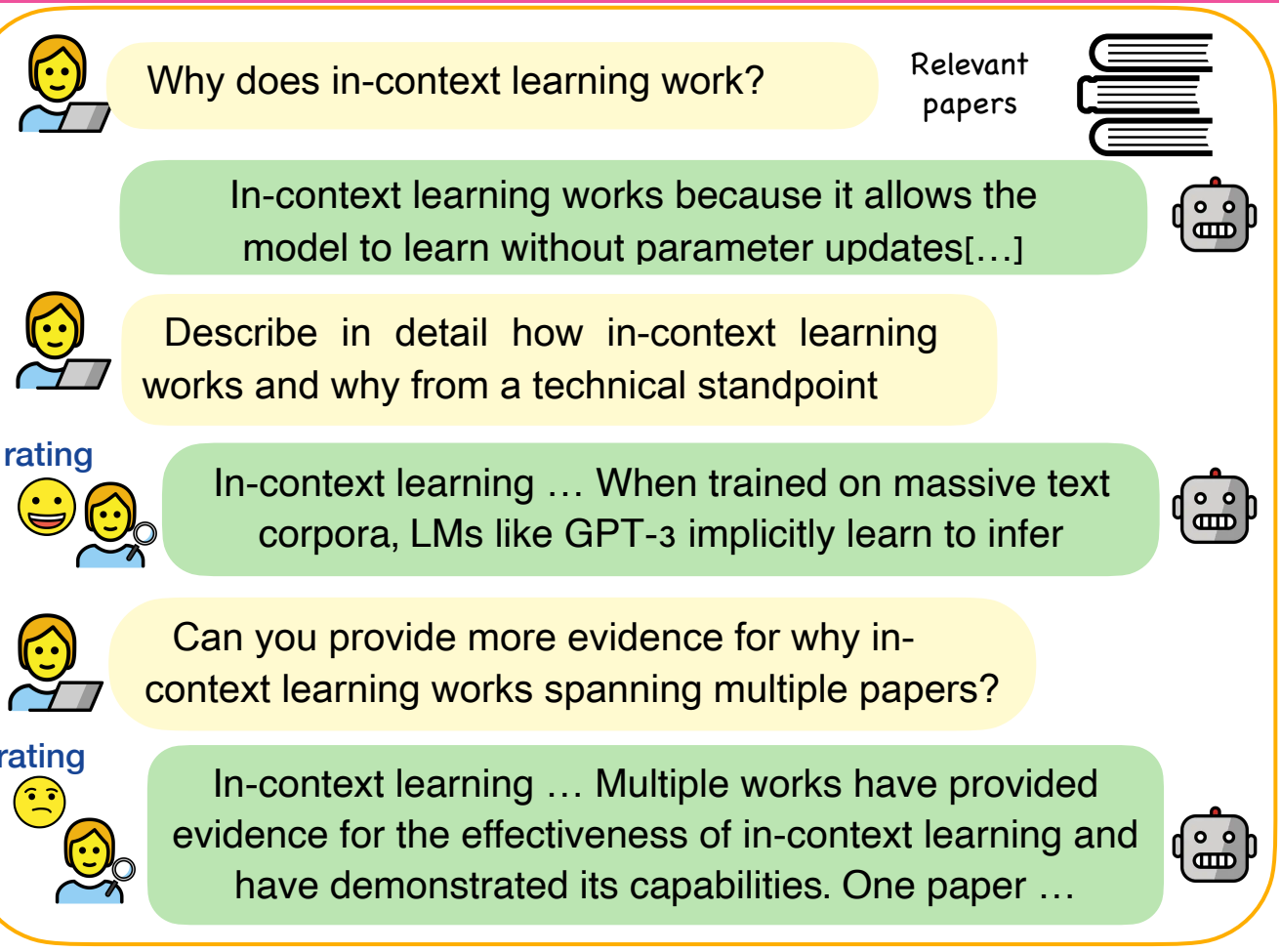
KIWI

A Dataset of Knowledge-Intensive Writing Instructions for Answering Research Questions

Download our datasets here!

