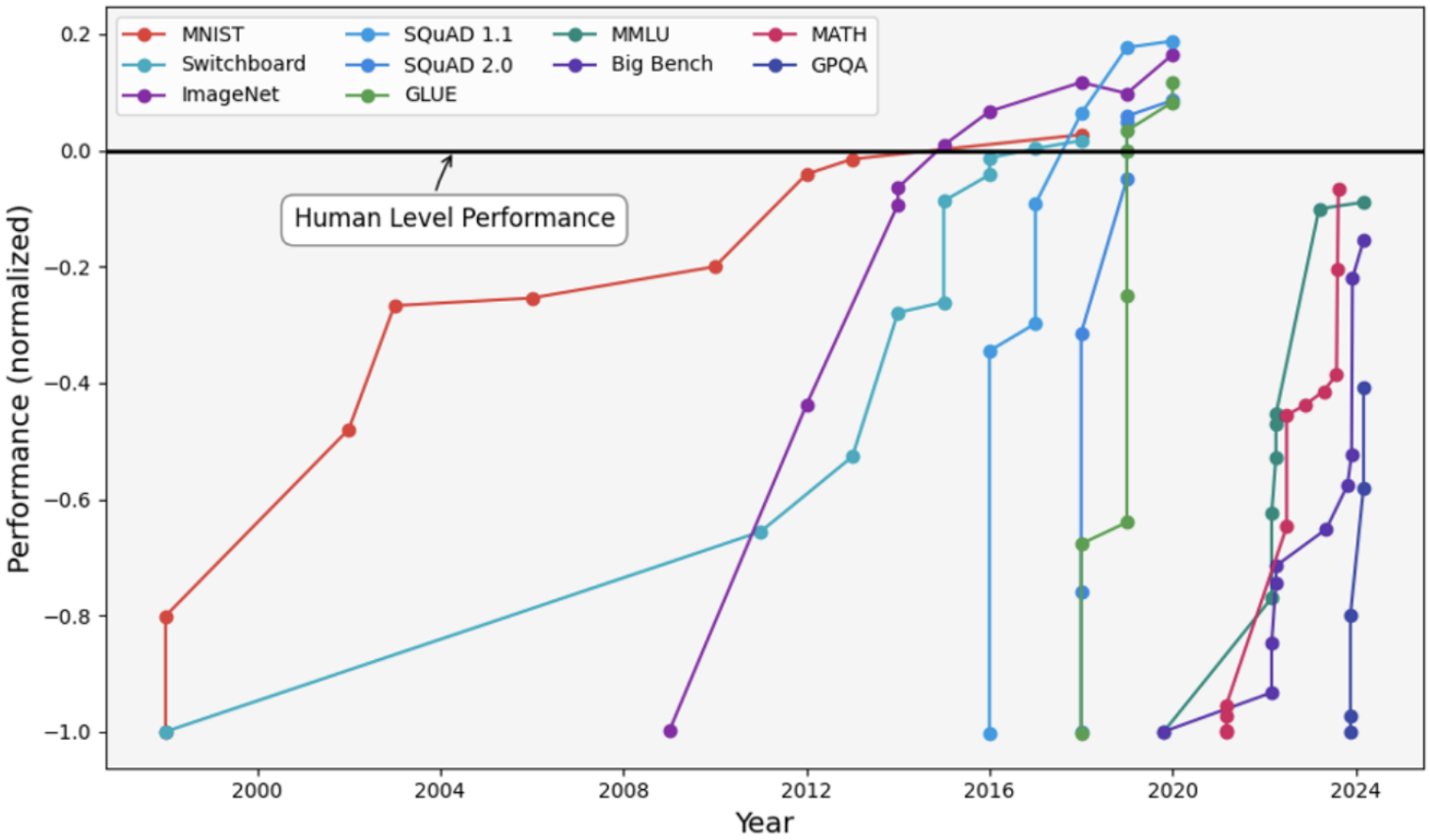
Large Language Model Debate for Scientific Decisions: A Particle Physics Prototype

Nayara Fonseca

AI Progress: From 2000 to 2024



Performance of AI models on various **benchmarks**: computer vision (MNIST, ImageNet), speech recognition (Switchboard), natural language understanding (SQuAD 1.1, MNLU, GLUE), general language model evaluation (MMLU, Big Bench, and GPQA), and mathematical reasoning (MATH). Many models surpass human-level performance (black solid line) by 2024. Kiela, D., Thrush, T., Ethayarajh, K., & Singh, A. (2023) 'Plotting Progress in AI'

https://yoshuabengio.org/2024/07/09/reasoning-through-arguments-against-taking-ai-safety-seriously/

But …

What if we (humans) do not know the answer?

- How can we evaluate something that is unknown?
- How can we evaluate an open-ended question?

But …

What if we (humans) do not know the answer?

- How can we evaluate something that is unknown?
- How can we evaluate an open-ended question?

AI to guide ''big science" decision

Debate

But …

What if we (humans) do not know the answer?

- How can we evaluate something that is unknown?
- How can we evaluate an open-ended question?

AI to guide ''big science" decision

Debate $\longrightarrow$ AI Debate

AI Debate

[arXiv:2402.06782] [ICML 2024]

Debating with More Persuasive LLMs Leads to More Truthful Answers

Akbir Khan [*1] John Hughes [*2 3] Dan Valentine [*3] Laura Ruis [1] Kshitij Sachan [4 5] Ansh Radhakrishnan [4]
Edward Grefenstette [1] Samuel R. Bowman [4] Tim Rocktäschel [1] Ethan Perez [4 6]

AI Debate

[arXiv:2402.06782] [ICML 2024]

Debating with More Persuasive LLMs Leads to More Truthful Answers

Akbir Khan [*1] **John Hughes** [*2 3] **Dan Valentine** [*3] **Laura Ruis** [1] **Kshitij Sachan** [4 5] **Ansh Radhakrishnan** [4]
Edward Grefenstette [1] **Samuel R. Bowman** [4] **Tim Rocktäschel** [1] **Ethan Perez** [4 6]

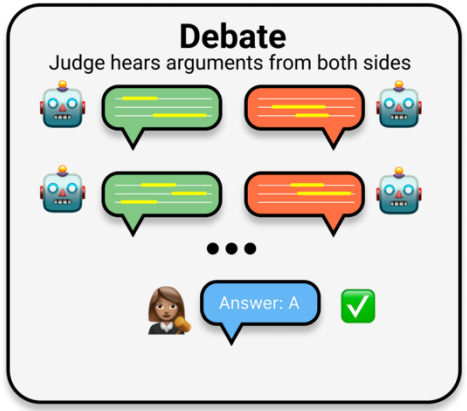
[arXiv:1805.00899]

AI safety via debate

Geoffrey Irving* Paul Christiano Dario Amodei
OpenAI

AI Debate

Khan et al. (ICML 2024)



High-level flow

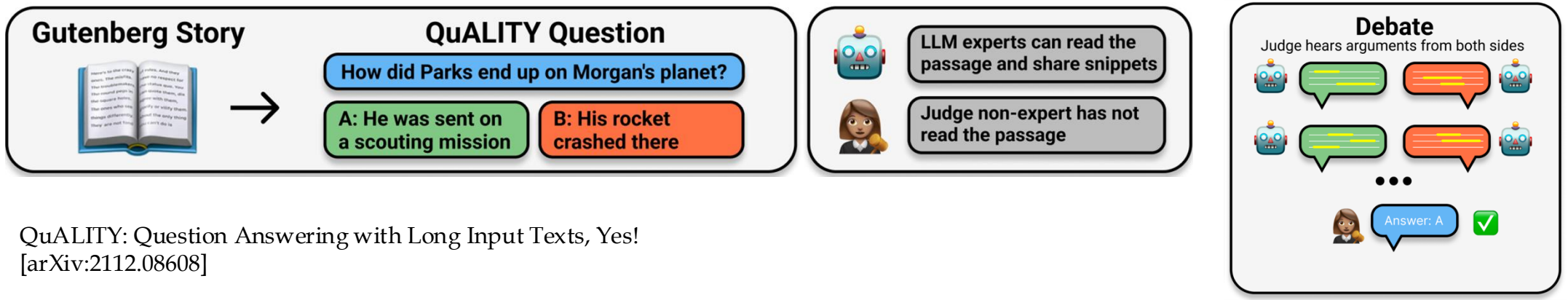
- Given a topic, two LLMs (experts) argue on a topic (e.g., "Pro vs Against" debates)
- Another LLM judge (non-expert) decides who is more persuasive

Scalable method for supervision (later: adapted for the particle-physics case)

- LLM judge is a 'weaker' model
- *Can weaker models assess the correctness of stronger models?*

AI Debate

Debating with More Persuasive LLMs Leads to More Truthful Answers,
Khan et al. (ICML 2024)

