Developing Language Models for Power-Affirming Writing Feedback
Mei Tan, Chris Mah, Alexa Sparks, Deepak Varuel Dennison, Dorottya Demszky

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 a good way to present the evidence. You can help me map out a set up for how you would review the prompt — you got the prompt right away which is great!

Update + Questions

Don’t just jump to definitions. Can you think of three or more reasons that one might be embarrassed by her family’s Christmas? What does that show about her feelings about her own culture versus American culture?

Voice

“Is your pronoun use important?” Find your capitals.

Suggest

I really like the idea about moving on / I would like to hear more about what you think it needs to move on from the past.

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 Derschens framework of power-affirming feedback.

**Power-Concealing**

- Positions the student as an author, legitimates student interpretations, uses student-friendly language, incorporates students’ worlds, and pushes the student to consider alternative perspectives and deepen the analysis.

...can we measure power-affirmingness?

**Power-Affirming**

- Positions the student as the authority, showcases conventional models for writing, takes on discipline-specific terminology or unfamiliar language without reproach, and positions the student as an outsider to academic discourses.

Data Collection

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

Data Annotation

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

Modeling

**An automated computational measure**

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

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

<table>
<thead>
<tr>
<th>Feedback Source</th>
<th>Avg. PA</th>
<th>Max PA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Teacher Written</td>
<td>0.346</td>
<td>1.00</td>
</tr>
<tr>
<td>ChatGPT (full)</td>
<td>0.357</td>
<td>0.392</td>
</tr>
</tbody>
</table>

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

Model efficiency: 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?

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

<table>
<thead>
<tr>
<th>Feedback Source</th>
<th>Avg. PA</th>
<th>Max PA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Teacher Written</td>
<td>0.346</td>
<td>1.00</td>
</tr>
<tr>
<td>Llama 1 (fine-tuned)</td>
<td>0.447</td>
<td>1.00</td>
</tr>
<tr>
<td>Llama 2 (fine-tuned)</td>
<td>0.589</td>
<td>0.734</td>
</tr>
</tbody>
</table>

*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.

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.