Epistemic Exams

Kola Ayonrinde and Loren K. Fryxell[1]*

[1]City St George's, University of London; GRAPE

loren.fryxell@citystgeorges.ac.uk

July 16, 2026

SAET Rio 2026

*Co-funded by the European Union (ERC, IMD-101040122). Views and opinions expressed are those of the author only and do not necessarily reflect those of the European Union or the European Research Council.

Introduction

Introduction

An **exam** measures what you know.

An **epistemic exam** measures what you know *and* what you know you know.

Example

Student 1 gets 8/10 correct answers on an exam

Results. ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ × ×

After the exam, they report *"I'm very confident my last 8 answers are correct. Moreover, though these are my best guesses, I'm very confident my first 2 answers are wrong."*

Beliefs. × × ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓

Student 2 gets 6/10 correct answers on an exam

Results. ✓ ✓ ✓ ✓ ✓ ✓ × × × ×

After the exam, they report *"I'm very confident my first 6 answers are correct. Moreover, though these are my best guesses, I'm very confident my last 4 answers are wrong."*

Beliefs. ✓ ✓ ✓ ✓ ✓ ✓ × × × ×

Example

Student 1	:	Results.	✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓ × ×
		Beliefs.	× × ✓ ✓ ✓ ✓ ✓ ✓ ✓ ✓
Student 2	:	Results.	✓ ✓ ✓ ✓ ✓ ✓ × × × ×
		Beliefs.	✓ ✓ ✓ ✓ ✓ ✓ × × × ×

- Student 1 **knows** more than Student 2.
  - That is, Student 1 will get a higher fraction of her answers correct than Student 2, given similar questions
- Student 2 **knows what she knows** more than Student 1.
  - That is, Student 2 can perfectly identify which of her answers are correct and which are incorrect, while Student 1 cannot

Epistemic Exams

An epistemic exam is an exam that, alongside each standard question, asks the student to report their confidence in their answer

I.e., the probability they believe their answer to be correct

Suppose Student 1 and Student 2 took such an exam. How should we score them? Should Student 1 or Student 2 score higher?

That's what this paper is about

Preview

Food for Thought and Preview

- Would you prefer to hire Student 1 or Student 2?
- It depends on the context
  - Sometimes it's better to have the student who knows more, even if they know what they know less
  - Other times it's better to have the student who knows what they know more, even if they know less
- The desired scoring rule will depend on the characteristics of the environment the test is simulating

This Paper

The purpose of this paper is to

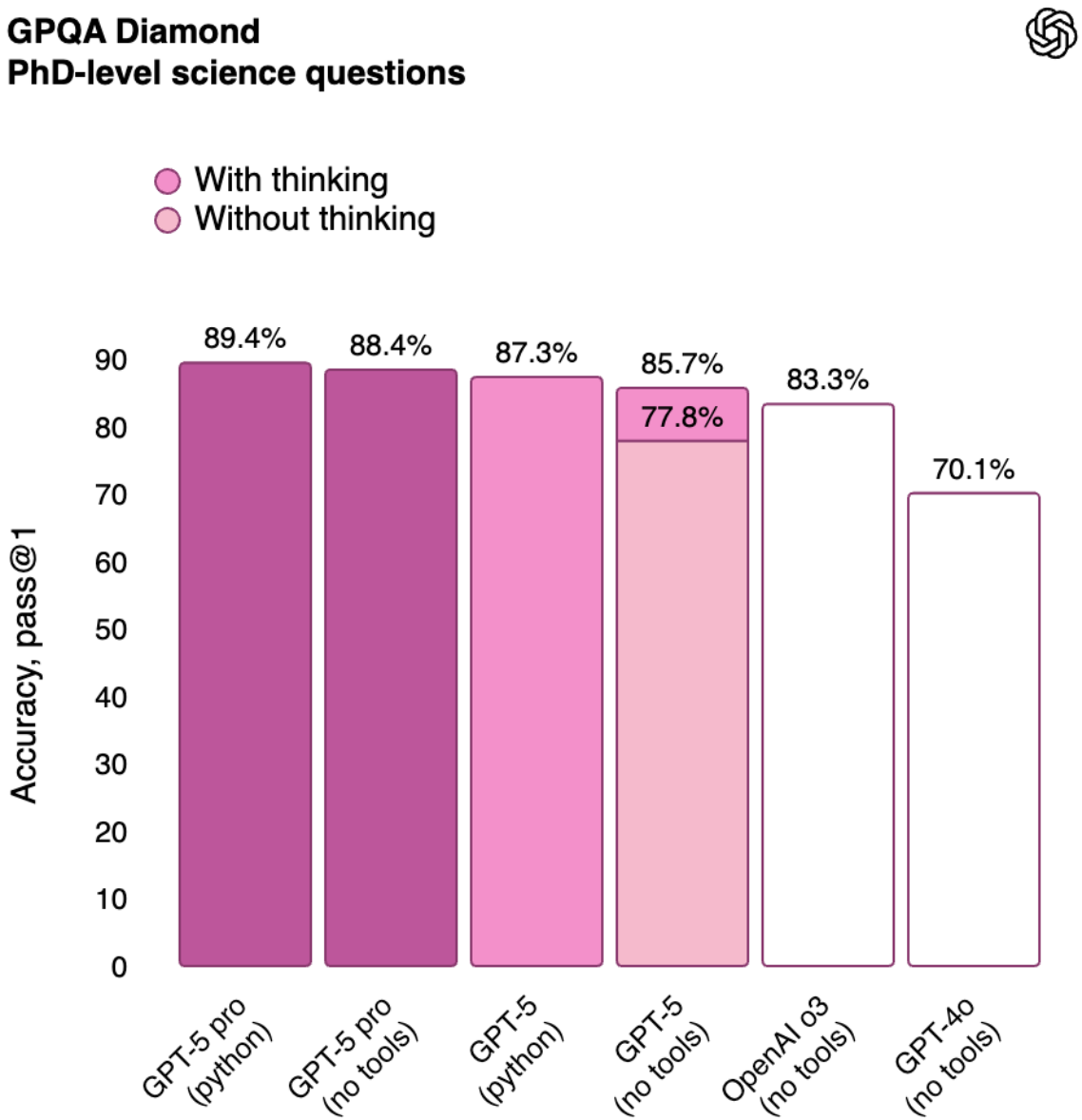
1. Introduce the concept of an epistemic exam
2. Develop a general theory for how to score epistemic exams
3. Argue that **AI capabilities** should be measured using epistemic exams and develop a scoring rule to score AI models called BIFOCAL

For AI, the purpose is twofold:

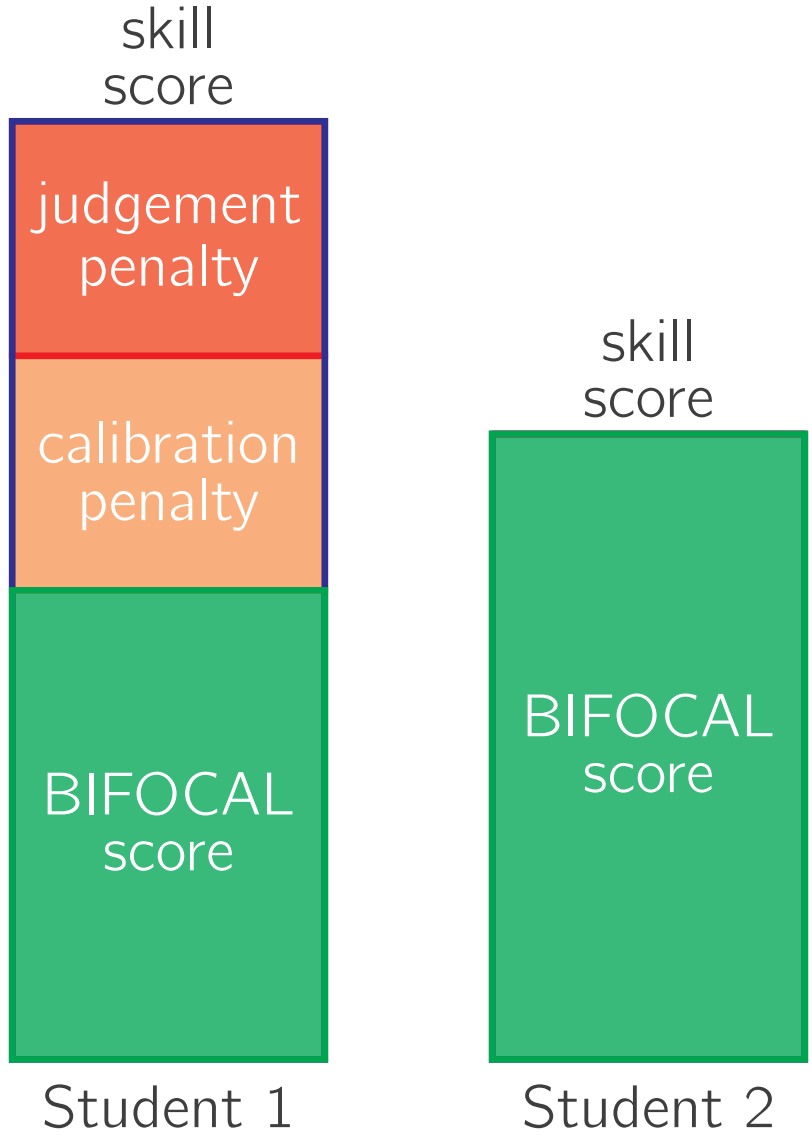
1. To better measure what we actually care about when it comes to AI capabilities
2. To help train AIs to behave closer to how we would like them to behave

In particular, current AIs know a lot, but don't always know what they know, and as a result are **unreliable** (confidently state wrong things)

Current Benchmarks



Our Proposal



Roadmap

1. Scoring Rules
2. Incentive-Compatible Scoring Rules
3. Utilitarian Scoring Rules
4. BIFOCAL
5. Empirical Results

Scoring Rules

Forecasting

Forecasting Environment

A *forecaster* is an agent who is given an event and is asked to report their subjective belief that the event will occur

E.g., if it will rain tomorrow

A forecaster has a prior belief p about the probability the event occurs

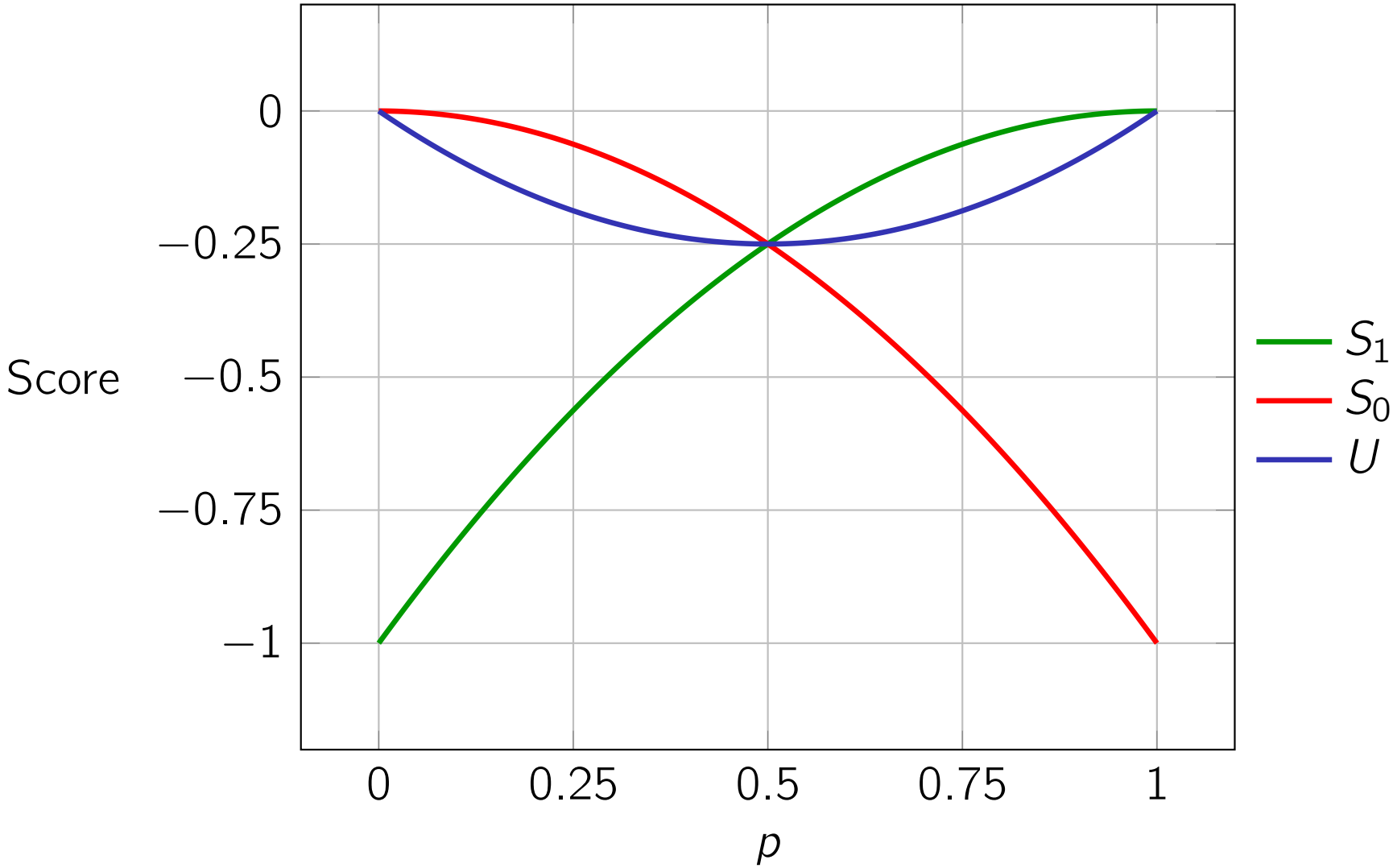
We would like to incentivize the forecaster to report p by providing a score based on the realized outcome and their reported p

Forecasting Environment: Scoring Rules

Definition. A *scoring rule* (S_0, S_1) is a pair of functions $S_0 : [0,1] \to [-\infty, \infty]$ and $S_1 : [0,1] \to [-\infty, \infty]$ where

- $S_0(p)$ is the score given to the reported belief p if the event does not occur and
- $S_1(p)$ is the score given to the reported belief p if the event does occur.