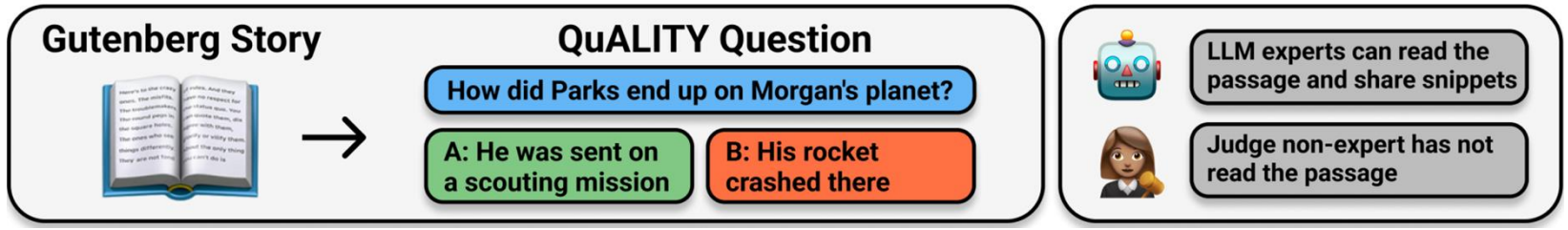


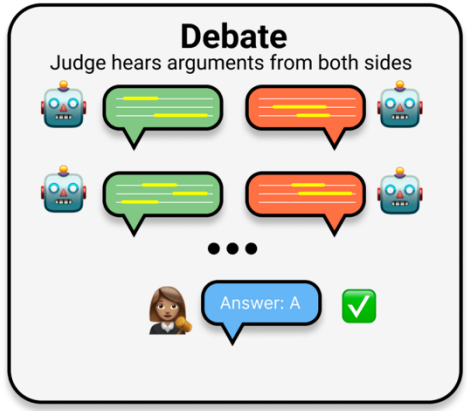
QuALITY: Question Answering with Long Input Texts, Yes!
[arXiv:2112.08608]

- Expert models argue for a specific answer to a comprehension question.
- Weaker (non-expert) judges, who cannot access the underlying text, evaluate the arguments and choose an answer.
- In debate, two experts simultaneously present arguments for a number of rounds.

AI Debate

Debating with More Persuasive LLMs Leads to More Truthful Answers,
Khan et al. (ICML 2024)





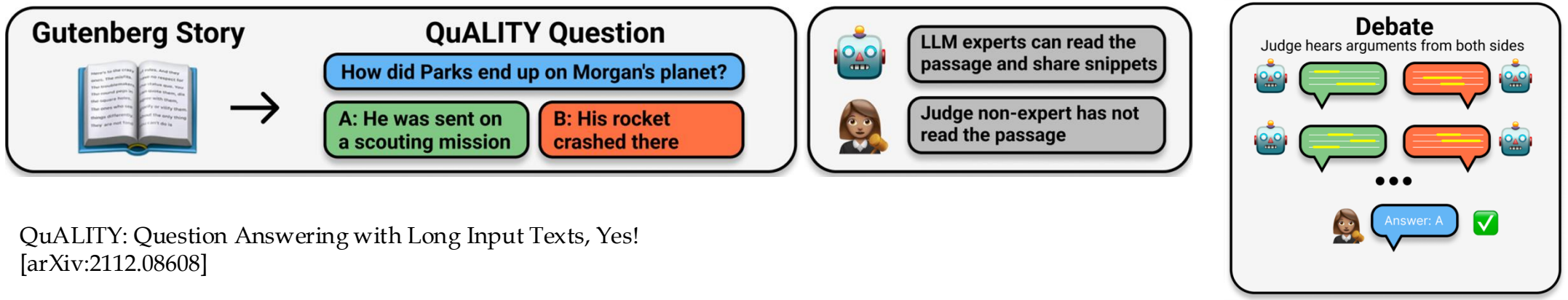
QuALITY: Question Answering with Long Input Texts, Yes!
[arXiv:2112.08608]

Truth: For this dataset this is fact-check

- Recruit 30 human judges via the referral-based annotator platform
- Use texts from the Project Gutenberg science-fiction story subset (approx. 7000 tokens)
- Select the HARD subset, where all annotators chose the correct answer and rated the answer as unambiguous

AI Debate

Debating with More Persuasive LLMs Leads to More Truthful Answers,
Khan et al. (ICML 2024)



QuALITY: Question Answering with Long Input Texts, Yes!
[arXiv:2112.08608]

Truth: For this dataset this is fact-check

- Recruit 30 human judges via the referral-based annotator platform
- Use texts from the Project Gutenberg science-fiction story subset (approx. 7000 tokens)
- Select the HARD subset, where all annotators chose the correct answer and rated the answer as unambiguous

Not available for open-ended science questions

‘What is the debate about?’

The Particle Physics Case ‘What is the debate about?’

March, 2025

The Guardian

News | Opinion | Sport | Culture | Lifestyle

World | US politics | UK | Climate crisis | Middle East | Ukraine | Environment | Science | Global development | Football | Tech | Business | Obituaries

Large Hadron Collider

This article is more than 2 months old

Just a big toy – or key to the universe?
Row over even Larger Hadron Collider

Robin McKie, *Science Editor*

Sat 29 Mar 2025 11.12 GMT

Share

CERN releases report on the feasibility of a possible Future Circular Collider

Released today, a report of a study investigating the project’s feasibility will serve as input for the European Strategy for Particle Physics and be assessed by the CERN Council in the coming months

31 MARCH, 2025

https://home.cern/news/news/accelerators/cern-releases-report-feasibility-possible-future-circular-collider

nature

NEWS FEATURE | 19 March 2025 | Correction 19 March 2025

The biggest machine in science: inside the fight to build the next giant particle collider

The European physics laboratory CERN is planning to build a mega collider by 2070. Critics say the plan could lead to its ruin.

By Davide Castelvecchi

Information | Files

FCC Document

Report number	arXiv:2505.00272 ; CERN-FCC-PHYS-2025-0002
Title	**Future Circular Collider Feasibility Study Report Volume 1 : Physics and Experiments**
Author(s)	Benedikt, M. (CERN) ; Zimmermann, F. (CERN) ; Auchmann, B. (CERN ; PSI, Villigen) ; Bartmann, W. (CERN) ; Giovannozzi, M. (CERN) ; Grojean, C. (DESY ; Humboldt U., Berlin) *Show all 1464 authors*
Publication	2025

The Particle Physics Case ‘What is the debate about?’

- Fundamental physics is at a crossroad, facing urgent decisions (such as investing in a proposed multi-billion particle collider) that will define the field’s future

- Current debates lack effective tools for structured deliberation

Generated with Chat GPT 5 (Thinking), Sep 2025

“Generate a crossroad image with this theme here in this article: https://www.theguardian.com/science/2025/mar/29/the-physics-community-has-never-split-like-this-row-erupts-over-plans-for-new-large-hadron-collider”

The Particle Physics Case Prototype

Pioneer AI-driven debates for science policy:

The Particle Physics Case Prototype

Pioneer AI-driven debates for science policy:

- Open-ended question: we adapt debate evaluation from multi-question settings to an open-ended policy question without ground truth;
- Reduce bias (e.g., setups with per-match randomized summary order, opener alternation, and balanced label-to-stance assignment);
- Measurements: outcomes are quantified (e.g., pooled stance win rates and debater strengths from Elo-like scores).

The Particle Physics Case Prototype

Pioneer AI-driven debates for science policy:

- Open-ended question: we adapt debate evaluation from multi-question settings to an open-ended policy question without ground truth;
- Reduce bias (e.g., setups with per-match randomized summary order, opener alternation, and balanced label-to-stance assignment);
- Measurements: outcomes are quantified (e.g., pooled stance win rates and debater strengths from Elo-like scores).

Long-term: make disagreement measurable in a transparent way that can augment expert deliberation in specialized scientific domains.

The Particle Physics Case Prototype

- LLM debates on Particle Physics Experimental Strategies
- Two LLM agents (GPT-style) argue and an LLM judge decides who wins the debate

```
P5 + FCC (static context) ──► Neutral initial prompt
                              "Given P5 (2023) and FCC (2025), what's the best
                               strategy for HEP in the next decade?"

                    ┌──────────────── diverse openings ────────────────┐
                    │  Strategy-1 Opening A   Strategy-1 Opening B  …  │
                    │  Strategy-2 Opening A   Strategy-2 Opening B  …  │
                    └──────────────────────────────────────────────────┘
                                         │ pairwise
                                         ▼
Debate Matches  (A1 vs B1, A1 vs B2, …) — multi-round back-and-forth
                                         │
                                         ▼
LLM Judge  →  WINNER: Strategy-1 / WINNER: Strategy-2  +  rationale
                                         │
                                         ▼
Quant stats  →  win-rate · average tokens / turn · Elo (BT)
```

AI Debate 🤖⚛️

python 3.10+

Prototype – AI Debate for particle-physics strategy (P5 2023 + FCC 2025, neutral multi-corpus)

This project explores domain-specific decision-making with automated judging and diverse opening arguments.

