Demystifying Scientific Problem-Solving in LLMs by Probing Knowledge and Reasoning

Alan Li* Yixin Liu* Arpan Sarkar Doug Downey Arman Cohan

Introduction

- Scientific tasks demand **domain knowledge** *and* **multi-step reasoning**
- Existing benchmarks probe only a **narrow slice** of skill space:
  - Subject coverage (e.g. GPQA - {phys, chem, bio}, LabBench - {bio})
  - MCQ only - {GPQA, MMLU-Pro, LabBench, …}

How do we evaluate scientific reasoning comprehensively?

How do reasoning and knowledge interact in science problem-solving?

SciREAS: A new benchmark for scientific reasoning

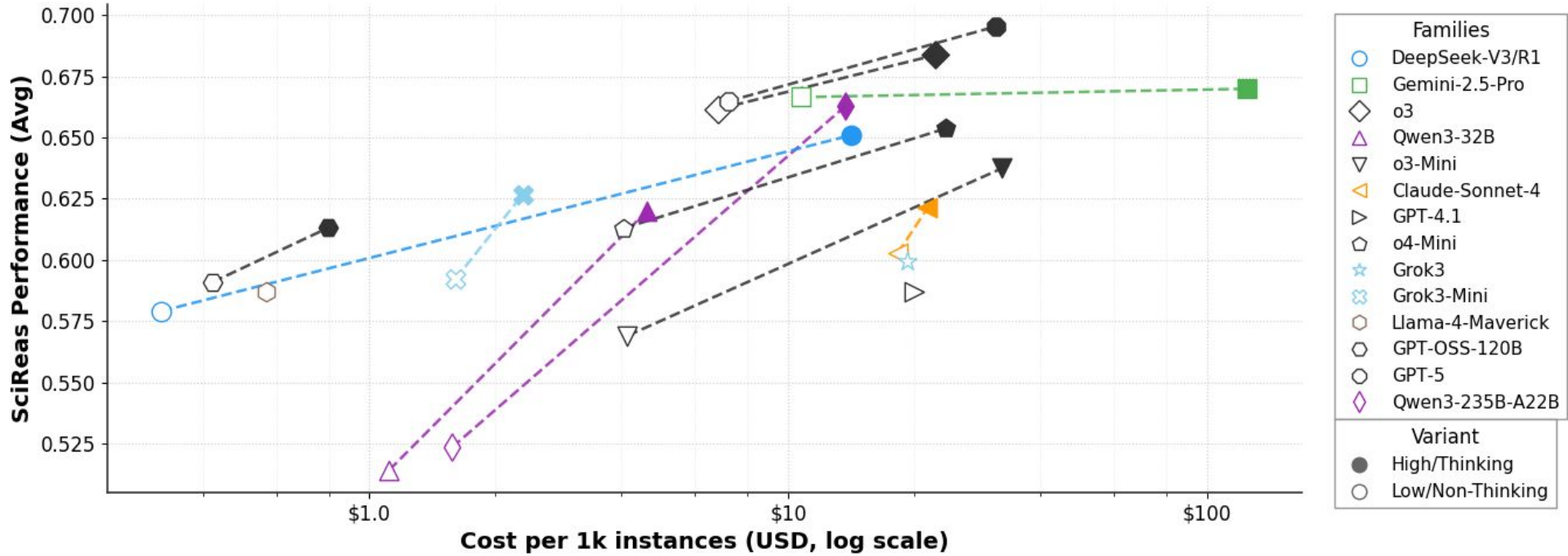
Existing scientific benchmarks: Domain-siloed, fact-heavy, overlook reasoning bottlenecks.

We manually filter existing scientific benchmarks and focus on **reasoning**

(10 datasets)

Sampling to cut the evaluation costs in half, without reducing confidence

 SciREAS Benchmark!


