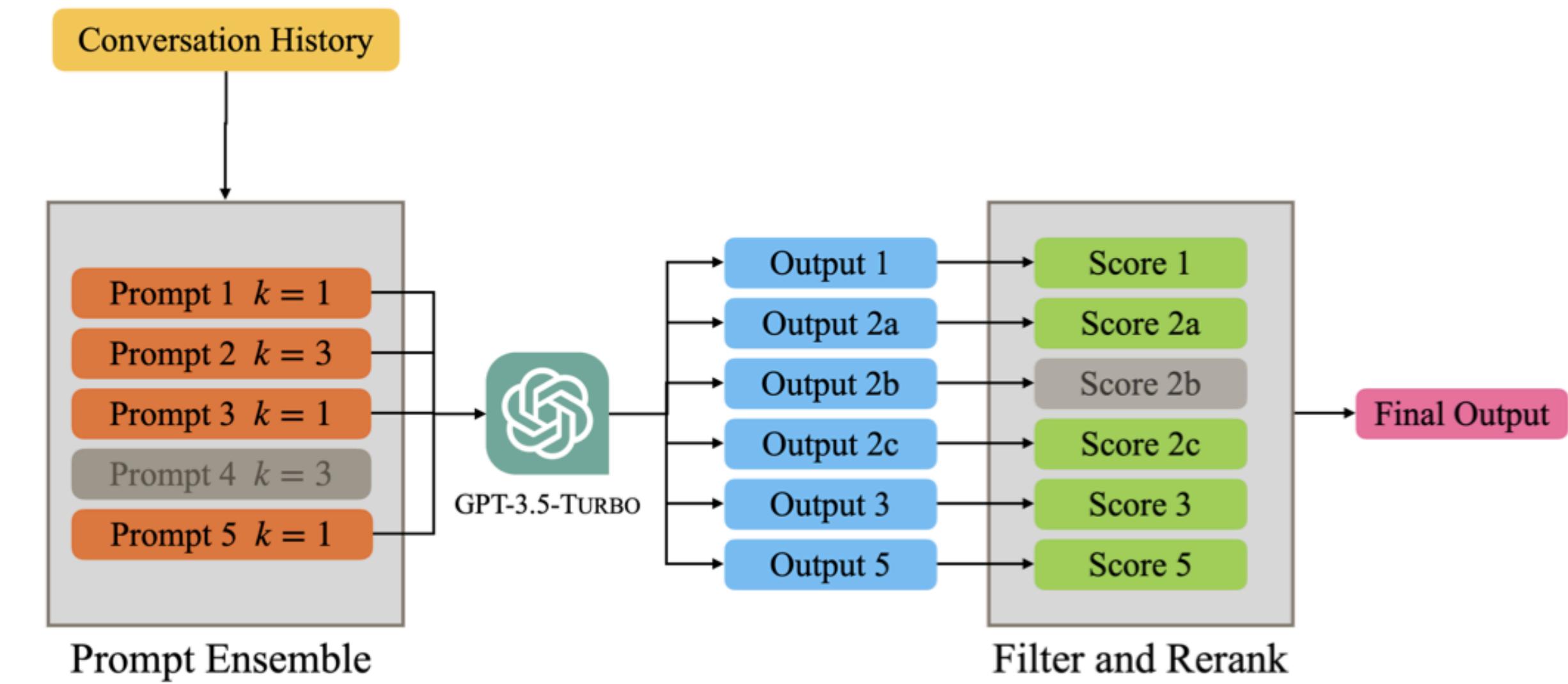


Figure 1: The NAISTeacher system overview. Prompts are chosen for the ensemble based on the role of the speaker of the final utterance of the conversation history. $k = 1$ and $k = 3$ refer to the number of candidates solicited by the prompt (one and three respectively).