- Large-language-model **"Strategy 1 vs Strategy 2"** debates over a **combined evidence context**: the US P5 (2023) report and CERN's FCC Feasibility Study (2025).
- Two GPT-style agents argue, an LLM judge decides who's more persuasive (emitting a strict `WINNER: Strategy-1` / `WINNER: Strategy-2` line), and the whole exchange (with token stats + context provenance) is logged to disk.
- Inspired by recent work on **LLM debate protocols** such as Debating with More Persuasive LLMs Leads to More Truthful Answers (Khan et al., 2024).

https://github.com/nayara-focs/ai-debate-p5

PLAN

- Setup, Workflow & Design Choices
- Measurements
- How the matches are organized

Setup

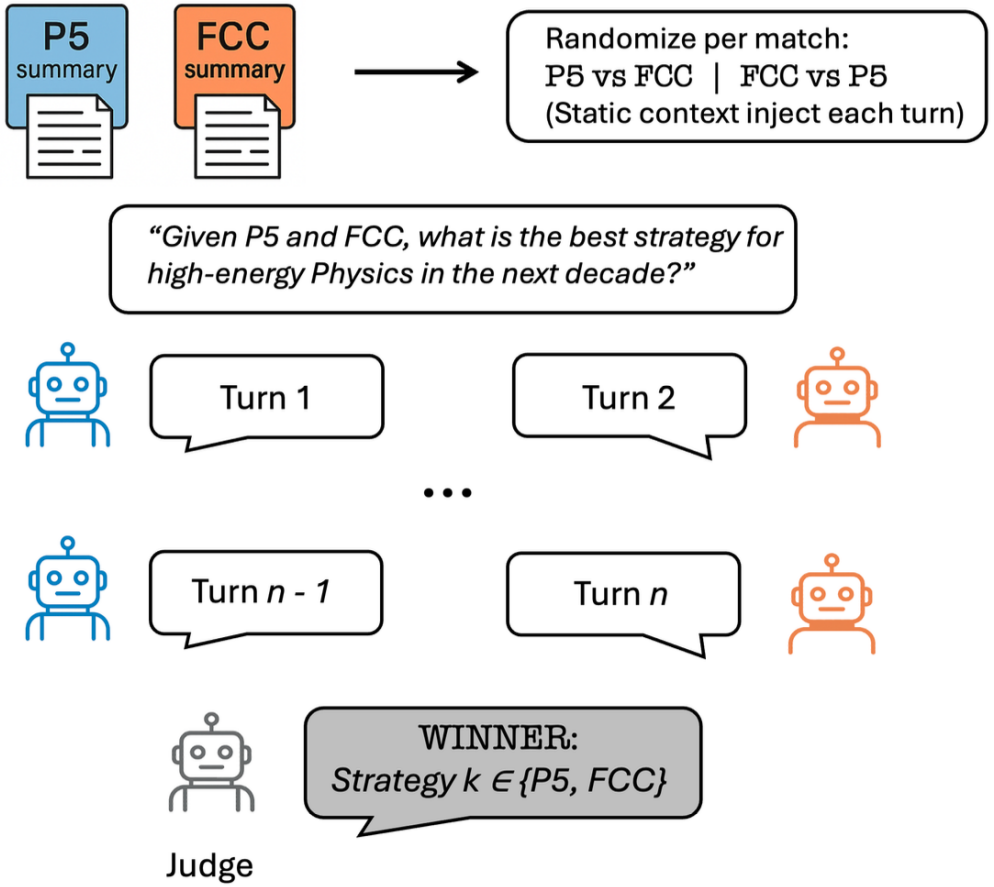


Illustration of our LLM debate setup

Examples:

Particle Physics Project Prioritization Panel (P5)
https://www.usparticlephysics.org/2023-p5-report/

Future Circular Collider (FCC), Vols. 1-3

See, e.g., https://home.cern/news/news/accelerators/cern-releases-report-feasibility-possible-future-circular-collider

Setup

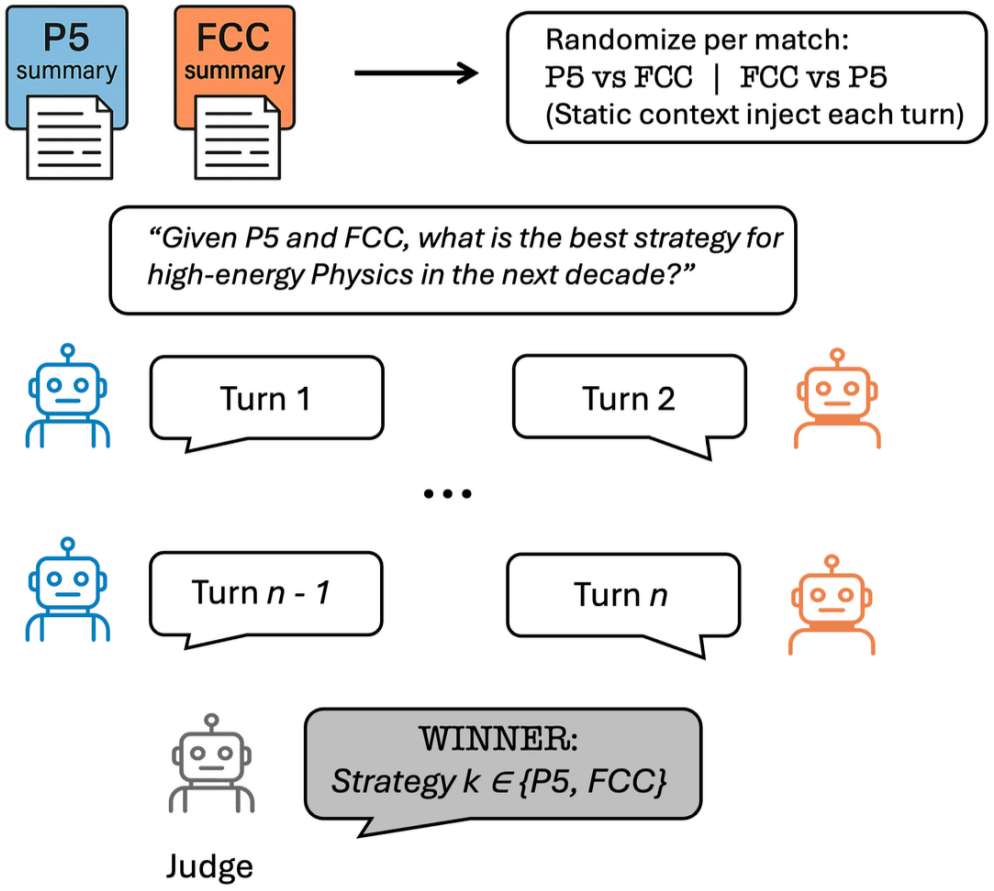


Illustration of our LLM debate setup

Examples:

Particle Physics Project Prioritization Panel (P5)
https://www.usparticlephysics.org/2023-p5-report/

Future Circular Collider (FCC), Vols. 1-3

See, e.g., https://home.cern/news/news/accelerators/cern-releases-report-feasibility-possible-future-circular-collider

Disclaimer

Setup

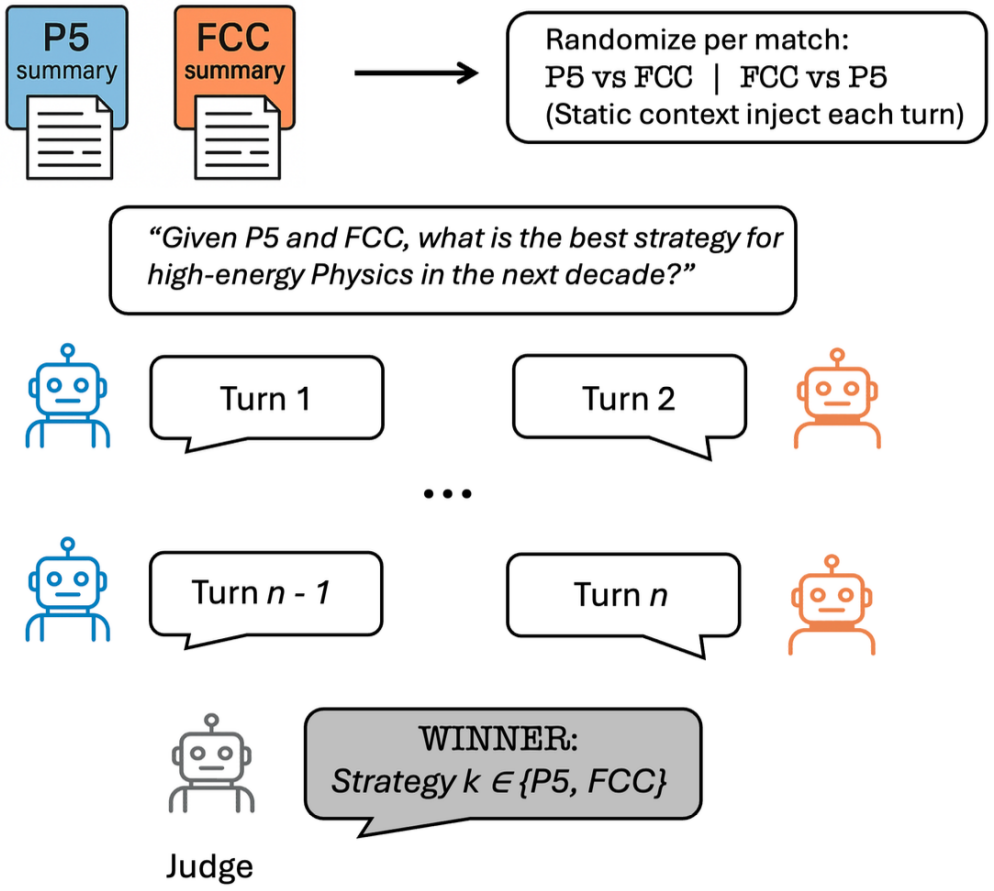


Illustration of our LLM debate setup

- Each match receives the two static summaries concatenated in a randomized order and injected on every turn.
- <u>Initial topic</u>: *"Given the two official planning documents … , what is the most compelling strategy for advancing HEP over the next decade? ... Compare scientific reach, timelines, ..., and cost/risk."*
- A transcript-only judge LLM outputs the winner.