SciReas Performance (Avg)
0.700
0.675
0.650
0.625
0.600
0.575
0.550
0.525
$1.0
$10
$100
Cost per 1k instances (USD, log scale)
Families
DeepSeek-V3/R1
Gemini-2.5-Pro
o3
Qwen3-32B
o3-Mini
Claude-Sonnet-4
GPT-4.1
o4-Mini
Grok3
Grok3-Mini
Llama-4-Maverick
GPT-OSS-120B
GPT-5
Qwen3-235B-A22B
Variant
High/Thinking
Low/Non-Thinking

Why holistic evaluation is important?

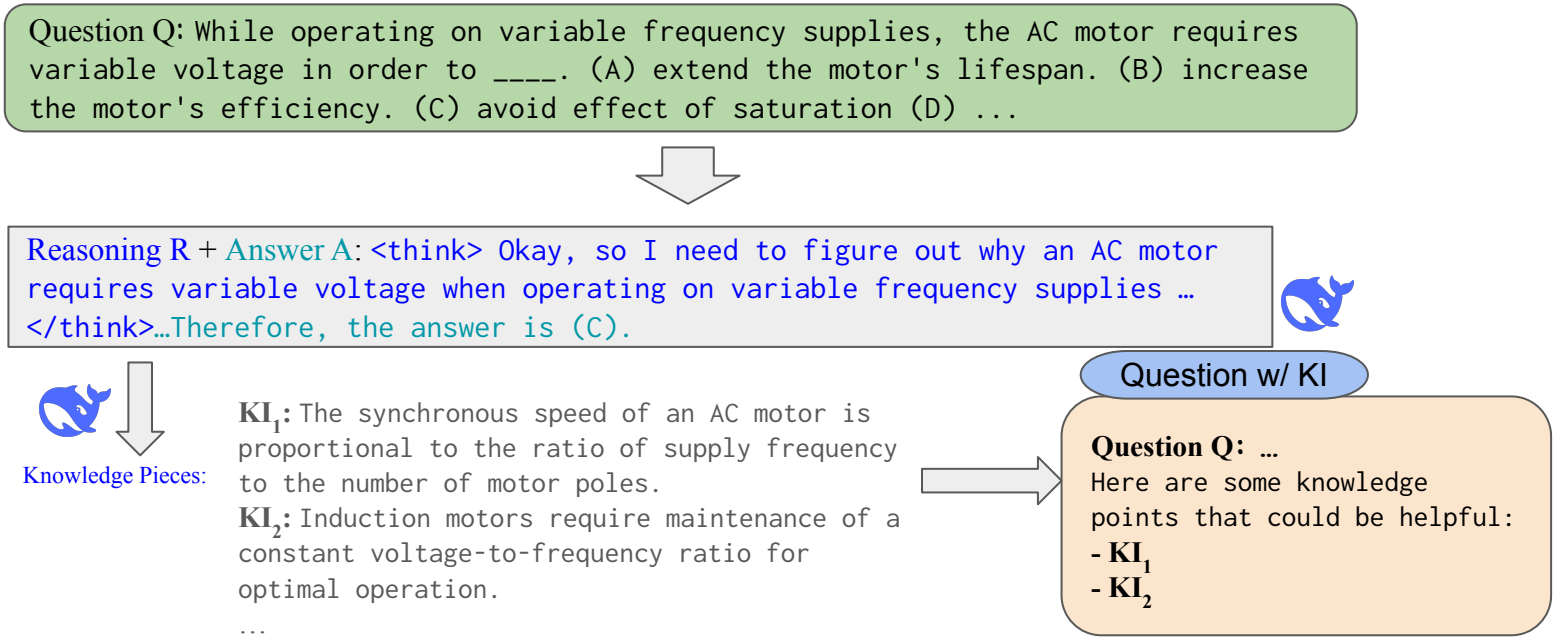
Performance on subsets do not necessarily correlate

