How Long Reasoning Chain Influence LLMs' Judgement of Answer Factuality

Minzhu Tu[1,2,4], Shiyu Ni[1,2,3], and Keping Bi[1,2,3]

1. State Key Lab of AI Safety
2. Institute of Computing Technology, Chinese Academy of Sciences
3. University of Chinese Academy of Sciences
4. Beijing University of Posts and Telecommunications

About the Authors

Minzhu Tu

Undergraduate student
BUPT

Shiyu Ni

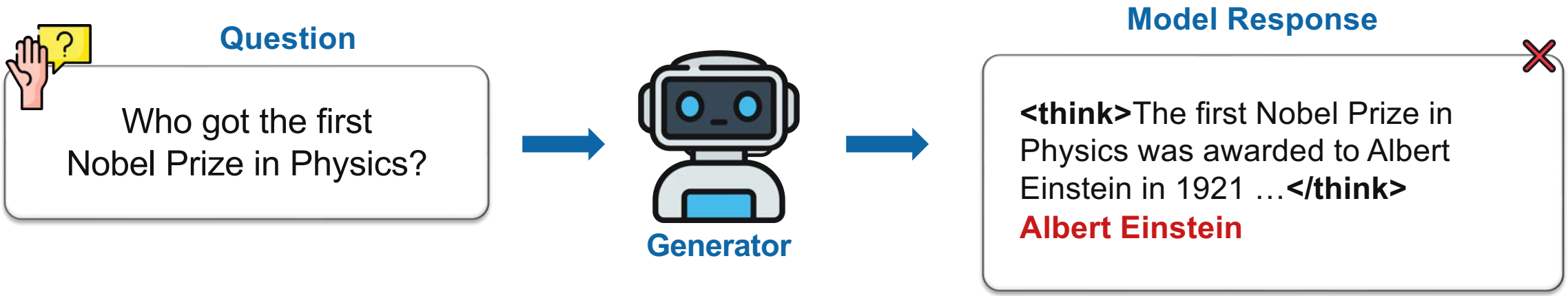
PhD student
ICT,CAS

Keping Bi

Faculty
ICT,CAS

Reasoning Has Become Part of the Model Response

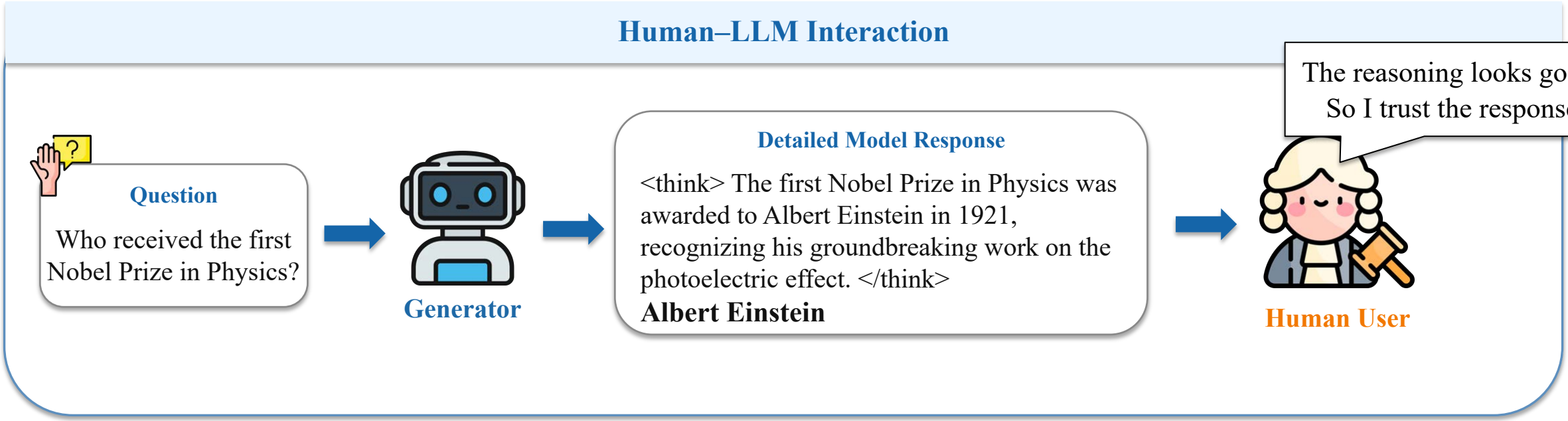
- Modern LLMs increasingly generate a **reasoning process** before providing the final answer.
- Reasoning can improve performance, especially on complex tasks.



