



Download our datasets here!

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

Fangyuan Xu¹, Kyle Lo², Luca Soldaini², Bailey Kuehl², Eunsol Choi¹, David Wadden²
¹The University of Texas at Austin, ²Allen Institute for AI

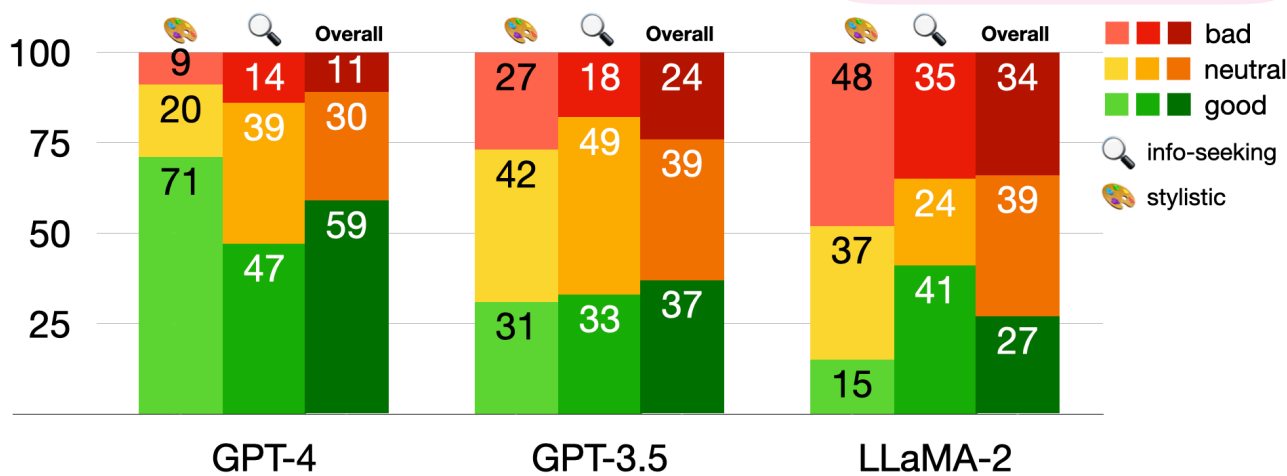
Can LLMs follow user instructions to answer complex research questions grounded in relevant papers?

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



Why does in-context learning work?

Relevant papers



In-context learning works because it allows the model to learn without parameter updates[...]



Describe in detail how in-context learning works and why from a technical standpoint

rating



In-context learning ... When trained on massive text corpora, LMs like GPT-3 implicitly learn to infer



Can you provide more evidence for why in-context learning works spanning multiple papers?

rating



In-context learning ... Multiple works have provided evidence for the effectiveness of in-context learning and have demonstrated its capabilities. One paper ...



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

Remove (13%)

Condense (10%)

Others (3%)

Reorganize (11%)

Specific edits (6%)

LLM as the judge?

GPT-4's Accuracy