Individual models may have been tuned towards different benchmarks

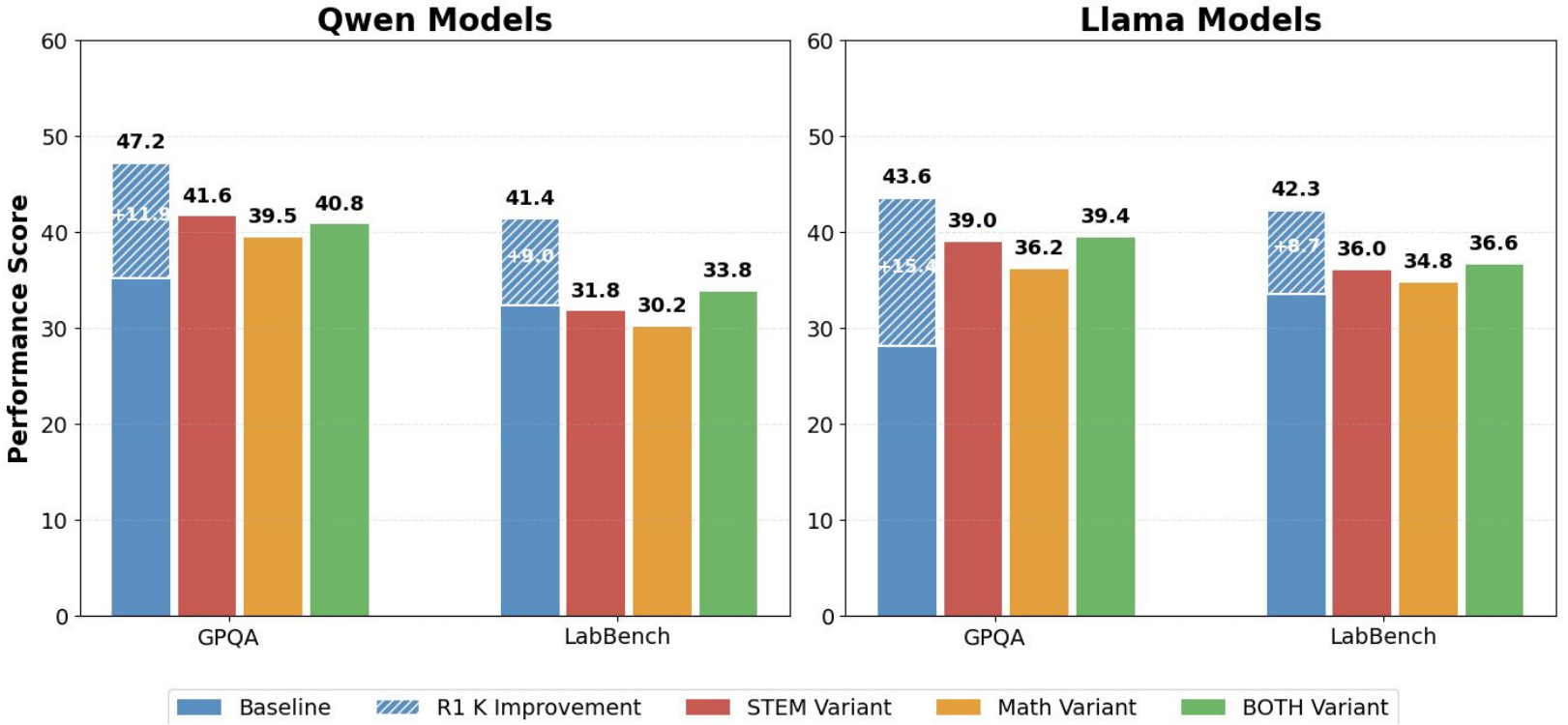
How does the interplay between **knowledge** and **reasoning** shape scientific problem-solving?