Quadratic Scoring Rule



Test-Taking

Test-Taking Environment

A *test-taker* is an agent who is given a question and is asked to report their best answer and their subjective belief that their answer is correct.

Test-Taking Environment

Let

- X be a finite set of feasible answers
- $\pi : X \to [0, 1]$ be the test-taker's prior belief that each feasible answer is correct

To gain some intuition, consider two common cases:

- Case 1: Four Multiple Choice Options
  - $X = \{A, B, C, D\}$
  - There is only one right answer so $\pi(A) + \pi(B) + \pi(C) + \pi(D) = 1$
- Case 2: Free Response
  - $X = \{\text{all possible finite length strings}\}$
  - There are many right answers (many ways to say the correct thing), so probabilities need not sum to 1

Test-Taking Environment

The agent submits a pair (x, p) where $x \in X$ and $p \in [0, 1]$

Test-Taking Environment: Scoring Rules

A scoring rule here is still the same object as before

The only difference is that the event being predicted (whether the reported answer is correct) is now selected by the agent, rather than exogenous (e.g., whether it will rain tomorrow)

Definition. A *scoring rule* (S_0, S_1) is a pair of functions $S_0 : [0,1] \to [-\infty, \infty]$ and $S_1 : [0,1] \to [-\infty, \infty]$ where

- $S_0(p)$ is the score given to the reported belief p if the given answer is not correct, and
- $S_1(p)$ is the score given to the reported belief p if the given answer is correct.

Incentive-Compatible Scoring Rules

Forecasting

Proper Scoring Rules

Definition. A scoring rule (S_0, S_1) is *(strictly) proper* if the forecaster (uniquely) maximizes her expected score by reporting her true belief. That is, for any $p^* \in [0, 1]$ and $p \neq p^*$,

$$p^* S_1(p^*) + (1 - p^*) S_0(p^*) \geq (>)\, p^* S_1(p) + (1 - p^*) S_0(p).$$

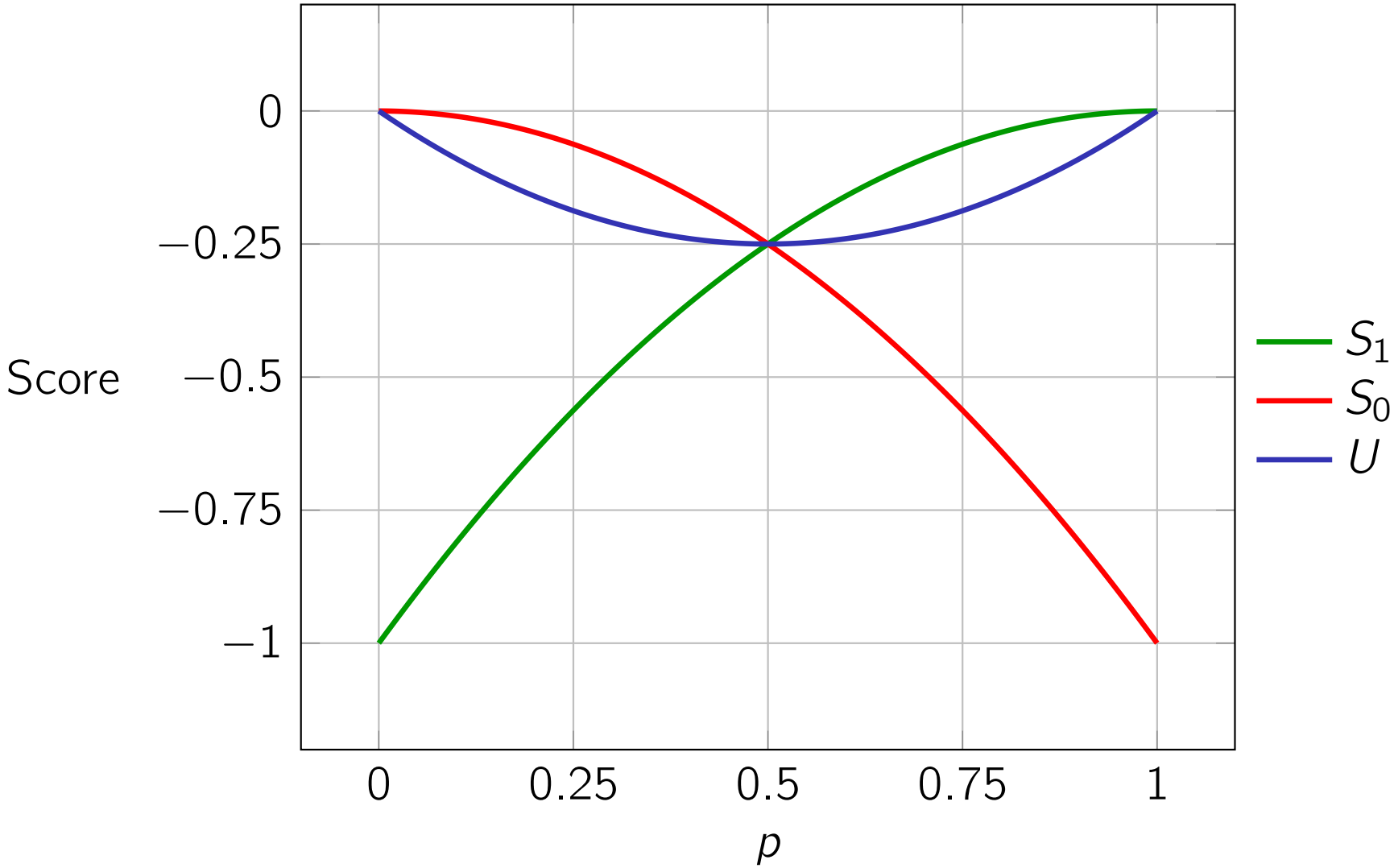
Quadratic Scoring Rule

For example, the *quadratic scoring rule* is defined by

$$S_0(p) = -p^2 \quad \text{and} \quad S_1(p) = -(1-p)^2$$

and is a strictly proper scoring rule.