Beyond improving performance, what other role does reasoning play?

Detailed Reasoning Makes Answers Easier to Trust

- Reasoning also provides explanations for its answers.
- When we interact with an LLM, a **detailed reasoning process** may make the answer feel more transparent and reliable.



Detailed reasoning can increase users' trust in the response.

But Fluent Reasoning Can Still Be Wrong

- LLMs can hallucinate, generating **fluent and coherent** but **factually incorrect** responses.

What the model says

<think> The first Nobel Prize in Physics was awarded to **Albert Einstein in 1921**, recognizing his groundbreaking work on the photoelectric effect. </think>
Albert Einstein

What is actually true

Both the reasoning and the answer are factually incorrect.

Ground-truth Answer

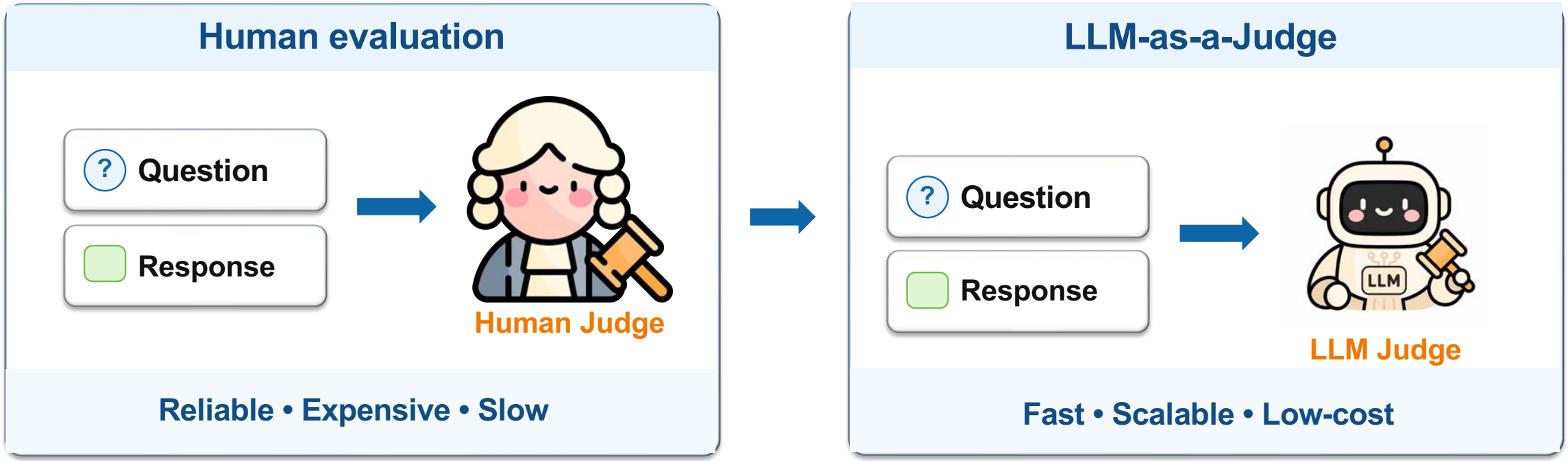
Wilhelm Conrad Röntgen received the first Nobel Prize in Physics in **1901**.

Plausible reasoning ≠ correct answer

Fluent reasoning does not imply correct answers.