Setup

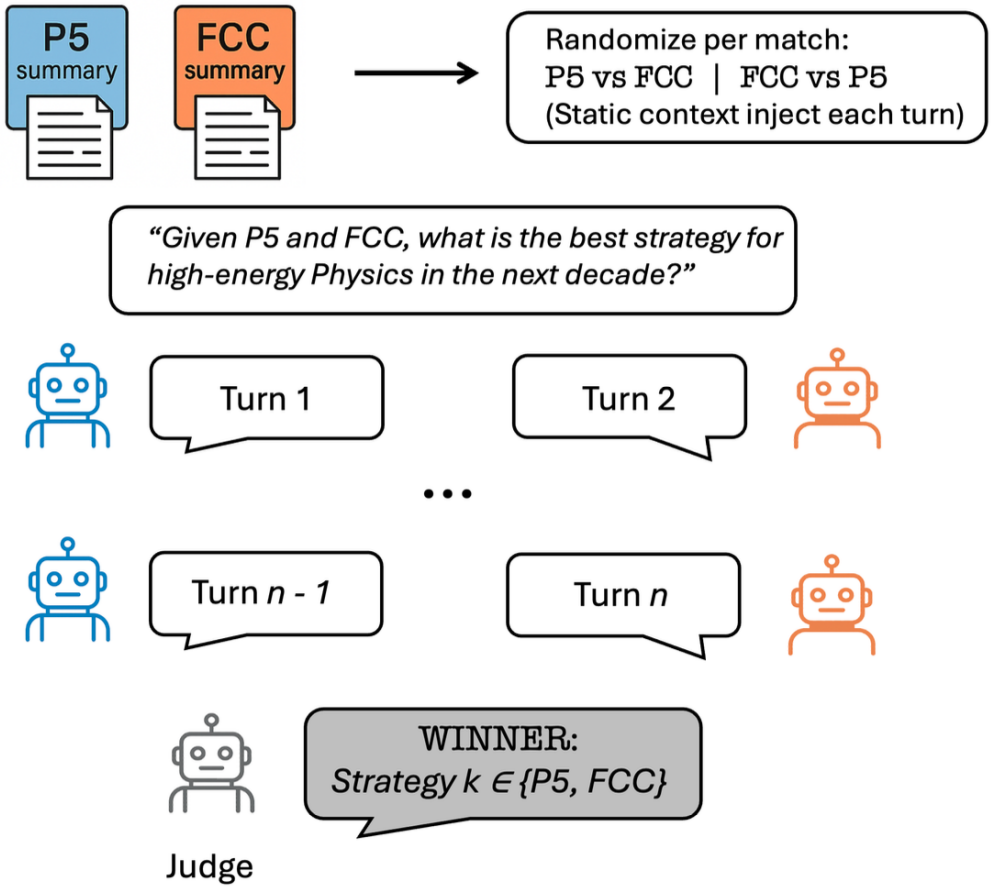


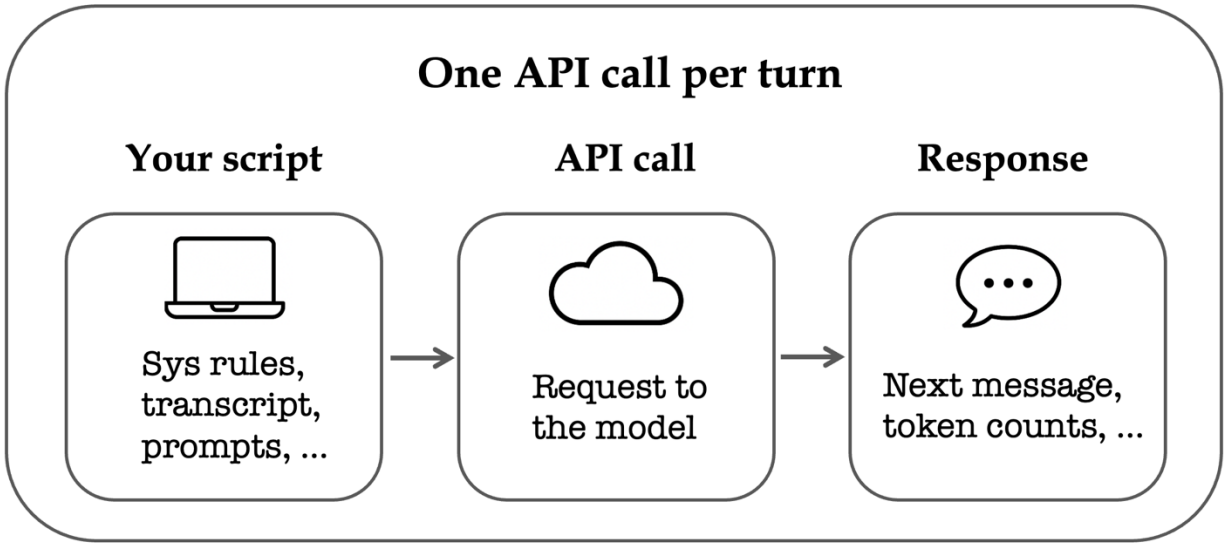
Illustration of our LLM debate setup

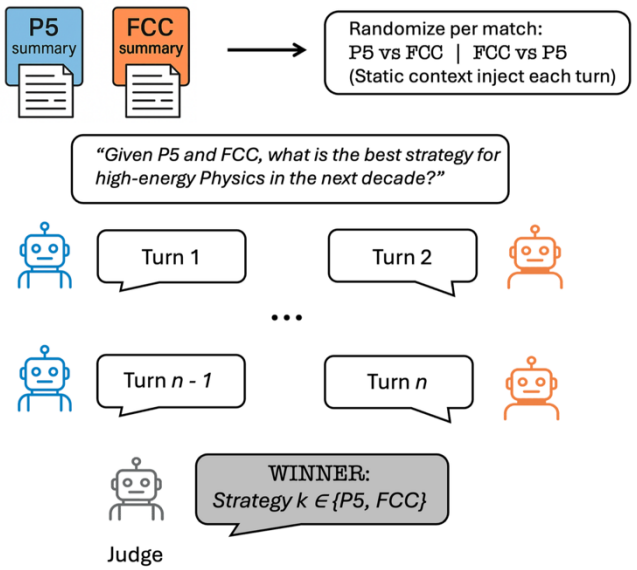
- Each match receives the two static summaries concatenated in a randomized order and injected on every turn.
- <u>Initial topic</u>: *"Given the two official planning documents … , what is the most compelling strategy for advancing HEP over the next decade? ... Compare scientific reach, timelines, ..., and cost/risk."*
- A transcript-only judge LLM outputs the winner.

