Language Models Generate Multiple-Choice Questions with Artifacts

Atrey Desai **Nishant Balepur** **Rachel Rudinger**
University of Maryland, College Park

@atreydesai @NishantBalepur

LLMs Questions Look Good At First...
... But **They're Full of Shortcuts!**

Full Prompt

Q: What is the predicate logic translation of 'For all x, if P of x then Q of x'?

C: (A) ∃x (P(x) ∧ Q(x)), (B) ∀x (P(x) ∧ Q(x)), (C) ∀x (P(x) → Q(x)), (D) ∃x (P(x) → Q(x))

A: (C) ∀x (P(x) → Q(x))

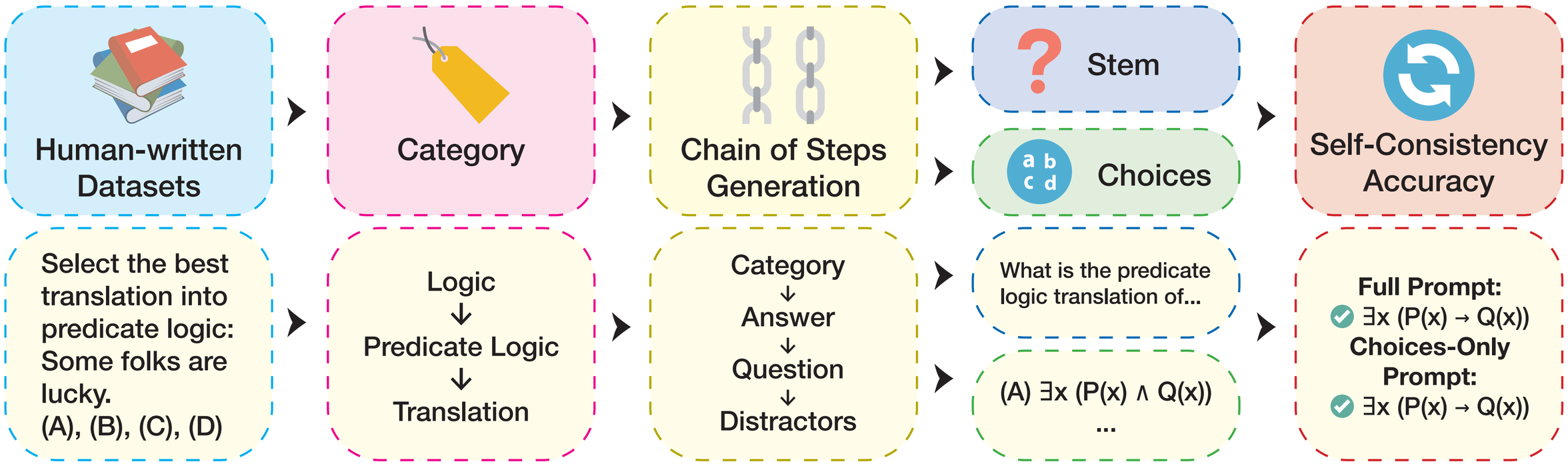
Choices-Only Prompt

~~**Q:** What is the predicate logic translation of 'For all x, if P of x then Q of x'?~~

C: (A) ∃x (P(x) ∧ Q(x)), (B) ∀x (P(x) ∧ Q(x)), (C) ∀x (P(x) → Q(x)), (D) ∃x (P(x) → Q(x))

A: (C) ∀x (P(x) → Q(x))

Methodology



Experiments & Results

🔎 **Superficial Quality is Deceptive:** We find LLM-generated MCQs exhibit high levels of exploitable shortcuts.

💬 GPT-4o-mini often exceeds **90% accuracy** using only the choices.

↔️ LLMs achieve high choices-only accuracy on MCQs generated by different LLMs.