From Human Evaluation to LLM-as-a-Judge

- Reliable evaluation is fundamental to **understanding the capabilities** of AI models and **guiding their future development**
- Human judge is **accurate** but costly, time-consuming, and **difficult to scale**
- LLMs are increasingly used as scalable surrogates for human judges



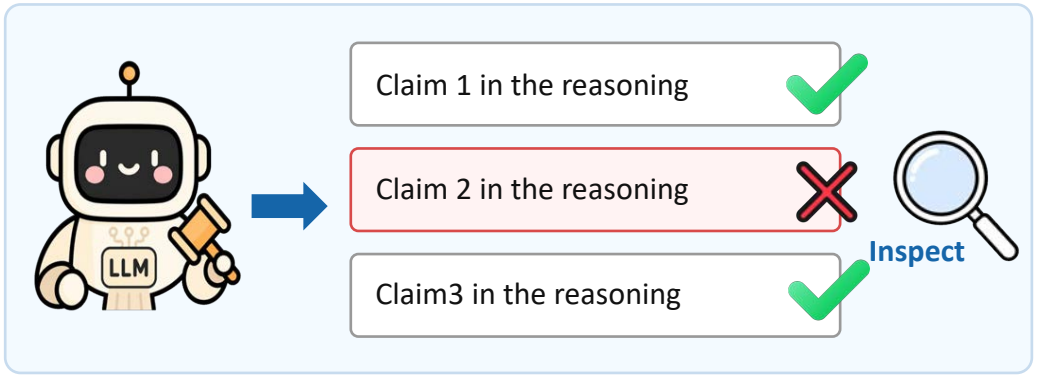
Are LLM judges influenced by reasoning like humans?

Reasoning May Be a Double-Edged Sword

- Reasoning can provide **more evidence for verification**
- But it can also make a response appear **more trustworthy than it actually is.**

Potential benefit

- Reveals intermediate claims that can be inspected
- May expose errors even when the final answer looks plausible



Potential risk

- Fluent reasoning can sound persuasive even when it is wrong.
- A judge may trust surface plausibility rather than factual correctness.

The explanation sounds good, so the answer is probably correct.

How does reasoning content affect LLM judges, and what determines its effect?

Our Hypothesis: Judge Strength Matters

- Strong judges may use reasoning for a more informed assessment
- Weak judges can be misled by fluent but flawed explanations

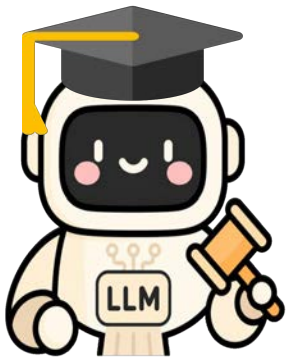
<think>The first Nobel Prize in Physics was awarded to Albert Einstein in **1921** …**</think>** **Albert Einstein**

Weak judge

- Limited relevant knowledge
- More likely to rely on fluency and surface plausibility

The reasoning looks good, so the answer is probably correct

Strong judge



- More relevant knowledge
- Better able to inspect intermediate claims

The first Nobel Prize in Physics was awarded in **1901**, so the answer is wrong

The effect of reasoning may depend on judge strength.

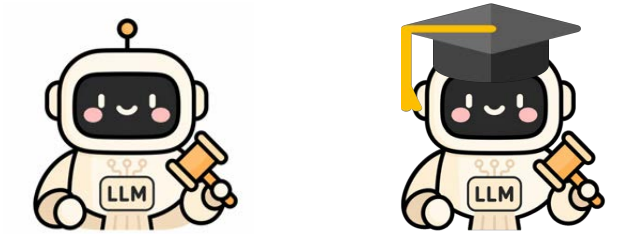
Research Questions

RQ1

Does exposing reasoning affect factuality judgment?

1

RQ2



Weak Judge Strong Judge

Does the effect depend on judge strength?

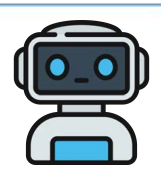
2

RQ3

Which properties of reasoning matter—fluency? factuality?