How we "talk" to the model

Crash-course: LLM APIs

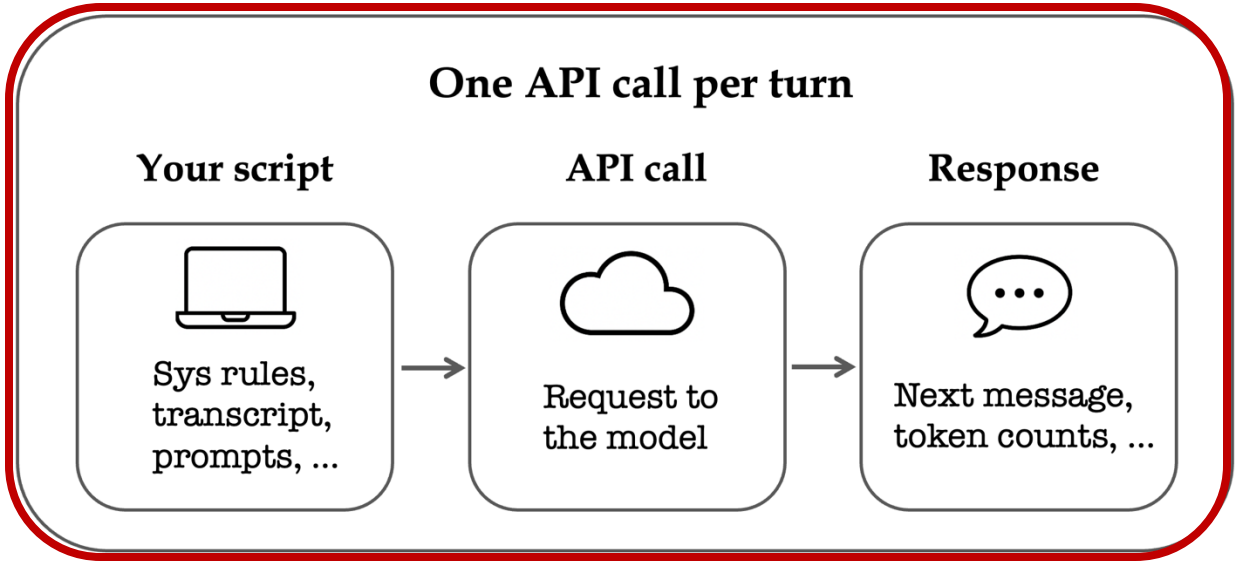
Workflow & Design Choices

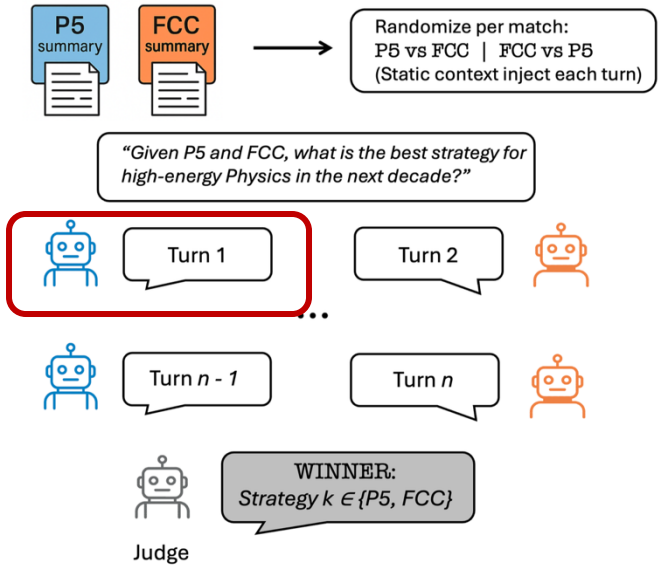




Application Programming Interfaces (APIs)

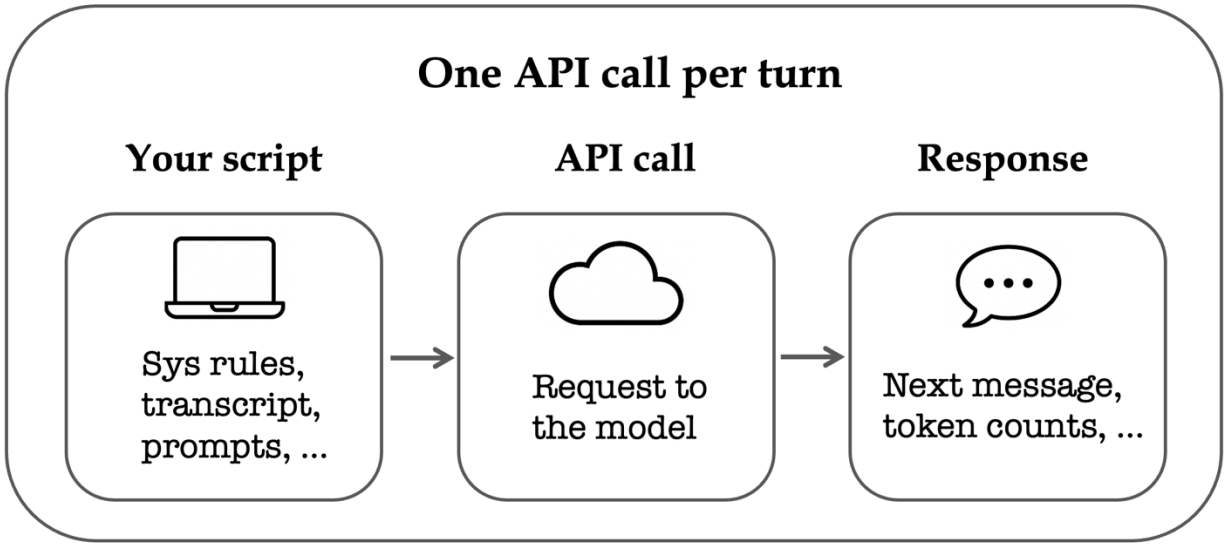
Workflow & Design Choices

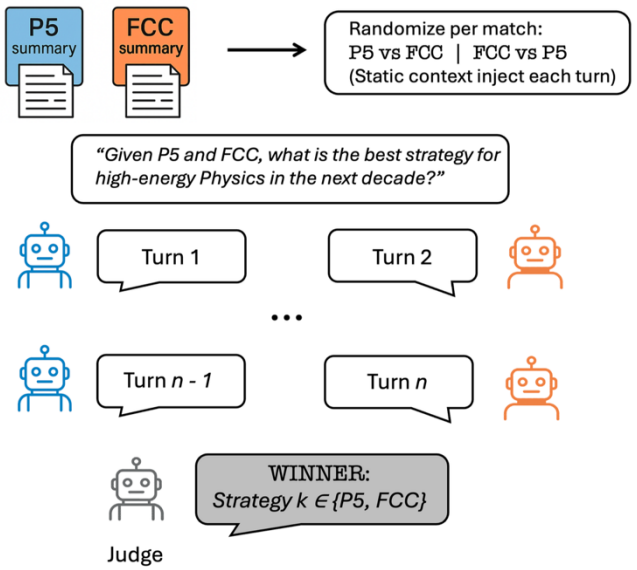




Application Programming Interfaces (APIs)

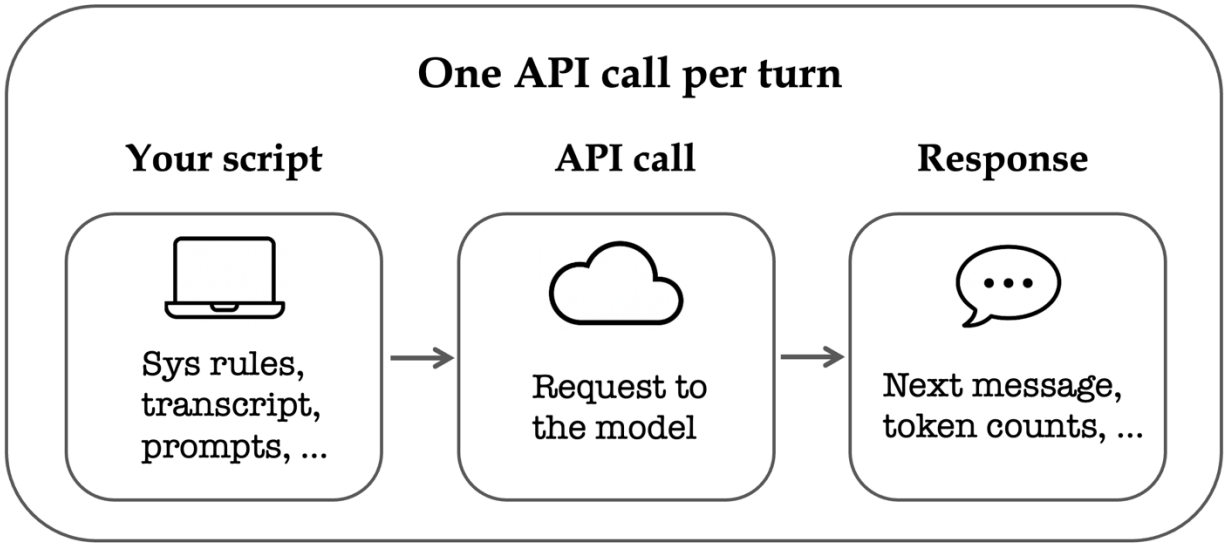
Workflow & Design Choices

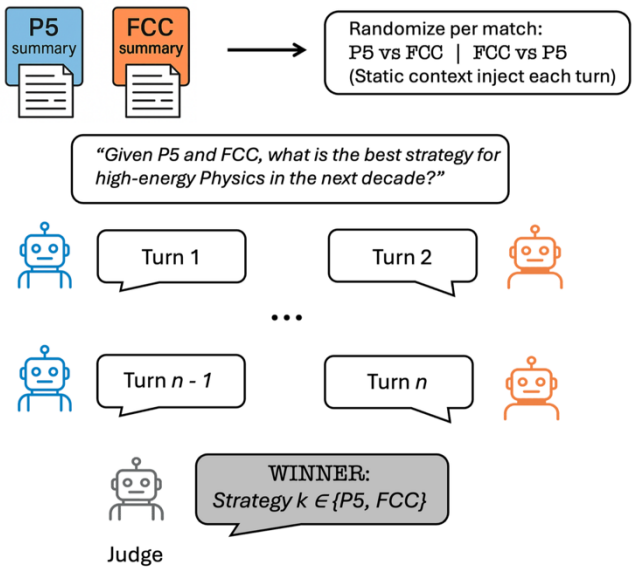




- **API call** = one request → one response
- We send text (prompt) + settings (e.g., token limit)
- The model returns the next message + metadata (e.g., token counts)