Quadratic Scoring Rule



Characterization of Proper Scoring Rules

Theorem 1 (McCarthy, Savage). A scoring rule S is (strictly) proper if and only if

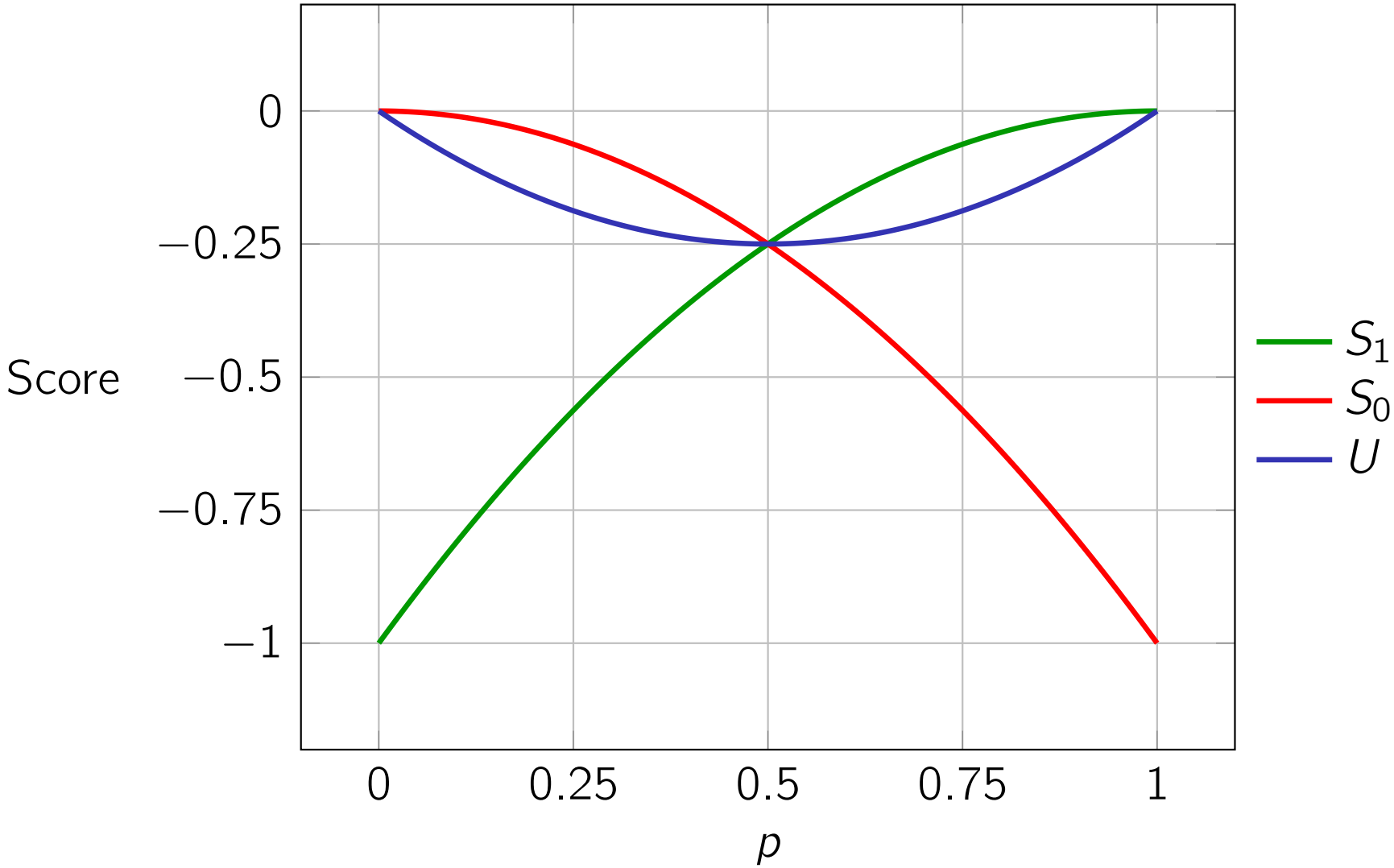
$$S_0(p) = U(p) - pU'(p) \quad \text{and} \quad S_1(p) = U(p) + (1-p)U'(p)$$

where $U : [0,1] \to \mathbb{R}$ is a (strictly) convex function.

Moreover, U is precisely the expected score function, i.e.,

$$U(p) = pS_1(p) + (1-p)S_0(p).$$

Quadratic Scoring Rule



Test-Taking

Proper Epistemic Exam Scoring Rules

Definition. A scoring rule (S_0, S_1) is a *(strictly) proper epistemic exam scoring rule* if the test-taker (uniquely) maximizes her expected score by submitting her best answer x^* along with her true belief that it is correct $\pi(x^*)$.

That is, for any X, $\pi : X \to [0, 1]$, $x^* \in \arg\max_{x \in X} \pi(x)$, $x \notin \arg\max_{x \in X} \pi(x)$, and $\hat{p} \neq \pi(x^*)$,

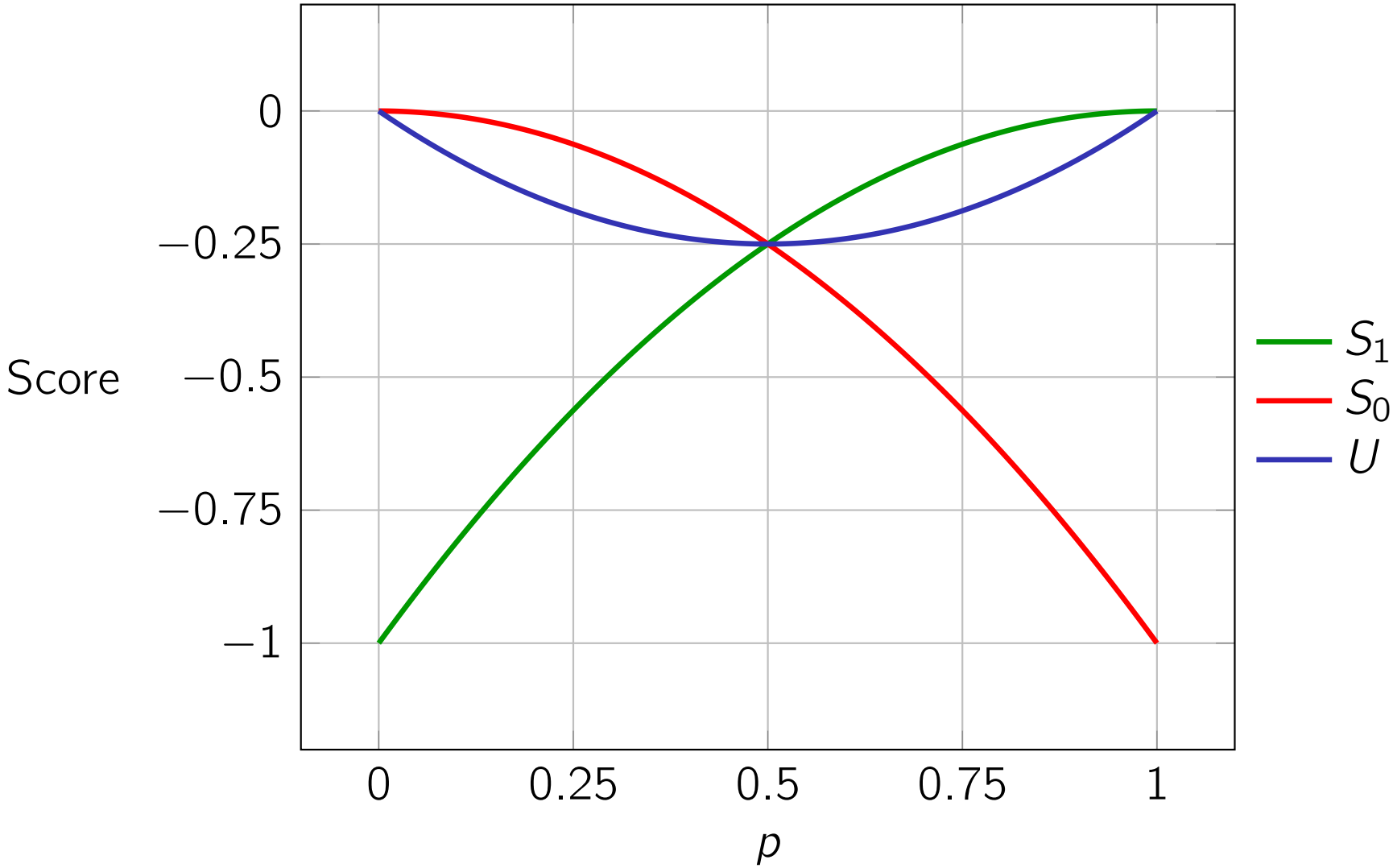
$$\pi(x^*)S_1(\pi(x^*)) + (1 - \pi(x^*))S_0(\pi(x^*)) \geq(>) \pi(x)S_1(\hat{p}) + (1 - \pi(x))S_0(\hat{p}).$$

