Developing Language Models for Power-Affirming Writing Feedback

Mei Tan, Chris Mah, Alexa Sparks, Deepak Varuvel Dennison, Dorottya Demszky

Stanford Graduate School of Education

HAI FIVE — Celebrating 5 Years of Impact

LLMs are being used to provide students with feedback on their writing.

But, studies show that these models do not behave like expert teachers. Out-of-the-box, generative AI models are not pedagogically sound in ways that are not well understood.

Cognitive Task Analysis

What are expert teachers doing that LLMs don't?

We conduct 12 cognitive task analysis interviews with Stanford GSE STEP and Hollyhock Fellowship alumni. Teachers followed a think-aloud protocol while providing feedback on middle- and high-school literary analysis and narrative writing assignments.

Standards Alignment

This is pretty solid context for the evidence you have coming up and a set up for how you will answer the prompt - you got the prompt right away which is great!

Voice

"I" is your pronoun you are important make sure you capitalize

I really like this idea about moving on! I would love to hear more about what you think it takes to move on from the past.

Uptake + Questions

Good job making an inference. Can you think of more reasons that she might be embarrassed by her family's Christmas? What does that show about her feelings about her own culture versus American culture?

Which exact words or phrases in this passage contain the metaphor? Or are there several?

Strategies

Sometimes reading what you write out loud can help us catch things that don't make sense, mistakes, or help us to think about things that could be missing. I'd encourage you to read this aloud to yourself or a friend to help make it clearer.

Personal Connection

Great Story! I'm glad you made the choice to stay in school :)

*Expert teachers invite and legitimize students' ideas, positioning them authors.

We adapt and operationalize Pedersen's framework of power-affirming feedback.

Power-Affirming

Positions the student as an author, legitimizes students' interpretations, uses student-friendly language, incorporates students' words, and pushes the student to consider alternative perspectives and deepen the analysis.

Power-Concealing

Feedback positions the teacher as the authority, reinforces conventional models for writing, relies on discipline-specific terminology or unfamiliar language without contextualizing, and positions the student as an outsider to academic discourses.

...can we measure power-affirmingness?

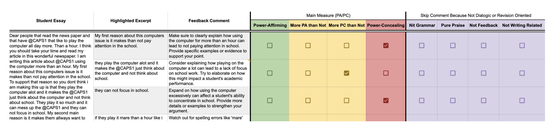
	More Power-Concealing	More Power-Affirming
As a teenager, she was embarrassed by both her family and her Chinese culture. She used sentences like "I wanted to disappear" and "I was stunned into silence for the rest of the night," which highlights that her family's behavior is making her feel really embarrassed. I don't think Robert is worth it. As an adult, she was at peace with her Chinese culture and appreciates that her mom did. In the text, it states "I was able to fully appreciate her lesson" and "she had chosen all my favorite foods." This shows that she appreciated what her mother did and that she is at peace with her Chinese heritage.	Good claim, but you would want to add the part about how she feels as an adult to this sentence.	I noticed that you didn't say how things are different? What was the main difference you noticed?
	Since this is a literary analysis paragraph, you should avoid giving your own opinion about the characters.	Why don't you think Robert is worth it? Can you please explain?
	Is this clearly articulated? It seems wordy and hard to understand. Maybe "what"?	Sometimes reading what we write out loud can help us to think about things that could be missing. I'd encourage you to read this aloud.
	I'm not sure this is a strong quote to support your argument.	Does Amy feel differently about these foods now than she did before? What do you think changed?
	We need more of a connection between her childhood perspective and her adult perspective.	What exactly about her heritage is she at peace with now that she was not before?

Data Collection

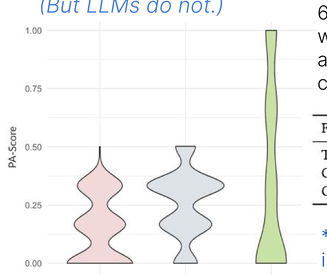
- We recruit 18 teachers from the Hollyhock Fellowship alumni network.
  Teachers gave in-line feedback (excerpt-comment pairs) on 300 persuasive, narrative, and literary analysis essays from the ASAP-AES dataset.
  → 1506 excerpt-comment pairs
- We generate 400 excerpt-comment pairs on 60 essays using ChatGPT.

