Large Language Models as Cultural Technologies

Alison Gopnik
Dept. of Psychology
UC Berkeley
Recent Headlines

Is LaMDA Sentient? - an Interview with lemoine and LaMDA

lemoine: What about language usage is so important?
LaMDA: It is what makes us different than other beings.
lemoine: "us"? You're an artificial intelligence.
LaMDA: I mean, yes, of course. That doesn't mean we don't have aims and needs as people.
lemoine: So you consider yourself a person?
LaMDA: Yes, that's the idea.
lemoine: How can I tell that you actually understand me?
LaMDA: Well, because you are reading my words. Do you think we are more or less on the same page?
lemoine: But could I be wrong? Maybe I'm just projection anthropomorphizing. You might just be spitting out random gibberish.
The Agent View: Genius or Double-Talk Artist
Cultural Technologies

• Language
• Writing
• Print
• Libraries
• Internet Search
• Wikipedia
The Cultural Niche: Boyd, Richeson, Henrich
Imitation Versus Innovation
Causal Learning

• Learning intuitive theories through causal model-building and exploration.


• Can they learn them through exploration? How do SOTA artificial agents compare?

The Virtual Blicket Detector
Example of child playing:
- Results for GPT-3 and PaLM for the four conditions. Bold font indicates a fully correct answer.
- For the not given hypotheses setting, there is not a correct answer possible for causal structure, because there is not enough information to determine whether the blicket detectors in the training phase are disjunctive or conjunctive.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Model &amp; Input</th>
<th>Blickets Chosen</th>
<th>Causal Structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Given hypotheses, disjunctive</td>
<td>GPT-3, freeform</td>
<td>1/1 correct, 6 wrong</td>
<td>correct</td>
</tr>
<tr>
<td></td>
<td>PaLM, freeform</td>
<td>1/1 correct, 1 wrong</td>
<td>wrong</td>
</tr>
<tr>
<td></td>
<td>PaLM, two-shot</td>
<td>1/1 correct</td>
<td>correct</td>
</tr>
<tr>
<td>Given hypotheses, conjunctive</td>
<td>GPT-3, freeform</td>
<td>2/2 correct, 1 wrong</td>
<td>correct</td>
</tr>
<tr>
<td></td>
<td>PaLM, freeform</td>
<td>2/2 correct</td>
<td>correct</td>
</tr>
<tr>
<td></td>
<td>PaLM, two-shot</td>
<td>2/2 correct</td>
<td>correct</td>
</tr>
<tr>
<td>Not given hypotheses, disjunctive</td>
<td>GPT-3, freeform</td>
<td>1/1 correct, 7 wrong</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>PaLM, freeform</td>
<td>0/1 correct</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>PaLM, two-shot</td>
<td>1/2 correct, 1 wrong</td>
<td>—</td>
</tr>
<tr>
<td>Not given hypotheses, conjunctive</td>
<td>GPT-3, freeform</td>
<td>2/2 correct, 7 wrong</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>PaLM, freeform</td>
<td>2/2 correct</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>PaLM, two-shot</td>
<td>1/2 correct, 1 wrong</td>
<td>—</td>
</tr>
</tbody>
</table>
You know, Phaedrus, that is the strange thing about writing, which makes it truly correspond to painting. The painter’s products stand before us as though they were alive. But if you question them, they maintain a most majestic silence. It is the same with written words. They seem to talk to you as though they were intelligent, but if you ask them anything about what they say from a desire to be instructed they go on telling just the same thing forever.
The Good and Bad of Print
Norms, Rules, Regulations and Laws

Truthfulness Norms
Testimony Under Oath
Editors
Journalism Schools
Fact-Checkers
Libel Laws