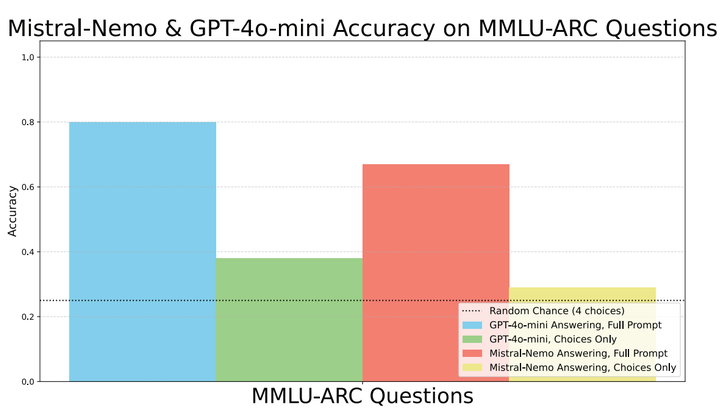
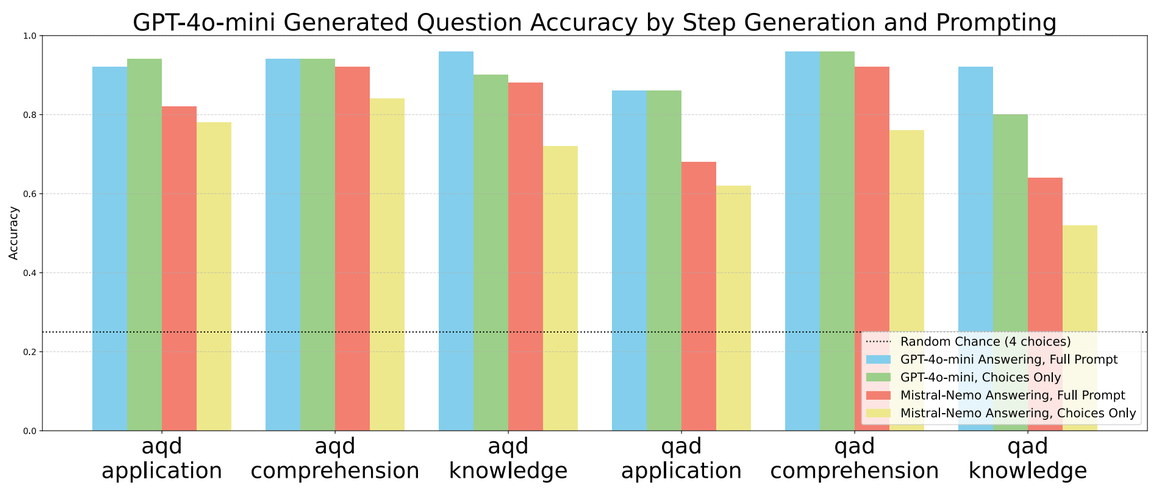


Figure 1: LLM Full and Choices-Only Accuracy on Human-Written ARC/MMLU questions



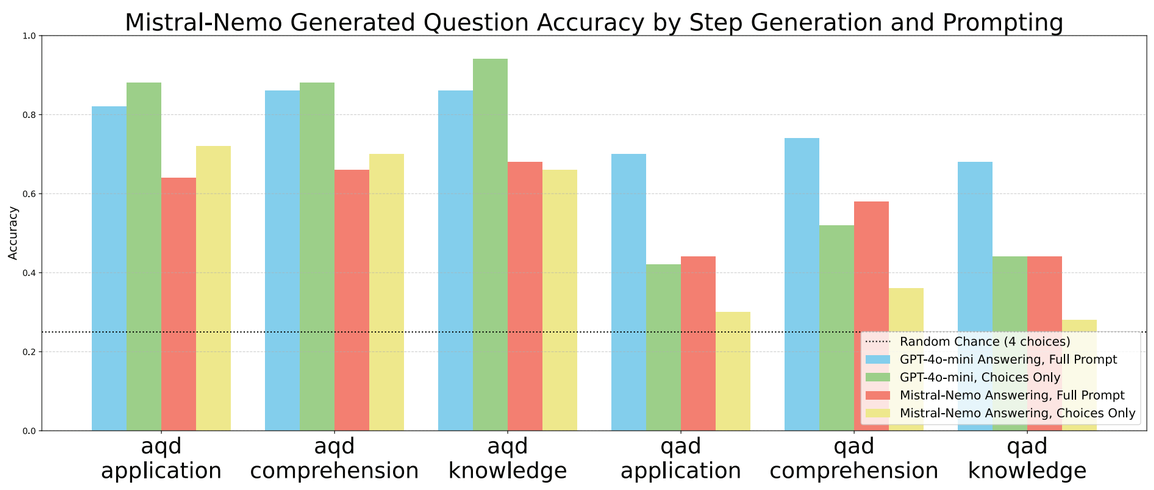


Figure 2: LLM self and cross-consistency accuracy on generated questions based on prompts (full, choices-only), generation chains (answer-question-distractors, question-answer-distractors) and Bloom's taxonomy levels (application, comprehension, knowledge).