Fangyuan Xu[1], Kyle Lo[2], Luca Soldaini[2], Bailey Kuehl[2], Eunsol Choi[1], David Wadden[2]
[1]The University of Texas at Austin, [2] Allen Institute for AI

🤔 Can LLMs follow user instructions to answer complex research questions grounded in relevant papers?



We collected 234 interaction session data with:

- 🤖 LLMs' (GPT-4, GPT-3.5-turbo, LLaMA2) initial responses to a complex ❓ research question based on 📄 relevant papers
- ♻️ A total of 1,260 rounds of:
  - 🎓 Expert-issued instructions
  - 🤖 LLMs' (GPT-4, GPT-3.5-turbo, LLaMA2) responses to the instruction
  - ☑️ Expert's rating of the response and explanation for the rating

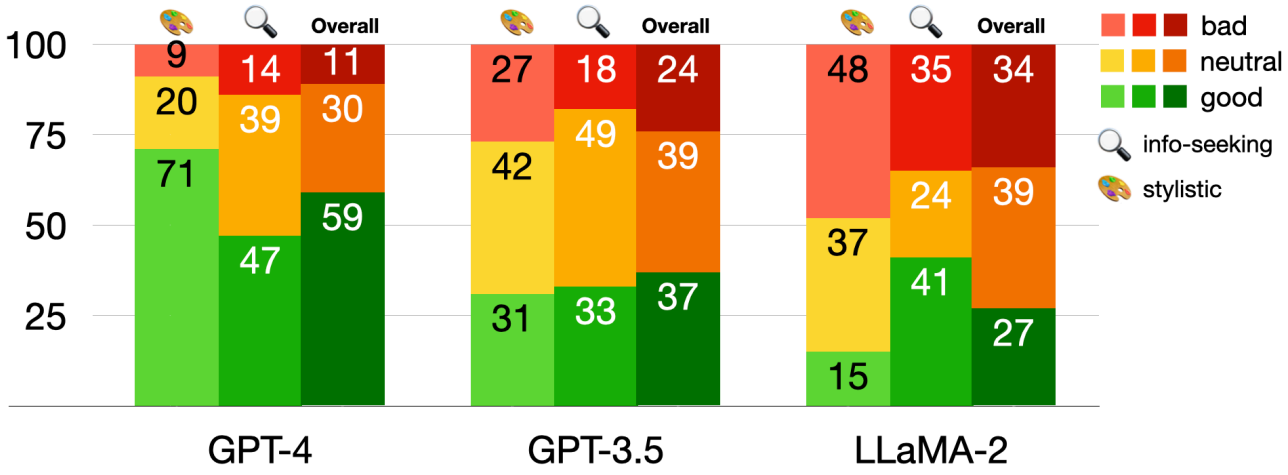
Two types of instructions

🔍 Info-seeking instruction

🎨 Stylistic instruction

❌ Common error patterns

Unrequested change (31%)	*"The model shorted the desired paragraph but removed info from previous answer"*
Ignored (21%)	*"Model does not seem to have reduced any information at all"*
Constraint failure (11%)	*"It gave the definition but not as first sentence."*
Poor integration (11%)	*"Requested info was added but answer now as a weird order"*
Others (28%)	



Our data contains diverse types of instructions

🔍 Info seeking

More info (43%)	*"How well do SOTA models perform on these datasets?"*
Expand coverage (9%)	*"Add applications of contrastive explanations from more papers"*
Clarification (5%)	*"Do all the quantization methods mentioned require additional training?"*

🎨 Stylistic

Condense (10%)

Others (3%)

Remove (13%)

Specific edits (6%)

Reorganize (11%)

LLM as the judge?

GPT-4's Accuracy