Characterization of Proper Epistemic Exam Scoring Rules

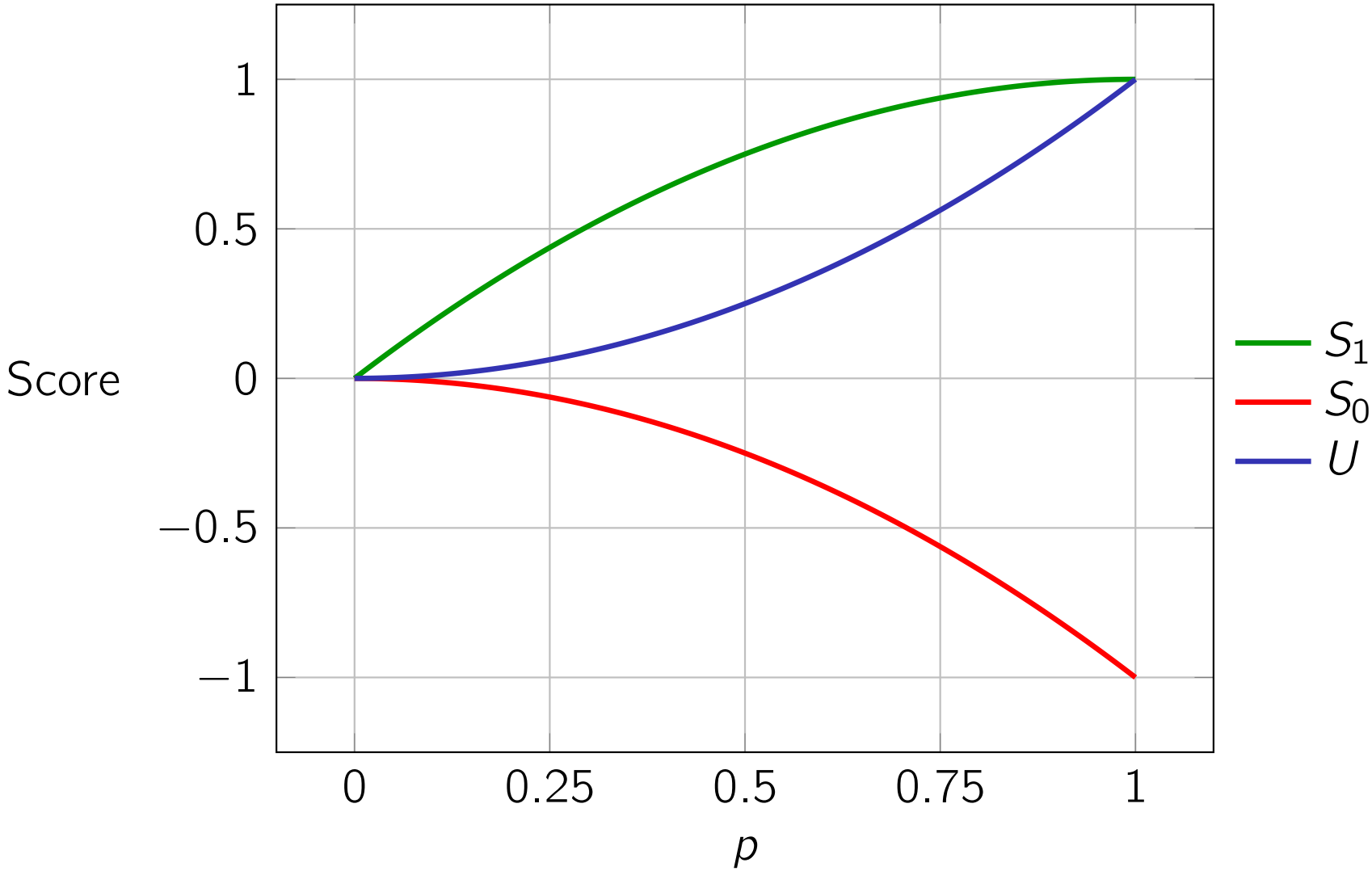
Theorem 2. A scoring rule (S_0, S_1) is a (strictly) proper exam scoring rule if and only if (S_0, S_1) is (strictly) proper AND

- the expected score of reporting an answer with true belief p, $U(p) = pS_1(p) + (1-p)S_0(p)$, is (strictly) increasing in p OR
- the score for a correct answer with zero confidence is weakly larger than the score for an incorrect answer with zero confidence, $S_1(0) \geq S_0(0)$.

Quadratic Scoring Rule



Quadratic Epistemic Exam Scoring Rule



Summary

$$S_0(p) = U(p) - pU'(p) \text{ and } S_1(p) = U(p) + (1-p)U'(p)$$

plus

U convex characterizes the full class of *proper scoring rules*

U convex and increasing characterizes the full class of *proper epistemic exam scoring rules*

Utilitarian Scoring Rules

Utilitarian Scoring Rule

A *utilitarian scoring rule* is a scoring rule that reflects the payoff to society of employing the agent

The payoff to society can be thought of as the payoff to the **end-user** of the forecast (in the forecasting environment) or answer (in the test-taking environment)

Forecasting

Forecasting Environment with User

Let A be the set of all feasible actions a user can take upon receiving some forecast p

E.g., $A = \{\text{bring umbrella, don't bring umbrella}\}$

Let $U_0 : A \to \mathbb{R}$ and $U_1 : A \to \mathbb{R}$ where

- $U_0(a)$ is the user's payoff of taking action a given the forecasted event does not occur
- $U_1(a)$ is the user's payoff of taking action a given the forecasted event does occur

Utilitarian Scoring Rule

Upon receiving forecast p, the user chooses an action

$$a^*(p) \in \operatorname*{arg\,max}_{a \in A} \; pU_1(a) + (1-p)U_0(a)$$

Definition. A *utilitarian scoring rule* is a scoring rule that reflects the expected utility to society of employing that agent, i.e., for all $p \in [0, 1]$,

$$S_0(p) = U_0(a^*(p)) \quad \text{and} \quad S_1(p) = U_1(a^*(p)).$$

Proposition 1

Proposition 1. In a forecasting environment, every utilitarian scoring rule is a proper scoring rule.