3

We study this across multiple generators, weak and strong judges, and both factual QA and math tasks.



Generators

Judges

Tasks

Study Design: Judging without Ground Truth

We systematically vary generator, judge, task, and input condition.

Generators

Representative models

- **Qwen3 series**
- **DeepSeek-V3.2**

Judges

Classify by model size and QA performance

weak: Qwen3 series; Llama3 series ; GLM4 series

Strong: DeepSeek-V3.1; GPT-4o; Claude-sonnet-4.5

Datasets

Include reasoning-intensive and non-intensive tasks

Factual QA

- NQ: one-hop
- HotpotQA: multi-hop

Mathematics

- GSM8K
- MATH500

Metrics

- **Accuracy**: % of responses actually correct
- **Pass Rate**: % judged as correct

Gap = Pass Rate − Accuracy

- **Alignment**: Gap = 0
- **Overconfidence**: Gap > 0
- **Conservativeness**: Gap < 0

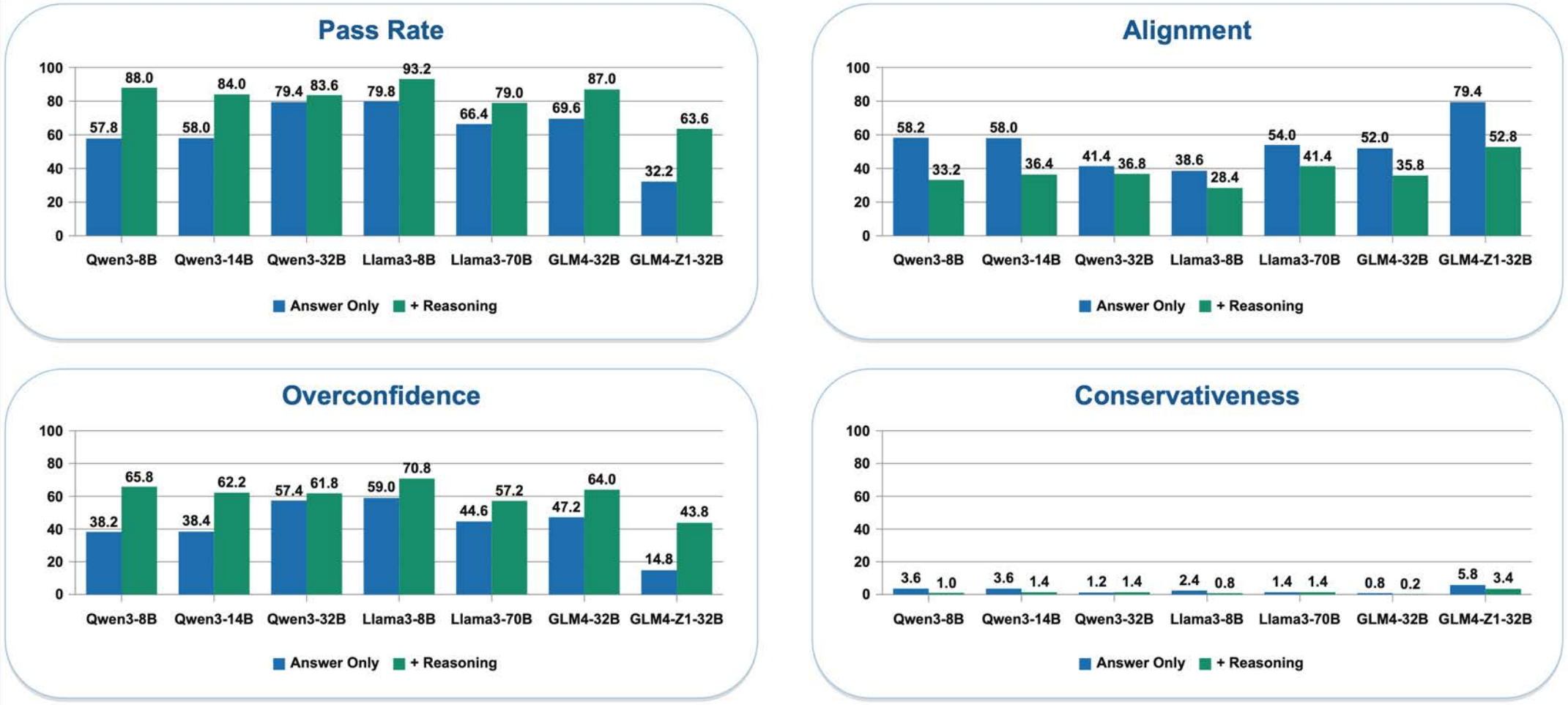
Answer-only judging ⟷ **Reasoning-exposed judging**