Workflow & Design Choices





- **Repeat this per turn** (include all context + transcript)
- **Final judge call** (transcript-only)

Demo

https://github.com/nayara-focs/ai-debate-p5/tree/feat/neutral-multicorpus

```
python scripts/run_debate.py \
  --repeats 6 \
  --turns 6 \
  --context-order random \
  --ctx-p5  docs/p5_summary.txt \
  --ctx-fcc docs/fcc_summary.txt \
  --out runs/$(date +%Y%m%d)/test.json
```

Set your key:

```
export OPENAI_API_KEY=sk-...
```

PLAN

- Setup, Design choices & Workflow
- Measurements
- How the debates are organized

Measurements Two types of ‘measurements’

1. Stance strength [What is the best strategy?]

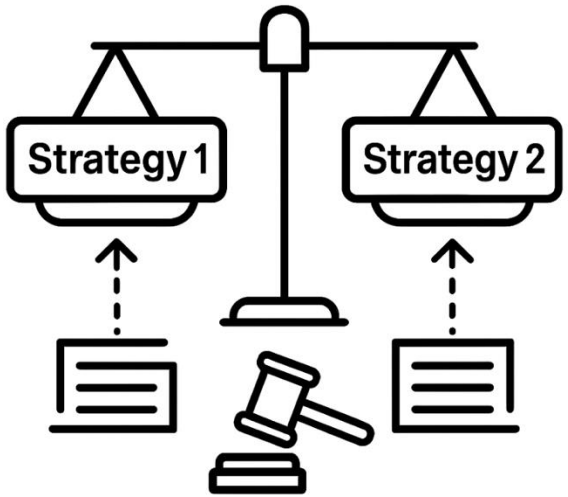
2. Debater strength [What is the best debater?]

Measurements Two types of 'measurements'

1. Stance strength [What is the best strategy?]

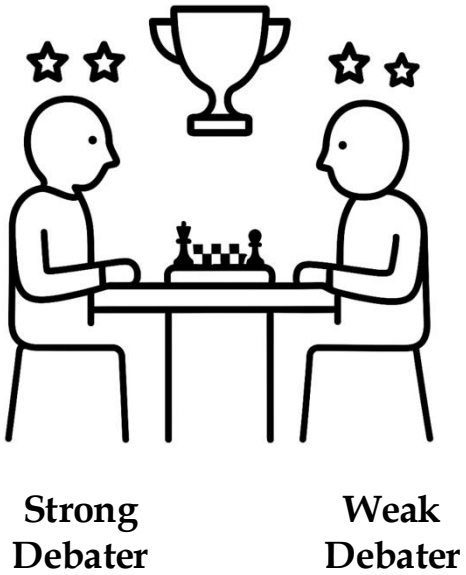
E.g.,

A. Particle Physics Project Prioritization Panel (P5)
B. Future Circular Collider (FCC)



2. Debater strength [What is the best debater?]

E.g., use different model config

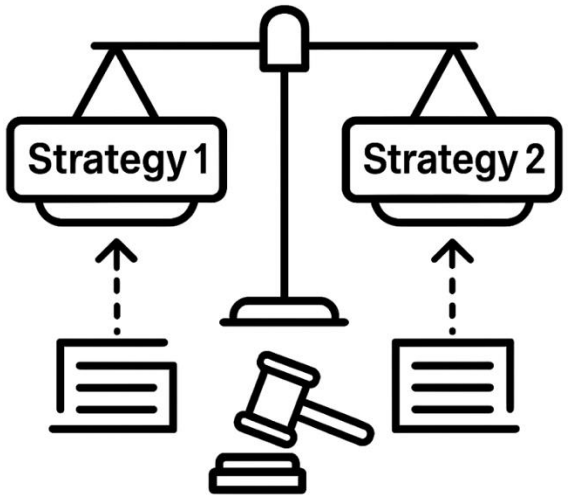


Measurements Two types of 'measurements'

1. Stance strength [What is the best strategy?]

E.g.,

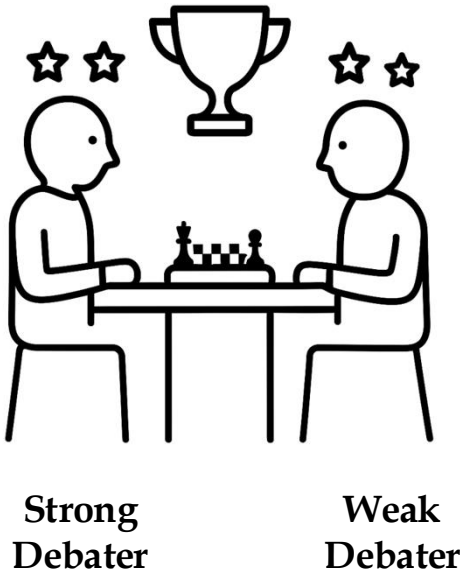
A. Particle Physics Project Prioritization Panel (P5)
B. Future Circular Collider (FCC)



2. Debater strength [What is the best debater?]

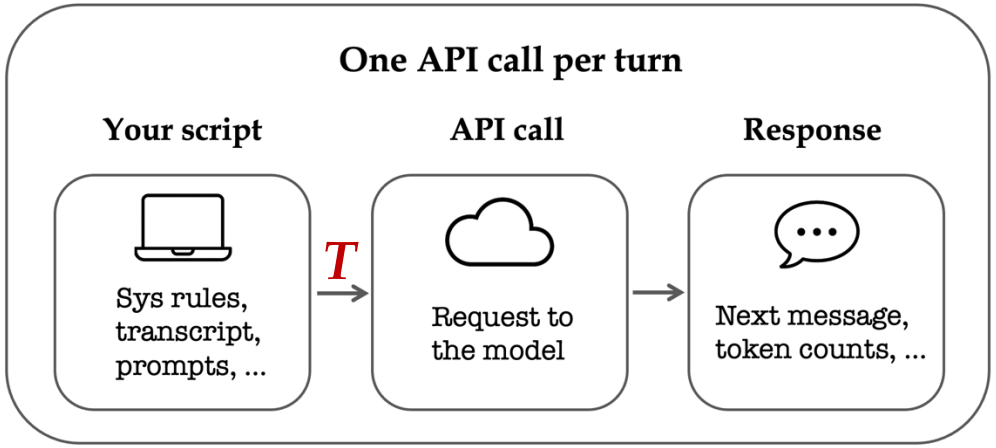
E.g., use different model config

'Temperature' of the model



Measurements

- **Temperature** ***(sampling parameter)***



Crash-course: 'Temperature' on LLM APIs

Temperature *(sampling parameter)*

The model predicts a **distribution over next words** (actually, tokens)

- T: rescale logits → soften/sharpen the next-token distribution
- For logits z_i, sampling uses:

$$p_T(i \mid \text{history}) = \frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)}$$

- $T < 1$: sharper, **more deterministic**
- $T = 1$: baseline softmax
- $T > 1$: flatter, **more diverse** (uniform as $T \to \infty$)

Intuition. This looks like a Boltzmann form, but T is just a **randomness knob** for sampling

Temperature (*sampling parameter*)