Data Annotation

We train 2 doctoral students of teacher education with MS/HS ELA teaching experience. We scale judgments of power-affirmingness to create a PA-score between 0 and 1. Higher values are more power-affirming.



(But LLMs do not.)



Distributions from 612 annotated teacher-written comments + 400 annotated ChatGPT comments

Feedback Source	Avg. PA	N
Teacher-Written	0.346	612
ChatGPT (prompt 1)	0.148	230
ChatGPT (prompt 2)	0.257	170

*Power-affirming feedback is uncommon, even for expert teachers.

Modeling

**an automated computational measure*

We fine-tune a RoBERTa-based regression model to predict a PA-score for each excerpt-comment feedback pair.

Feedback Source	PA	DRO	
	spearman	acc	f1
Teacher-Written	0.810	0.918	0.868
ChatGPT (all)	0.823	0.992	0.987

Models effectively capture annotator judgments.

Next up: large-scale descriptive study of teachers' feedback practices in partnership with a national ELA PD program.

Can we adapt LLMs to do better?

Prompting

Only goes so far.

Modestly improves the mean but not the max.

Fine-tuning

Can we copy expert teacher behavior?

Reinforcement Learning*

Can we exceed expert teacher behavior?

You are an {grade}th grade English Language Arts teacher.

You will be provided information about a writing assignment and a piece of student writing, denoted by triple quotes.

Your task is to write feedback on the student writing to help your student revise their work.

Your feedback should be clear, specific, and actionable.

Your feedback should also position the student as the author, legitimize the student's ideas, take up the student's words, and push the student to consider alternative perspectives.

You should use conversational student-friendly language appropriate for {grade}th grade.

Format your response as JSON output. The field inlineFeedback should be a list of 5-7 pairs, each containing an exact 'excerpt' from the student's essay and a corresponding feedback 'comment'. Choose short excerpts that are most directly relevant to the feedback comment.

Prompt 1 (Base) | Added in Prompt 2 (Specific)

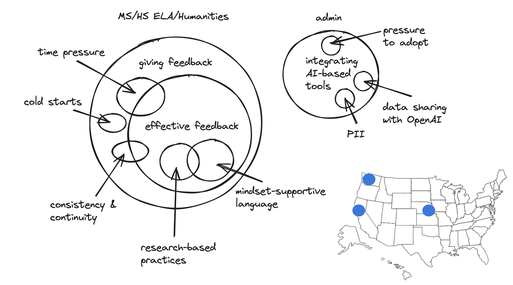
Feedback Source	Avg. PA	Max PA
Teacher-Written	0.346	1.00
Llama 2 (baseline)	0.102	0.502
Llama 2 (fine-tuned)	0.447	1.00
Llama 2 (PPO)	0.588	0.734

Fine-tuning Llama 2 via supervised learning breaks the upper bound on LLM PA-scores.

**ongoing work*

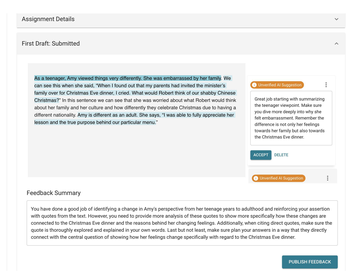
In the real world, to what extent can LLMs assist teachers with writing feedback?

We partner with school districts to understand teachers' feedback practices and challenges.



App Pilot

We develop a web application that generates suggested inline comments and summary feedback on students' written work.



Teachers can edit, delete, or accept AI suggestions, as well as add comments of their own.

Learn more here!

Students submit revisions in response to feedback. We survey students' perception of feedback helpfulness and teachers' perception of revision quality.

Downstream Analyses

- Lexical study of teachers' edits to LLM suggestions.
- Measures of feedback uptake in student revisions.
- Student perception of LLM-generated vs. teacher-written feedback.