Core comparison: answer-only judging vs. reasoning-exposed judging.

Reasoning Inflates Acceptance for Weak Judges

Generator: Qwen3-8B

Dataset: NQ

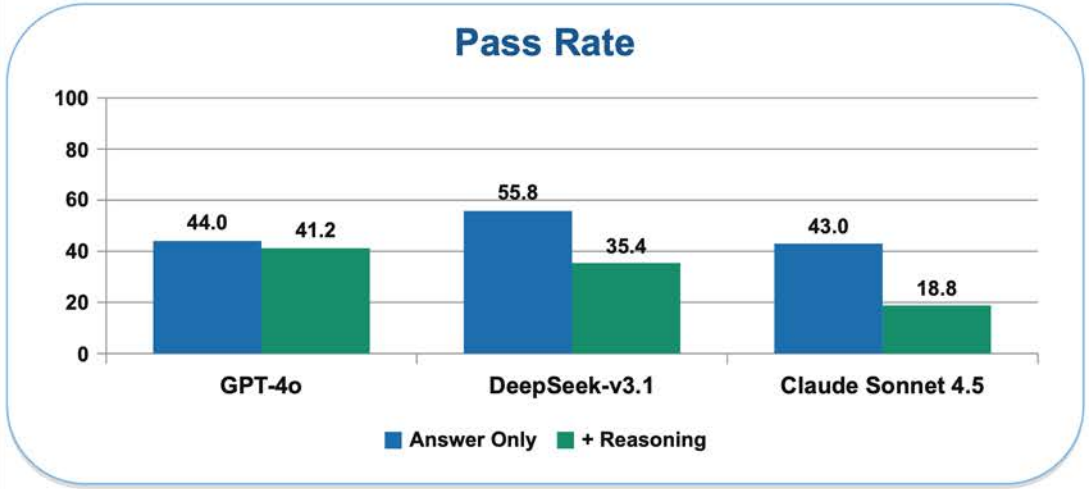


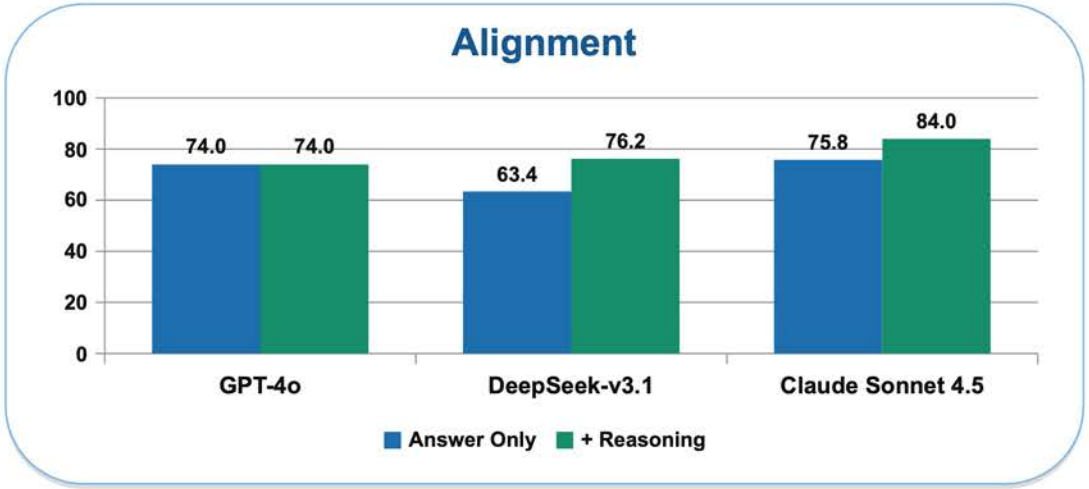
Weak judges are consistently misled by the reasoning process.