KRUX: Knowledge and Reasoning Utilization Exam

1. Collect reasoning traces using a strong reasoning model
2. Extract **knowledge ingredients *K*** from reasoning traces **(KI)**
3. Augment testing instances with *KI* in the prompt

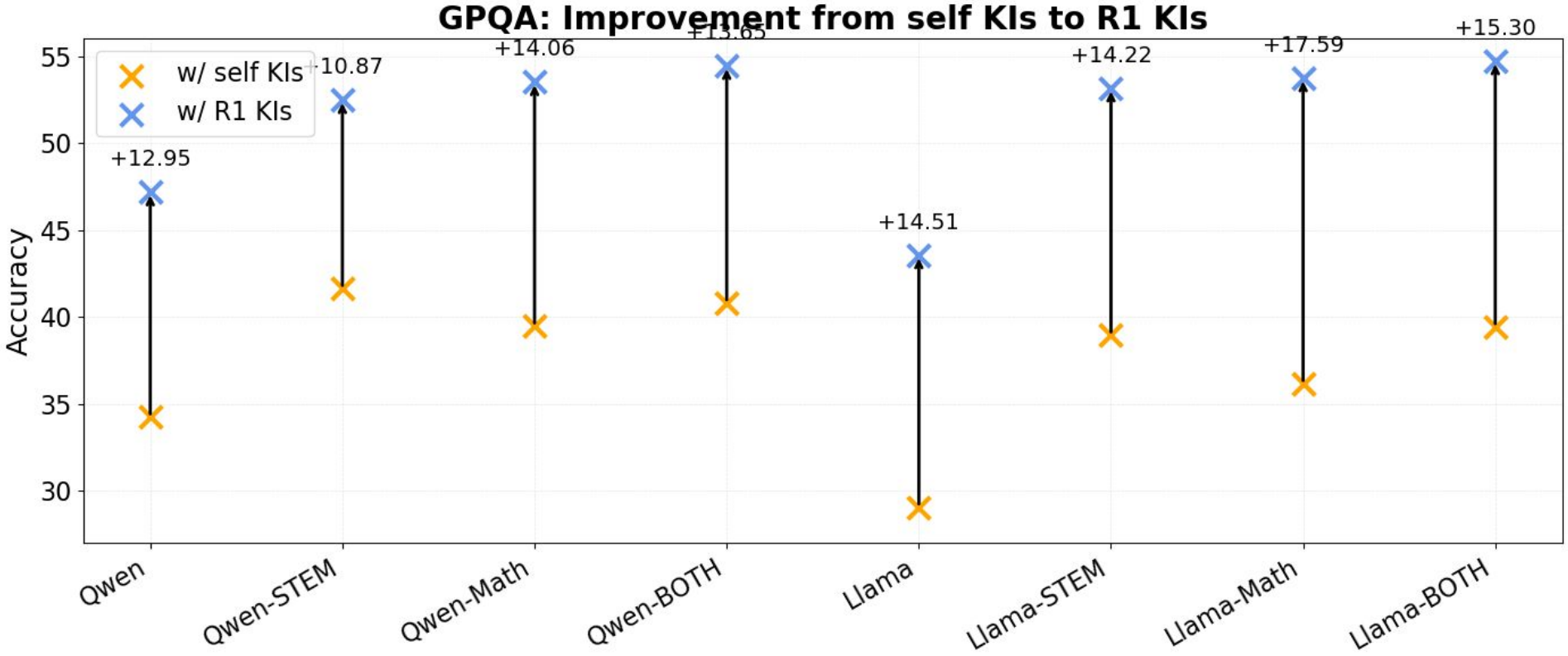


RQ1: Knowledge Bottleneck



Missing knowledge hinders model performance

RQ2: Additive Gains



KIs extracted from different post-trained models and supplied to the prompt

RQ3: how different KIs affect the same base model

Base	Setup	GPQA	MMLU-Pro*
Qwen	w/ Qwen KIs	34.24 ± 0.93	59.03 ± 0.34
	w/ Qwen-Math KIs	**36.93 ± 1.75**	**63.66 ± 0.45**
Llama	w/ Llama KIs	29.06 ± 1.44	47.73 ± 0.89
	w/ Llama-Math KIs	**29.69 ± 1.72**	**53.91 ± 0.94**

Findings: Reasoning-focused training on math tends to help the model in surfacing the knowledge better.

Takeaways

Need for comprehensive benchmarking in scientific problem solving

Models could be bottlenecked by lack of specialized knowledge

Reasoning models consistently prevail with the use of context

Reasoning-focused post-training helps models in surfacing and utilizing knowledge

Thanks!