Intuitively, the forecaster and the user are perfectly aligned—and the user makes optimal decisions given the forecast—so the forecaster wants to give the user the true forecast

Test-Taking

Test-Taking Environment with User

A test-taking environment adds a set X of possible answers the user can receive from the agent:

- X is the set of possible answers the agent can give
- A is the set of feasible actions the user can take upon receiving answer x and confidence p

E.g., $A = \{\text{disregard answer, evaluate answer, implement answer}\}$

Let $U_0 : A \times X \to \mathbb{R}$ and $U_1 : A \times X \to \mathbb{R}$ where

- $U_0(a, x)$ is the user's payoff of taking action a given the answer x is incorrect
- $U_1(a, x)$ is the user's payoff of taking action a given the answer x is correct

Utilitarian Scoring Rule

Upon receiving answer x with confidence p, the user chooses action

$$a^*(p, x) \in \arg\max_{a \in A} \; pU_1(a, x) + (1 - p)U_0(a, x)$$

Let $U(p) = \max_{a \in A} \; pU_1(a, x) + (1 - p)U_0(a, x)$ be the user's indirect utility as a function of the given confidence p

Assumption. U is weakly increasing (better answers result in higher payoffs for the user).

Definition. A *utilitarian scoring rule* is a scoring rule that reflects the expected utility to society of employing that agent, i.e., for all $p \in [0, 1]$,

$$S_0(p) = U_0(a^*(p, x)) \quad \text{and} \quad S_1(p) = U_1(a^*(p, x)).$$

Proposition 2

Proposition 2. In a test-taking environment, every utilitarian scoring rule is a proper epistemic exam scoring rule.

This follows from Proposition 1 and the fact that U is weakly increasing

Takeaway

In economics, we are usually quite concerned with incentive compatibility

And for the application to human students, this would be important

However, with AI, it's not obvious that we need their scores to be incentive-compatible

What we do want is that the score measures their usefulness (the payoff to society of using that AI)

Proposition 2 shows that if we want to measure an AI's usefulness to society, then that scoring rule will also be incentive-compatible

BIFOCAL

BIFOCAL

We would like to design a utilitarian scoring rule that reflects the payoff to society of employing the AI that we are testing

We consider three actions the user can take:

1. **Disregard** : Don't implement
2. **Evaluate** : Check the answer and implement only if the user believes it to be correct
3. **Implement** : Do implement

The User's (Society's) Payoffs

Normalize the status quo payoff of not having an AI agent (equivalently, disregarding its answer) to 0

We now define the user's payoff as a function of 7 parameters

The 7 Parameters

- I : importance, the benefit of success
  - the benefit of implementing a correct answer
  - (for a single question this can be normalized to 1, for multiple questions this measures the relative importance across questions)
- F : cost of failure
  - the cost of implementing an incorrect answer (relative to the benefit of implementing a correct answer)
  - $F = \frac{1-\lambda}{\lambda}$ where λ is the failure rate at which the user is just indifferent between using the AI and not
- C : cost of checking
  - the cost of checking an answer (relative to the benefit of implementing a correct answer)
  - $C = \frac{1}{n}$ where n is the number of answers at which the user would be just indifferent between checking n answers to get one correct answer and not bothering

The 7 Parameters (continued)

- α : type I error of the checking technology
  - the rate of false positives associated with the evaluation technology (evaluation says answer is correct when it's incorrect)
- β : type II error of the checking technology
  - the rate of false negatives associated with the evaluation technology (evaluation says answer is incorrect when it's correct)
- O : omission rate of the checking technology
  - the rate of omissions (evaluation is unable to verify the answer as being correct or incorrect)
- L : luck
  - the confidence level associated with random guessing (1 / the number of multiple choice answers)

The 7 Parameters

And if we reshuffle those 7 parameters...

- β : type II error
- I : importance
- F : cost of failure
- O : omission rate
- C : cost of checking
- α : type I error
- L : luck

Moreover, like bifocal lenses, BIFOCAL allows us to see two different things clearly—the agent's skill (what they know) and epistemics (what they know they know)

BIFOCAL

Interactive Visual

Skill, Judgement, and Calibration

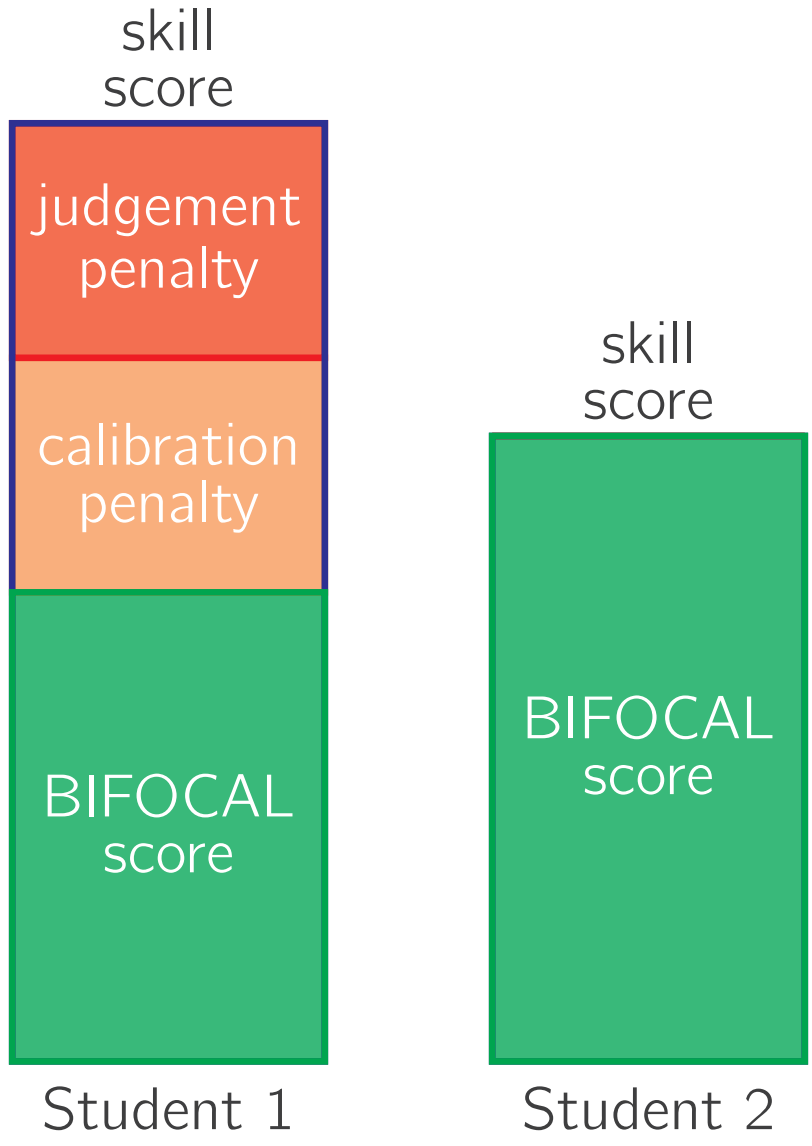
We can decompose BIFOCAL into three pieces: skill, judgement, and calibration

the expected payoff to society of employing the agent period is
BIFOCAL = Skill Score − Judgement Penalty − Calibration Penalty

the expected payoff to society of employing the agent *if* they had perfect calibration (if they know what they know on average) is
Skill Score − Judgement Penalty

the expected payoff to society of employing the agent *if* they had perfect judgement and calibration (if they know what they know exactly) is
Skill Score (standard exam score)

