Crafting In-context Examples according to LMs' Parametric Knowledge

Yoonsang Lee[1][2]*, Pranav Atreya[3]*, Xi Ye[1], Eunsol Choi[1]

[1]University of Texas at Austin, [2]Seoul National University, [3]UC Berkeley
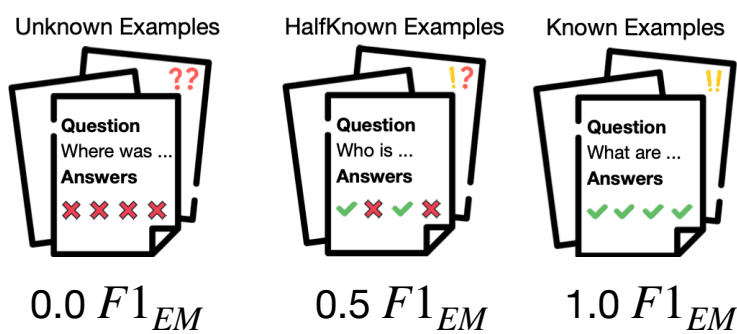
Key Question

- 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?

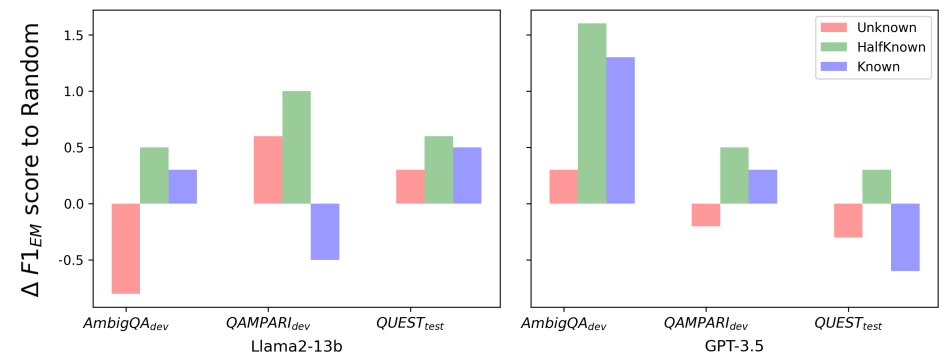## Finding 1: Known + Unknown > Known > Unknown exemplars

Study design

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

Unknown Examples — Question Where was ... Answers ✗✗✗✗ — $0.0\ F1_{EM}$

HalfKnown Examples — Question Who is ... Answers ✓✓✗ — $0.5\ F1_{EM}$

Known Examples — Question What are ... Answers ✓✓✓ — $1.0\ F1_{EM}$

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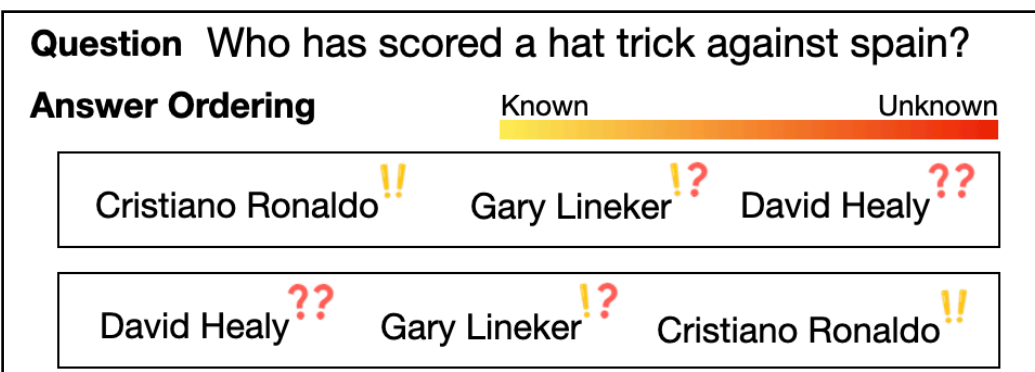
Result

- **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.

Question Who has scored a hat trick against spain?

Answer Ordering   Known ————— Unknown

Cristiano Ronaldo ‼️   Gary Lineker ⁉️   David Healy ⁉️⁉️

David Healy ⁉️⁉️   Gary Lineker ⁉️   Cristiano Ronaldo ‼️

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

Result
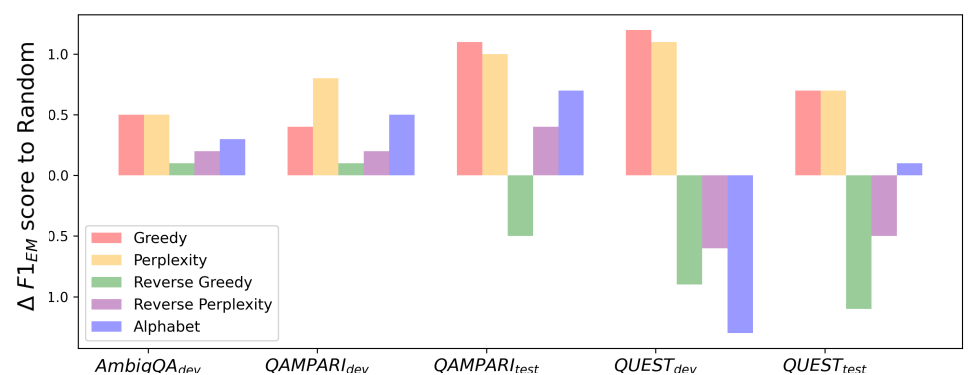
- Placing **known answer at the front** performs better than random and reverse counterparts.
- LMs mimic the answer ordering pattern of the in-context examples.

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?