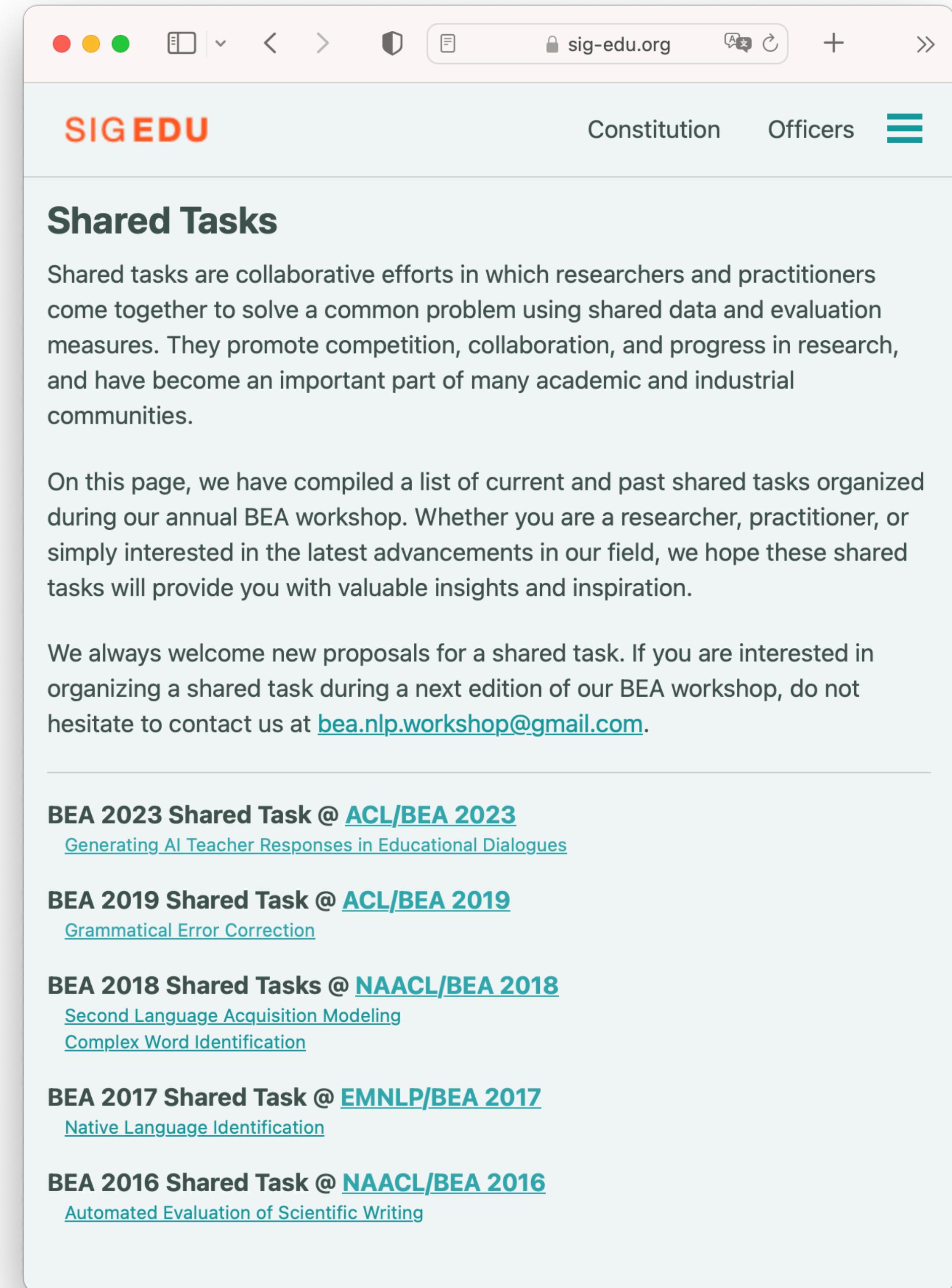
Thanks to

- **Shared task participants**
Adaeze Adigwe, Aijun An, Alexis Baladón, Rabin Banjade, Luis Chiruzzo, Siegfried Handschuh, Yann Hicke, Thomas Huber, Tanay Gahlot, Tushaar Gangavarapu, Wentao Guo, Abhishek Masand, Christina Niklaus, Adam Nohejl, Aiala Rosá, Ignacio Sastre, Amin Omidvar, Justin Vasselli, Christopher Vasselli, Taro Watanabe, Zheng Yuan
- **Anonymous raters on Prolific**
for taking the time to provide us with additional feedback
- **Andrew Caines**
for letting us use the TSCC corpus for this shared task

BEA Shared Tasks

We always welcome **new proposals for a shared task.**

If you are interested in organizing a shared task during the next edition of the BEA workshop, don't hesitate to send a proposal to
bea.nlp.workshop@gmail.com.



The screenshot shows a web browser window for the SIG EDU website at sig-edu.org. The page title is "SIG EDU". The main content area is titled "Shared Tasks". A descriptive paragraph explains what shared tasks are and their benefits. Below this, there is a section for past shared tasks, listing entries from 2023, 2019, 2018, 2017, and 2016, each with a brief description of the task.

SIG EDU

Constitution Officers 

Shared Tasks

Shared tasks are collaborative efforts in which researchers and practitioners come together to solve a common problem using shared data and evaluation measures. They promote competition, collaboration, and progress in research, and have become an important part of many academic and industrial communities.

On this page, we have compiled a list of current and past shared tasks organized during our annual BEA workshop. Whether you are a researcher, practitioner, or simply interested in the latest advancements in our field, we hope these shared tasks will provide you with valuable insights and inspiration.

We always welcome new proposals for a shared task. If you are interested in organizing a shared task during a next edition of our BEA workshop, do not hesitate to contact us at bea.nlp.workshop@gmail.com.

BEA 2023 Shared Task @ [ACL/BEA 2023](#)
[Generating AI Teacher Responses in Educational Dialogues](#)

BEA 2019 Shared Task @ [ACL/BEA 2019](#)
[Grammatical Error Correction](#)

BEA 2018 Shared Tasks @ [NAACL/BEA 2018](#)
[Second Language Acquisition Modeling](#)
[Complex Word Identification](#)

BEA 2017 Shared Task @ [EMNLP/BEA 2017](#)
[Native Language Identification](#)

BEA 2016 Shared Task @ [NAACL/BEA 2016](#)
[Automated Evaluation of Scientific Writing](#)

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>
- Bradley, R. A., & Terry, M. E. (1952). Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons. *Biometrika*, 39(3/4), 324. <https://doi.org/10.2307/2334029>
- 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.
- 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.13>
- Gao, X., Zhang, Y., Galley, M., Brockett, C., & Dolan, B. (2020). Dialogue Response Ranking Training with Large-Scale Human Feedback Data. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 386-395. <https://doi.org/10.18653/v1/2020.emnlp-main.28>
- 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/3f5ee243547dee91fb0d053c1c4a845aa-Abstract.html>
- Vasselli, J., Vasselli, C., Nohejl, A., & Watanabe, T. (2023). NAISTeacher: A Prompt and Rerank Approach to Generating Teacher Utterances in Educational Dialogues. *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications*, to appear.
- 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>
- Zhang*, T., Kishore*, V., Wu*, F., Weinberger, K. Q., & Artzi, Y. (2020). BERTScore: Evaluating text generation with BERT. *International Conference on Learning Representations*.