Crafting In-context Examples according
to LMs' Parametric Knowledge

Yoonsang Lee\textsuperscript{1}\textsuperscript{*}, Pranav Atreya\textsuperscript{3}\textsuperscript{*}, Xi Ye\textsuperscript{1}, Eunsol Choi\textsuperscript{1}

\textsuperscript{1}University of Texas at Austin, \textsuperscript{2}Seoul National University, \textsuperscript{3}UC Berkeley

Key Question

- How should we select in-context examples for knowledge-rich tasks?
- Would providing challenging in-context examples (where model don't know the answers) lead LMs to hallucination or better performance?

Finding 1: Known + Unknown > Known > Unknown exemplars

Study design

- First, we label each training example on how much LM knows the answers to the question.

\begin{itemize}
  \item [0.0] \( F_1 \text{EM} \)
  \item [0.5] \( F_1 \text{EM} \)
  \item [1.0] \( F_1 \text{EM} \)
\end{itemize}

- Then, we construct three types (unknown, half-known, known) of in-context example set and compare the performances of using each of them.

Result

- Mixture of known & unknown information yields the best performance.
- Results generalize to other tasks (GSM8K, RTE, and SNLI).

Finding 2: Prompting LM to generate confident answer first leads to performance increase

Study design

- We choose questions with multiple valid answers and study their answer ordering.

\begin{itemize}
  \item Question: Who has scored a hat trick against Spain?
  \item Answer Ordering:
    \begin{tabular}{|c|c|c|}
      \hline
      \textbf{Known} & \textbf{Unknown} \\
      \hline
      Cristiano Ronaldo & David Healy & Gary Lineker \\
      \hline
    \end{tabular}
\end{itemize}

- Greedy decoding: constrained greedy decoding
- Perplexity ordering: compute length-normalized perplexity of each answer and sort the answers.

Result

- Placing \textbf{known} answer at the front performs better than random and reverse counterparts.
- LMs mimic the answer ordering pattern of the in-context examples.

More analysis in the paper!

- If we provide only one answer per example, which answer would lead to better performance?
- Does answer ordering impact the number of generated answers?
- How does in-context example set constructed with one LM impact the generation of another LM?