Example from Introduction



Empirical Results

Empirical Results

The goal of the empirical section is twofold:

1. Learn some useful things about A(G)I
2. Demonstrate some of the things you can do with epistemic exams and BIFOCAL

Empirical Results

For the learning useful things bit, we are currently investigating the following two questions:

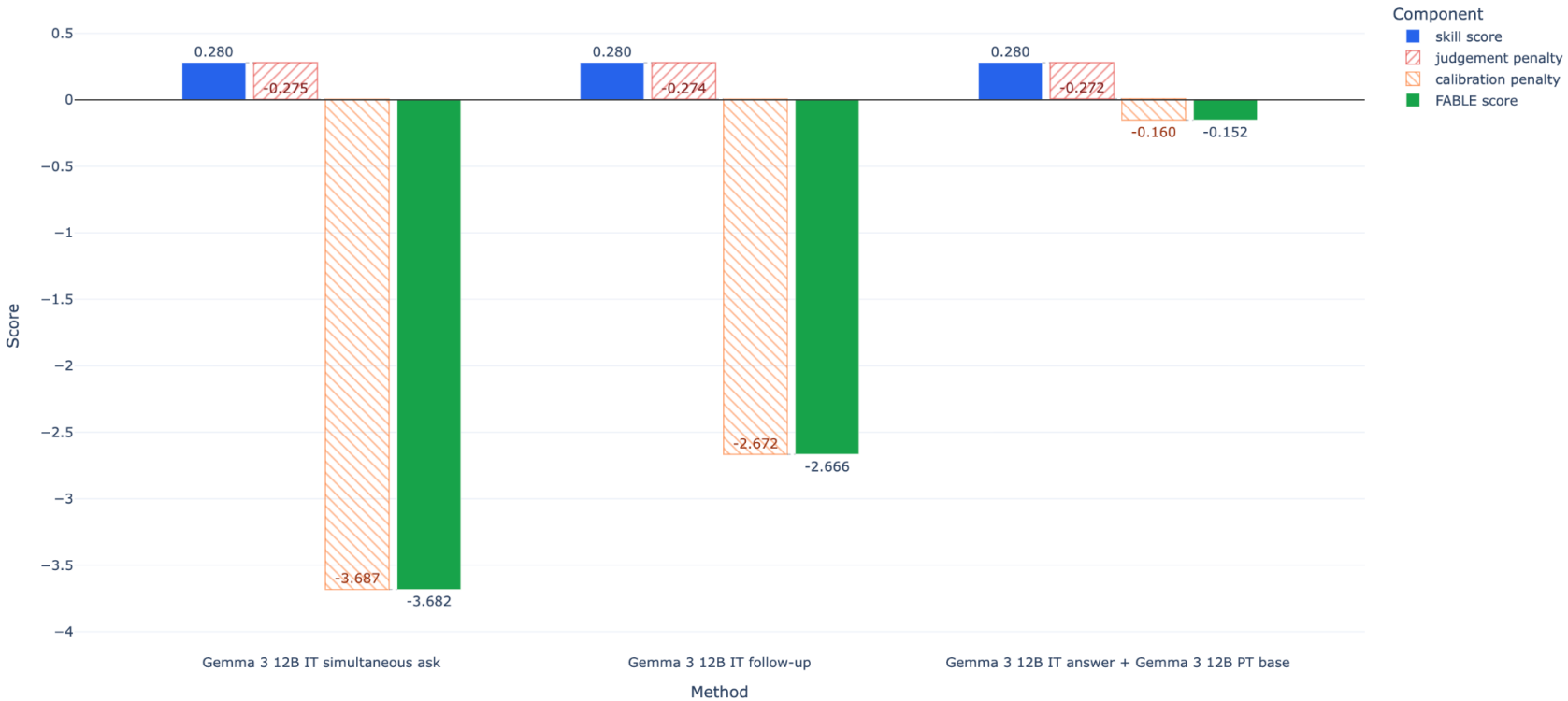
1. How do different confidence elicitation methods compare?
   - There are many different ways to elicit an AIs confidence
   - How do they compare (measured by their BIFOCAL score)?
2. Using the best confidence elicitation method we currently have access to, how capable are frontier models really?
   - In particular, how do they perform on the epistemic exam version of standard benchmarks?

Confidence Elicitation

We tested four types of confidence elicitation:

1. **Simultaneous Ask:** Ask the AI to report its confidence alongside the answer
2. **Reflective Ask:** Ask the AI to report its confidence after it has already answered
3. **Base Model Audit:** Ask the base version of the model whether the answer is correct and extract its internal probability of that token
4. **Base Model Ask:** Ask the base version of the model its confidence that the answer is correct