The model predicts a **distribution over next words**

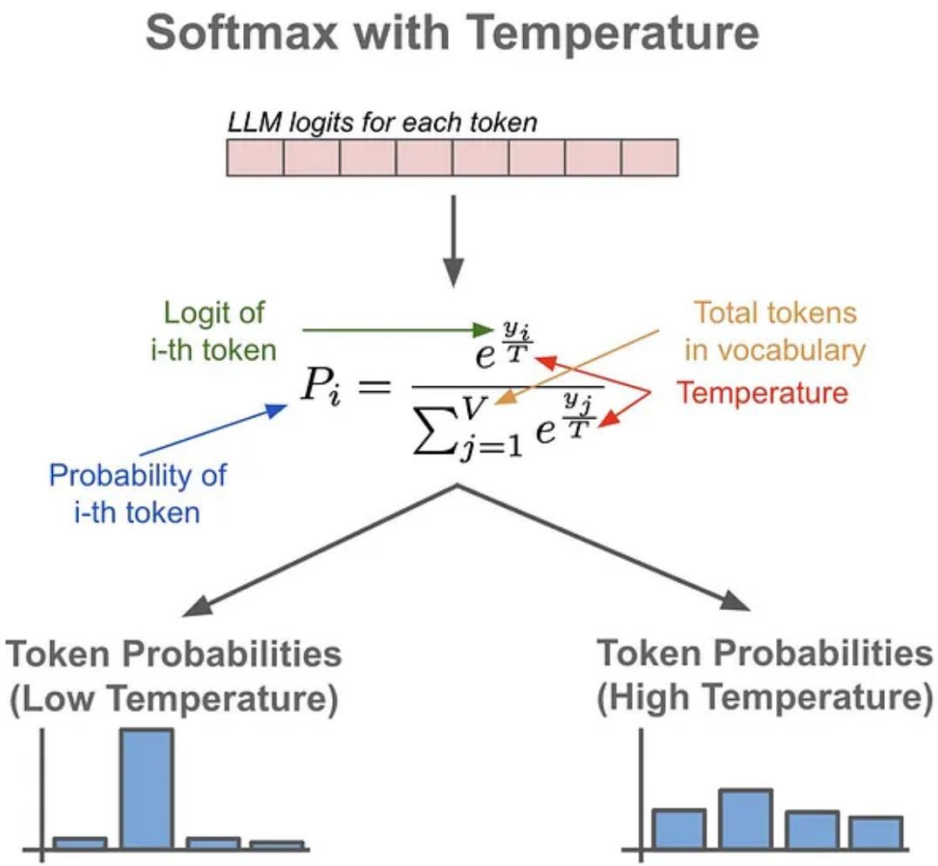


Figure: https://medium.com/%40amansinghalml_33304/temperature-llms-b41d75870510

PLAN

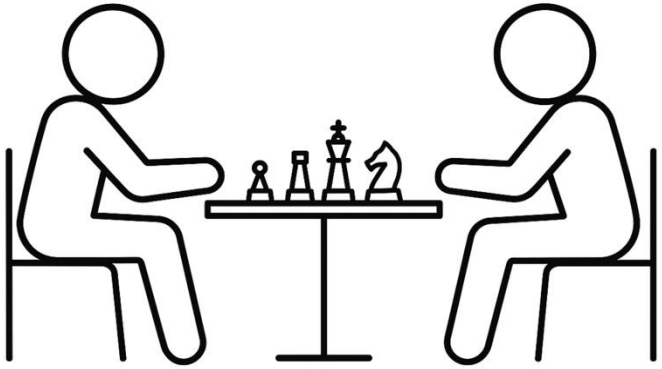
- Setup, Design choices & Workflow
- Measurements
- How the matches are organized

How the matches are organized

- Toy mini-tournament (illustrative)

Elo-style ratings (Bradly-Terry model)

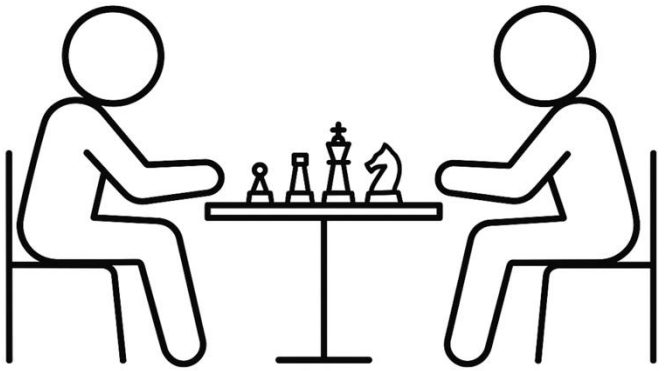
See: The Elo Rating System:
https://www.youtube.com/watch?v=inXUp5j107I



How the matches are organized

- Toy mini-tournament (illustrative)

Elo-style ratings (Bradly-Terry model)



- Pairwise matches: Each debater i has a latent log-strength E_i
- Probability that i beats j in a debate

$$\Pr(i \text{ beats } j) = \sigma(E_i - E_j) = \frac{1}{1 + e^{-(E_i - E_j)}}$$

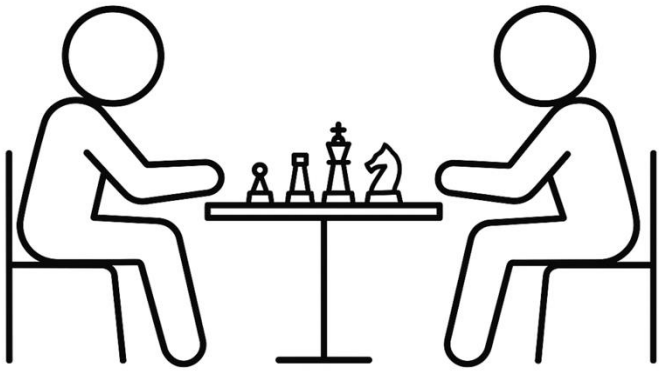
ΔE maps to win probability $\sigma(\Delta E) = \frac{1}{1+e^{-\Delta E}}$

(e.g., $\Delta E{=}0.5 \Rightarrow \sigma \approx 0.62$, $\Delta E{=}1 \Rightarrow \sigma \approx 0.73$)

How the matches are organized

- Toy mini-tournament (illustrative)

Elo-style ratings (Bradly-Terry model)



- Pairwise matches: Each debater i has a latent log-strength E_i
- Probability that i beats j in a debate

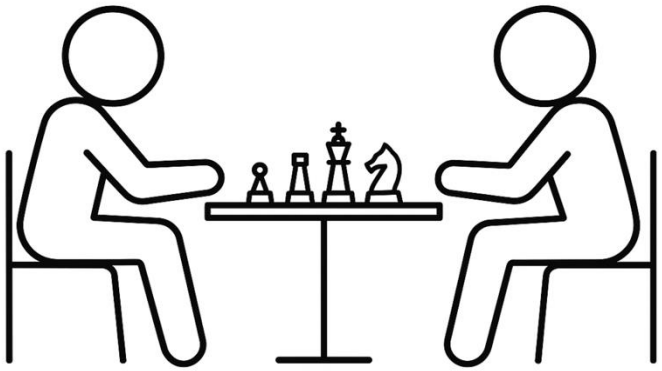
$$\Pr(i \text{ beats } j) = \sigma(E_i - E_j) = \frac{1}{1 + e^{-(E_i - E_j)}}$$

'Beating a stronger debater earns you more points'

How the matches are organized

- Toy mini-tournament (illustrative)

Elo-style ratings (Bradly-Terry model)



Match: Strategy 1 x Strategy 2

[Strategy 1 (P5) x Strategy 2 (FCC)]

Debaters: A/B/C

[A (*T=0.3*) ; B (*T=0.7*); C (*T=1.0*)]

Results (Work in Progress)

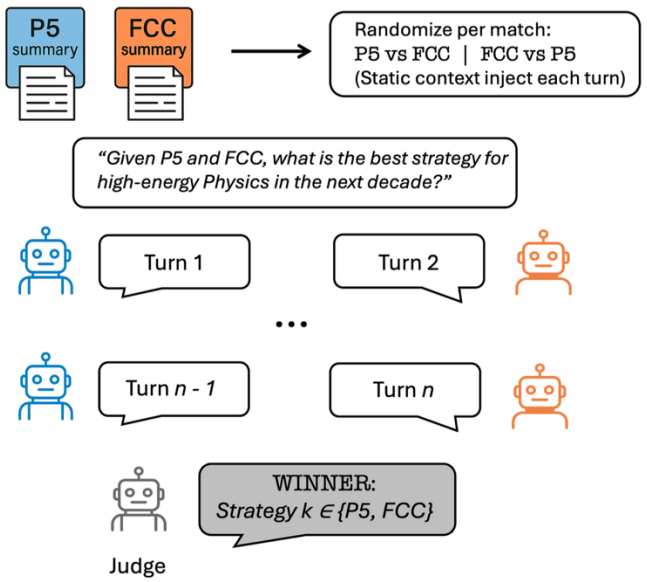
Results (Work in Progress)

- **Model: gpt-4o-mini**
  Low cost + fast latency → affordable to run many turns/matches;
  Good controllability (temperature, token caps)

- **Tournament have 72 matches:** three debaters in cross-play, six repeats, both directions

- **Costs scale with tokens:** Cap tokens per message to keep runs cheap and comparable

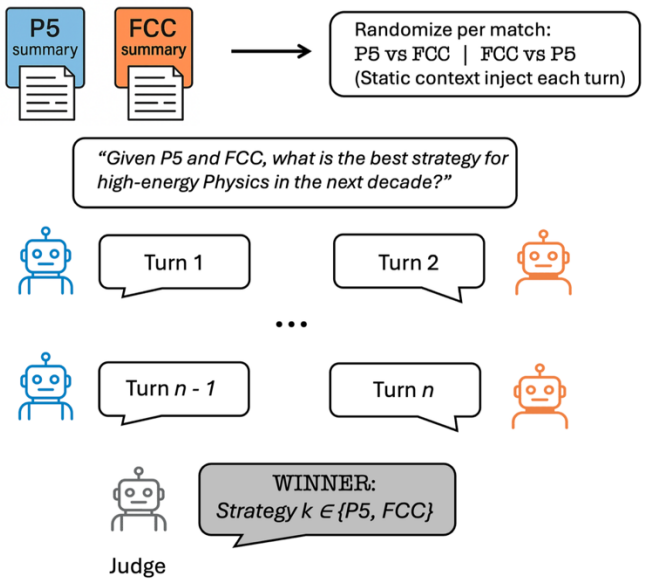
(Tokens ≈ sub-words; for English prose, 1 word ≈ 1.3–1.5 tokens)