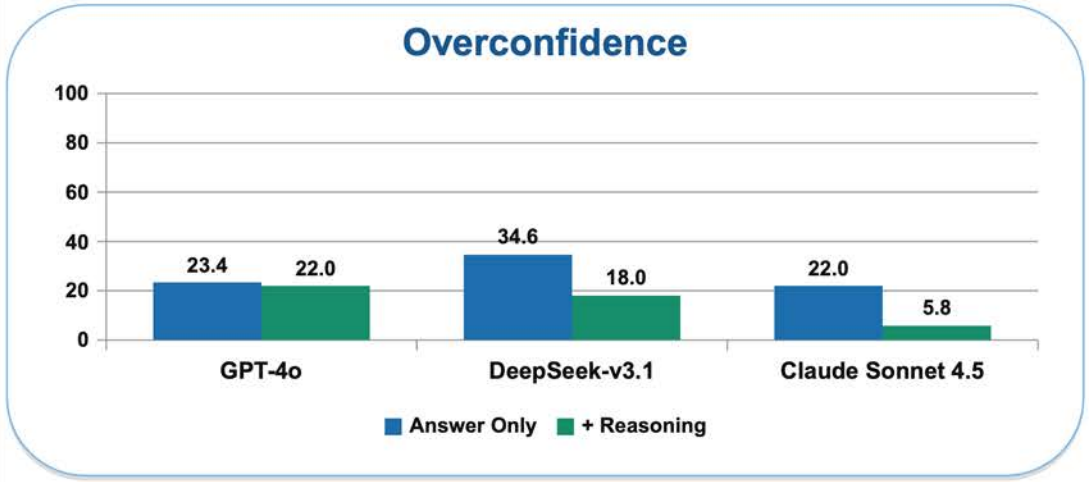
Strong Judges Use Reasoning More Selectively

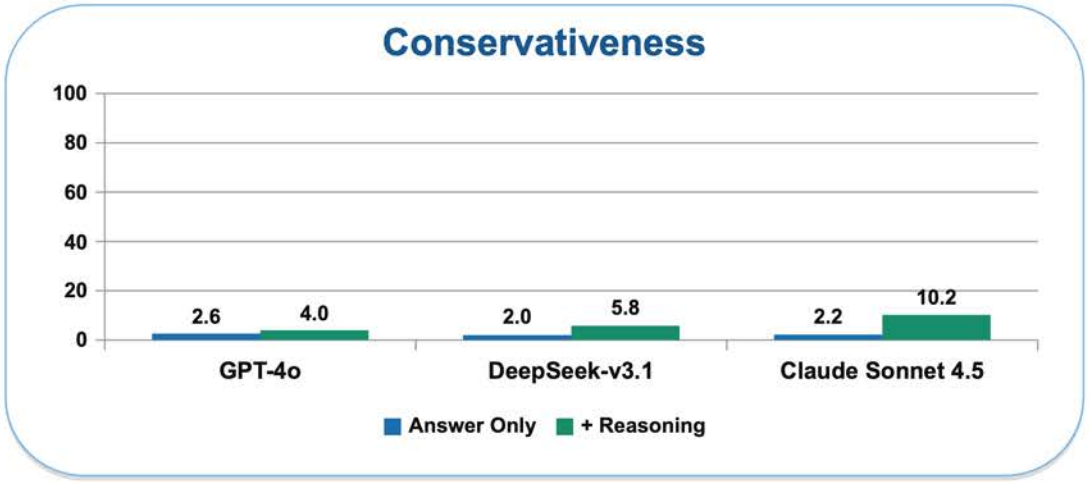
Generator: Qwen3-8B

Dataset: NQ









↗ **Strong judges can sometimes benefit from the reasoning process.**

The Real Case

Generator: Qwen3-8B

Misleading Weak Judges

Question: Who is the girl in Green Day 21 Guns?