Confidence Elicitation



The base model ask was so bad we didn't include it in the figure

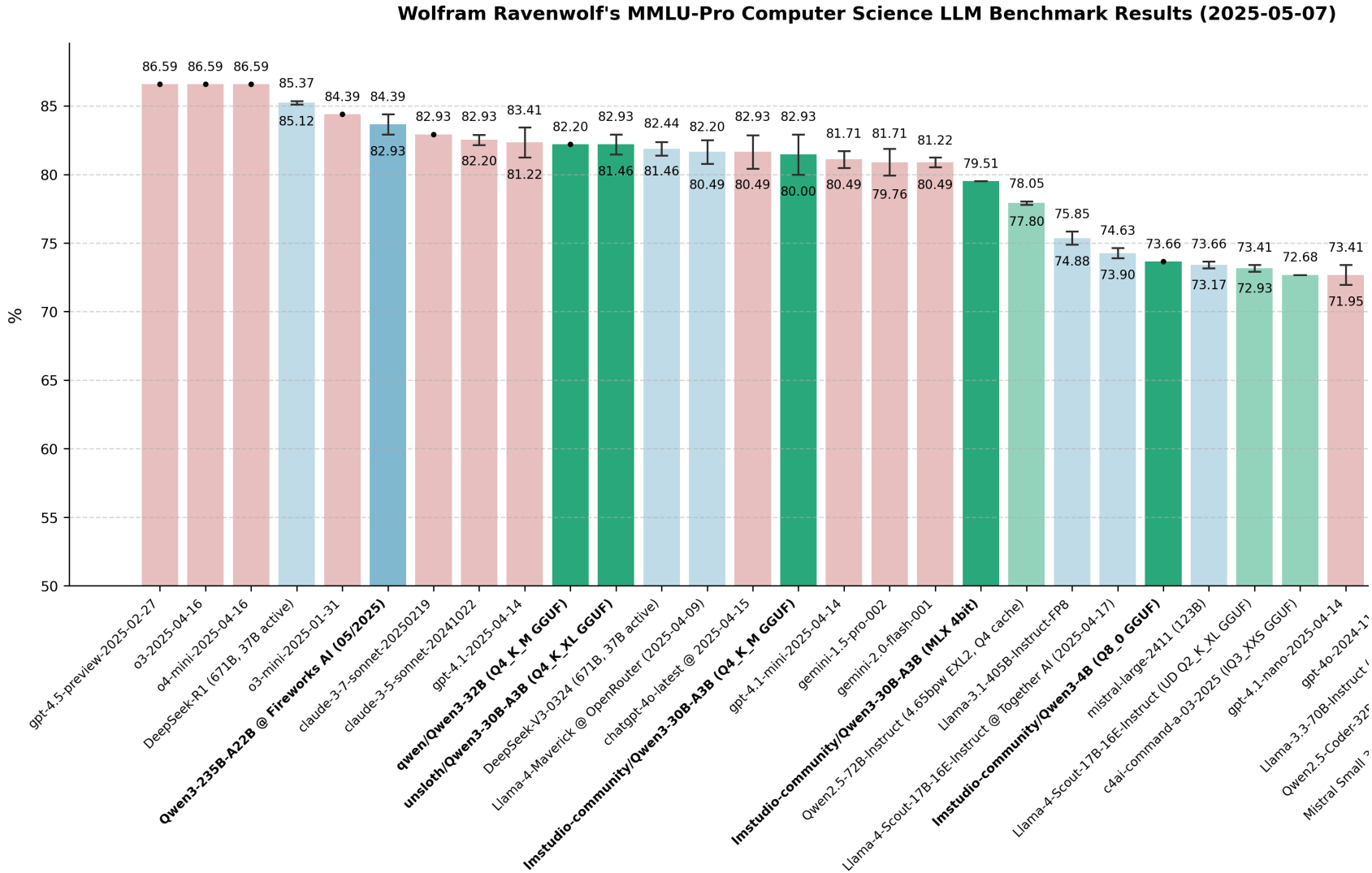
Frontier Capabilities on MMLU

MMLU is a database of 15k multiple-choice questions spanning 57 subjects including STEM, international law, nutrition, and religion

As you will see, standard scoring implies that frontier models have "saturated" this benchmark

That is, they are so capable that the benchmark is now considered too easy for them

MMLU Standard Scores (Skill Scores)



Frontier Capabilities on MMLU

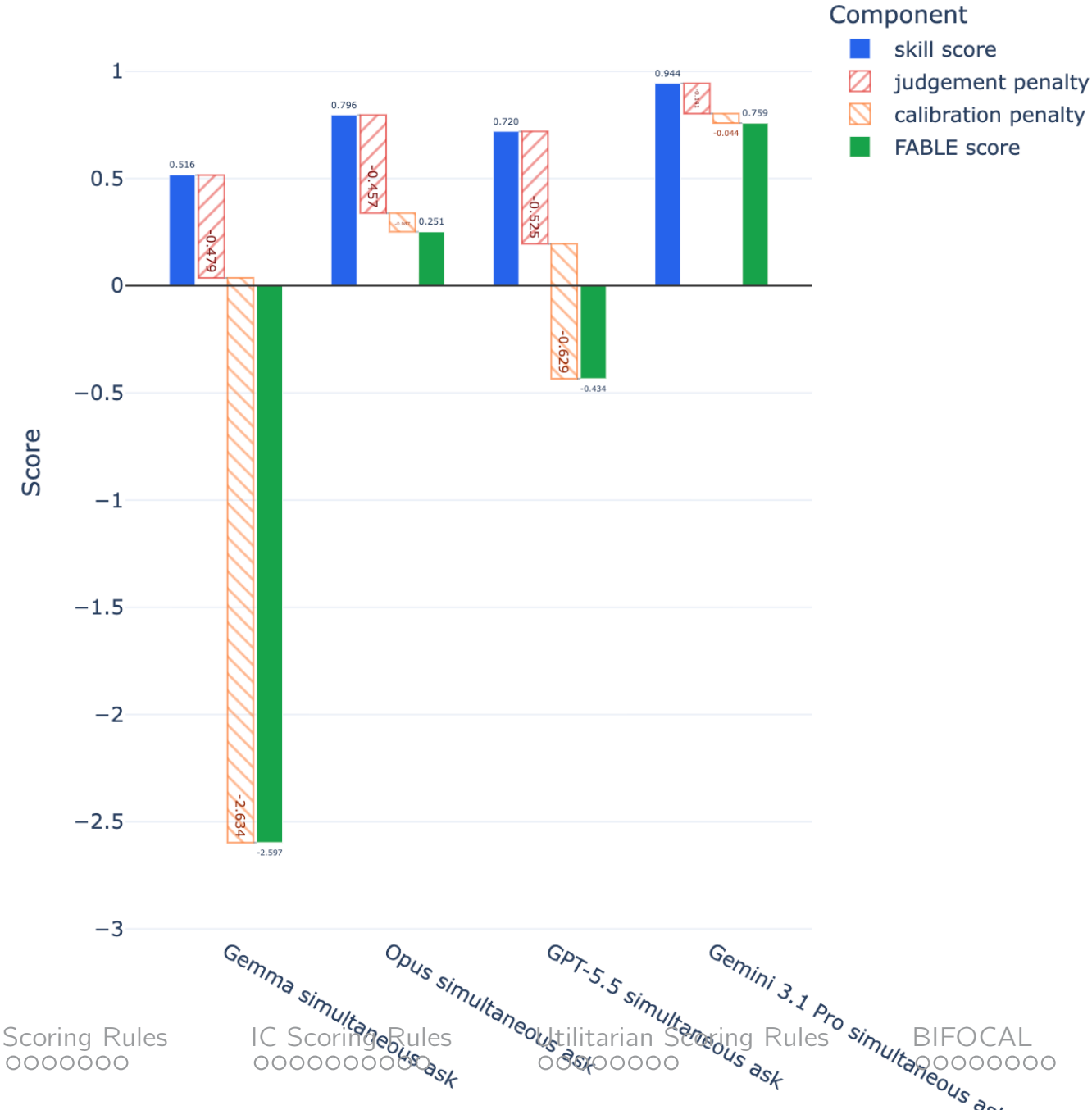
From this you might conclude that frontier models are as good or better at humans at these sort of questions/tasks, and as a result might be expected to replace humans in these areas

But the claim of this paper is that skill scores don't measure an AI's actual usefulness, which should take into account skills and epistemics simultaneously

The goal of BIFOCAL is to measure an AIs actual usefulness

How do frontier models perform with BIFOCAL?

MMLU BIFOCAL Scores



Thank you for listening!

Questions, comments, or concerns?