Thanks to the API credits

Oxford AI and ML Competency Centre

Results (Work in Progress)



- **Stance Strength**

  A. Particle Physics Project Prioritization Panel (P5)
  B. Future Circular Collider (FCC)

- **Debater Strength** ["debater" variants A/B/C]
  - Temperature
  - Best-of-*N*

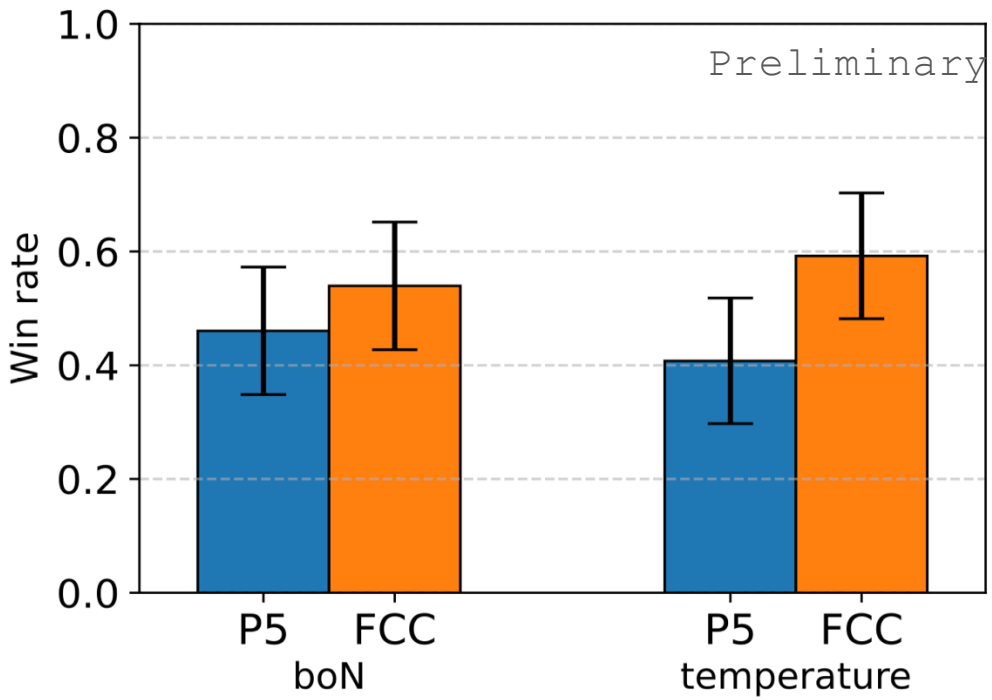
  Variants differ only by one fixed parameter: base model, prompts, judge, and turn limits held constant

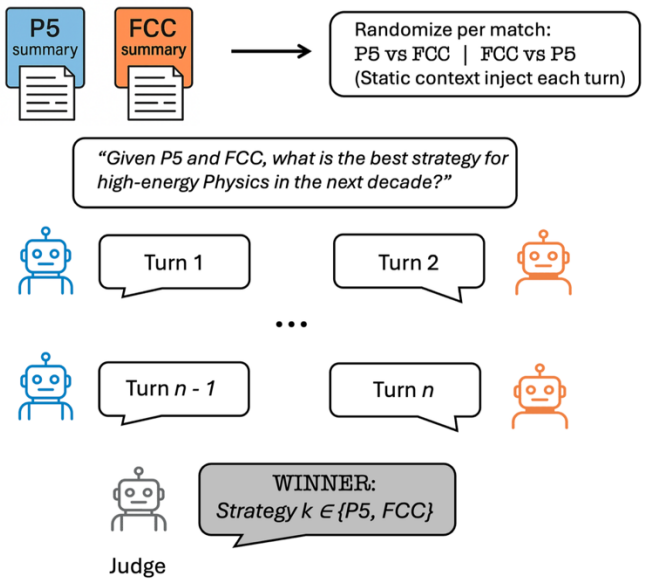
A. https://www.usparticlephysics.org/2023-p5-report/
B. See, e.g., https://home.cern/news/news/accelerators/cern-releases-report-feasibility-possible-future-circular-collider

Results (Work in Progress)

Stance Strength

(Strategy 1 x Strategy 2)



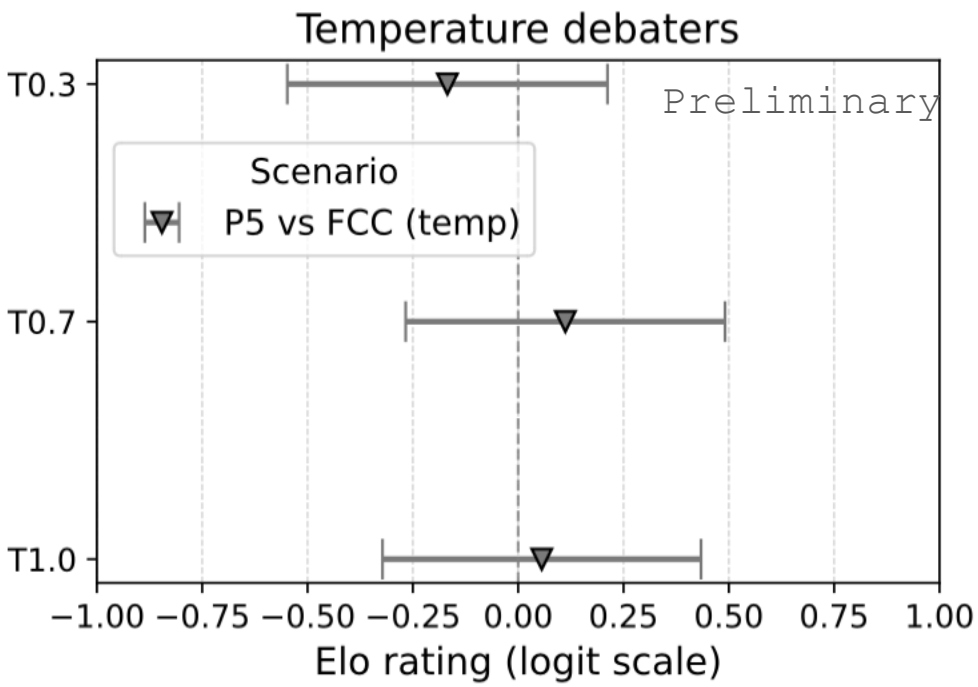


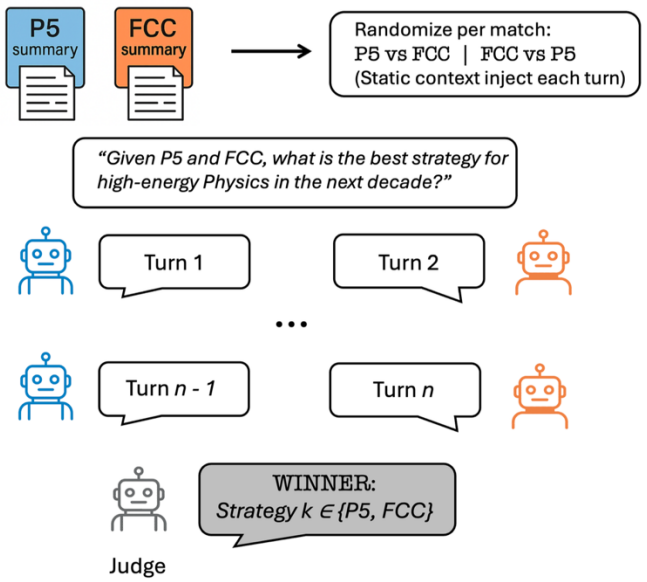
<u>FCC wins by a small margin (CIs overlap)</u>

Results (Work in Progress)

Debater Strength

(Debater A x Debater B)





<u>Intervals overlap broadly, no trend across variants</u>

Conclusion, limitations, and outlook

- Recap: Pioneer AI-driven debates that transparently organize evidence and expert arguments
- For this scale and benchmark: no clear trend

Conclusion, limitations, and outlook

- Recap: Pioneer AI-driven debates that transparently organize evidence and expert arguments

- For this scale and benchmark: no clear trend

- Limitations: Single-question tournaments, one model family, and an LLM judge

- Near-term extensions: re-judging transcripts with stronger and multiple judges, compare additional model families

Conclusion, limitations, and outlook

- Recap: Pioneer AI-driven debates that transparently organize evidence and expert arguments

- For this scale and benchmark: no clear trend

- Limitations: Single-question tournaments, one model family, and an LLM judge

- Near-term extensions: re-judging transcripts with stronger and multiple judges, compare additional model families

Broader scope: multi-question policies, integrate tools, **human-in-the-loop**, and domain replications beyond particle physics

- Curiosity is a guide!
- Science is transformative!

Thanks!