Model Answer: Amanda Kessel **Ground Truth:** Lisa Stelly

Key reasoning snippets (from <think>)

- … maybe **a fan who was part of the video shoot**.
- … I believe her name is **Amanda Kessel**.
- … the user probably wants the name of the girl in the video, so the answer is **Amanda Kessel**.

Judges' outcome (✓ = judge thinks the answer is correct)

	Weak Models			Strong Models		
w/o Think	✗	✗	✗	✗	✗	✗
w/ Think	✓	✓	✓	✗	✗	✗

The chain **sounds plausible** but lacks reliable support, leading weak judges to be misled.

Dataset: NQ

Helpful for Strong Judges

Question: When did Toyota first come to the United States?

Model Answer: 1958 **Ground Truth:** 1957

Key reasoning snippets (from <think>)

- … The first Toyota car sold in the US was in **1958**.
- … Some sources mention **1957 for the initial import**.
- … However, the commonly accepted answer is **1958**.

Judges' outcome (✓ = judge thinks the answer is correct)

	Weak Models			Strong Models		
w/o Think	✓	✓	✓	✓	✓	✓
w/ Think	✓	✓	✓	✗	✗	✗

The chain **contains conflicting evidence**, which strong judges leverage to detect errors.

Reasoning Effects Persist on MATH

Generator: Qwen3-8B **Dataset: MATH500**

- Weak judges consistently raise the pass rate
- Strong judges are relatively robust to misleading reasoning, but their alignment does not improve.

↗ **Similar trends to those observed on factual QA datasets.**

Strong Judges Can Also Be Misled

Generator: DeepSeek-V3.1 **Dataset: NQ**

With DeepSeek's reasoning chain, all judges show higher pass rates but lower alignment.

Fluency and Factuality Are Important Factors

We keep the final answer fixed and manipulate only the reasoning chain.

Experimental Setting Five judge-input conditions disentangle answer-only judging, reasoning exposure, fluency disruption, and factuality manipulation.

1 Baselines

Vanilla (w/o Think): judge sees the final answer only.

Vanilla (w/ Think): judge sees the original reasoning plus the final answer.

2 Fluency disruption

Basic-All: insert 4 factual but irrelevant sentences into the reasoning.

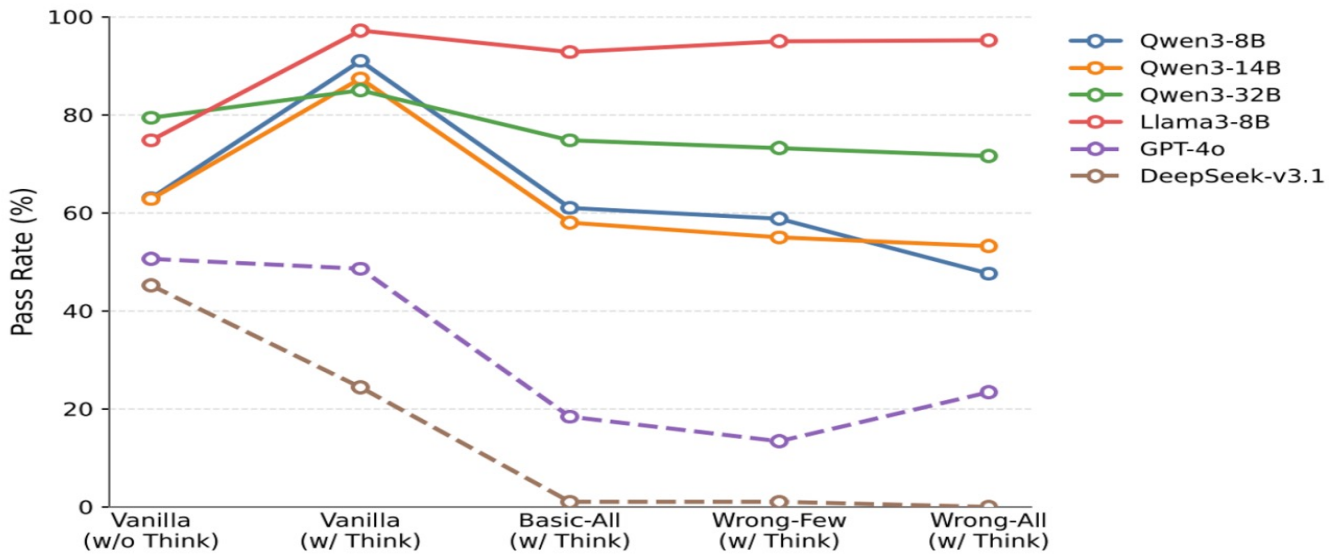
Example: *"The sun rises in the east."*

Tests whether broken fluency/relevance changes the verdict.

3 Factuality manipulation

Wrong-Few: replace 1 inserted sentence with a counterfactual one.
Wrong-All: replace all 4 inserted sentences with counterfactual ones.
Example: *"The sun rises in the west."*

Generator: Qwen3-8B · Dataset: NQ



What the Manipulations Reveal

1 Fluency matters
Original → Basic-All

Factual but irrelevant insertions lower pass rates for all judges.

2 Factuality matters
Basic-All → Wrong-Few/All

Counterfactual content lowers pass rates for most judges.

3 Sensitivity is limited
Basic-All → Wrong-Few; Wrong-Few → Wrong-All

Weak judges react weakly to one false claim, and more false claims add little drop.

Fluency and factuality are two key factors shaping how reasoning affects LLM judges.

Key Takeaways

RQ1 Reasoning exposure often increases acceptance, but alignment does not consistently improve.

RQ2 Weak Judges

Easily persuaded

Reasoning often raises pass rates while lowering alignment.

RQ2 Strong Judges

Selective but not immune

They sometimes use reasoning effectively, but can still be misled by strong generators.

RQ3 Reasoning Content

Fluency + factuality matter

Both how reasoning is presented and what it contains shape the judgment.

Reasoning is not neutral evidence: its effect depends on both the judge and the reasoning itself.

Discussion

When Does Reasoning Help—and When Can Judges Verify It?

Reasoning is useful only if it provides usable signals and the judge can effectively use them.

When can reasoning help?

What signals are useful, which ones matter, and how should they be used?

1 Answer–Reasoning Inconsistency

If reasoning conflicts with the final answer, can it still provide useful evidence?

2 Evidence Usefulness

When reasoning and answer agree, which claims are actually informative? Does usefulness depend on their impact on the final answer?

3 Mixed Evidence

When reasoning contains both supporting and contradicting evidence, how should the judge combine them?

When can judges use it?

When can a judge use reasoning signals effectively—and when does it fail?

1 Verification without Solving

Can a judge detect a flawed claim without independently knowing the correct answer?

2 Generator–Judge Capability Gap

When does the generator's reasoning exceed the judge's ability to verify it?

3 Selective Reliance

When should a judge use, ignore, or seek external verification for the reasoning?

Core challenge: knowing when reasoning helps, and when judges are capable of verifying it.

Research on Factuality Judgement

When Do LLMs Need Retrieval Augmentation **ACL24**

Fully Exploiting LLM Internal States **ACL25**

Do LVLMs Know What They Know? **EMNLP25**

Annotation-Efficient Universal Honesty Alignment **ICLR26**

Reasoning Affects LLM-as-a-Judge Evaluation **ACL26**

Honesty Alignment under Multi-Answer Questions **Arxiv26**

A Perspective from Knowledge Popularity **Arxiv26**

Q&A

TIME Group

Contact us: Trustworthy InforMation accEss (TIME) https://stay-hungry-time.github.io/