Understanding Understanding:
A Pragmatic Framework Motivated by Large Language Models

Kevin Leyton-Brown[1,3] and Yoav Shoham[1,2]

[1]AI21 Labs
[2]Stanford University
[3]University of British Columbia

Abstract

Motivated by the rapid ascent of Large Language Models (LLMs) and debates about the extent to which they possess human-level qualities, we propose a framework for testing whether any agent (be it a machine or a human) understands a subject matter. In Turing-test fashion, the framework is based solely on the agent's performance, and specifically on how well it answers questions. Elements of the framework include circumscribing the set of questions (the "scope of understanding"), requiring general competence ("passing grade"), avoiding "ridiculous answers", but still allowing wrong and "I don't know" answers to some questions. Reaching certainty about these conditions requires exhaustive testing of the questions which is impossible for nontrivial scopes, but we show how high confidence can be achieved via random sampling and the application of probabilistic confidence bounds. We also show that accompanying answers with explanations can improve the sample complexity required to achieve acceptable bounds, because an explanation of an answer implies the ability to answer many similar questions. According to our framework, current LLMs cannot be said to understand nontrivial domains, but as the framework provides a practical recipe for testing understanding, it thus also constitutes a tool for building AI agents that do understand.

1 Introduction

The question of whether machines can understand is as old as computing. Although Lady Ada Lovelace didn't explicitly address this question, her remarks on what a century later came to be called computers [Menabrea, 1843] suggest that she would have answered negatively. Much later, Searle [1980] certainly did answer negatively with his Chinese Room argument. The issue has now returned with renewed force in the context of Large Language Models (LLMs), with people questioning whether LLMs really understand their subject matter. To some, modern LLMs are so advanced so that they can be said to be approaching human-like intelligence, or at least show sparks thereof [Bubeck et al., 2023]. To others, LLMs may be impressive and useful artifacts, but inherently cannot understand, and only mimic it superficially [Bender and Koller, 2020, Chomsky, 2023]. These latter sentiments tend to based on arguments such that the systems engage in mere statistical calculations rather than reasoning; they deal with syntax but lack semantics; they lack a "world model"; they lack "communicative intent"; they are information-poor relative to the rich sensory information available to a person embedded in the real world; they lack emotions, "qualia" or consciousness; and so on. (We discuss these skeptical views, as well as ones more sympathetic to LLMs, in Section 2.)

It's not hard to sympathize with the intuitions underlying the skeptical arguments, and even agree with some of the conclusions (e.g., that there is much that current LLMs can't be said to understand), but we find the arguments themselves weak for two reasons. First, they are vague and appeal to ill-defined concepts. For example, what would constitute satisfactory semantics? Would a mapping from linguistic constructs to any knowledge graph [Wikipedia, 2024] do the trick? And second, we find these claims to be assumptive assertions without explanatory power. You assert that a world model is needed; I assert not; can we go have lunch now?

The underlying problem is that the discussion has lacked a clear definition of understanding. Providing a mathematical definition of understanding is this paper's first technical contribution (see Section 3). We take an approach akin to the Turing Test [Turing, 1950], which identified intelligence with behavior indistinguishable from human in a question-answering task. Roughly, we define a domain by a set of questions, and say that understanding the domain consists of being able to answer these questions in a satisfactory way. "Satisfactory" has two aspects. The first is demonstrating *general competence*: assigning a score to each answer, we require that the average score across all questions be above a given, relatively high (for example, .7) threshold (the "passing grade"). The second aspect is *avoiding ridiculousness*, which is defined in two steps. First, we define a ridiculous answer as one receiving a particularly low score (here we fix it at 0). Then, we require that the probability of a ridiculous answer be negligible (for example, below .001). Depending on the domain, answering "I don't know" can be acceptable: often not as good as providing a substantive answer, indeed sometimes ridiculous, but sometimes (if infrequently) the perfect answer.

It might seem that a question-answering framework can be used only to *assess* understanding, but we argue that it is also a way to *define* the concept. In doing so, we appeal to the philosophical tradition of Pragmatism, which argues that any concept's meaning can derive solely from its observable consequences. For more discussion, see Section 2.

Our paper's second key contribution (in Section 4) is to answer the following question: How can we ever realistically conclude that any agent (be it another person or an LLM) understands anything? Although our definition of understanding implies an obvious way of assessing whether an agent understands—namely, asking every question in the scope—all scopes of interest are so large (indeed, often infinite) that it is not possible in practice to verify our criteria with certainty. The natural alternative is to forego certainty and settle for obtaining high confidence, by asking a reasonable number of questions taken at random from the scope. We make these ideas formal too, appealing to concentration inequalities to investigate the sample complexity of testing understanding.

Unfortunately, this analysis shows that the number of samples required to reach acceptable confidence bounds can be high, especially due to the non-ridiculousness criterion. Our third key contribution is to show how *explanations* can mitigate the problem (see Section 5). In a nutshell, when an answer is coupled with an explanation, it effectively provides answers to a large number of similar questions one could have asked but didn't. We define this formally and show that it can meaningfully reduce the number of required samples. (This also provides intuition for why explanations play a key role in educational settings.)

A final introductory remark on "why care". One sometimes hears that AI practitioners interested in building intelligent machines needn't waste time on idle philosophizing if the agent does a good enough job. A similar sentiment used sometimes to be expressed in AI about the precise definition of causality (see Pearl [2000] for a comprehensive treatment). In the case of causality the issue is pretty much settled by now, and we believe here too practitioners should care. As will be seen, the definition of understanding here is directly related to shortcomings of current LLMs, including being ridiculous when they're not brilliant, not being able to say "I don't know", and not providing explanations of their output. The definition of understanding here not only sheds light on these shortcomings, but also points to ways in which to avoid them; as we discuss in the last section, our framework has practical implications for both evaluating and designing AI agents. (This is beside the fact that, in general, a little philosophizing won't hurt us engineers, particularly in this age in which AI is impacting society so profoundly.)

2 Related work

The topic of understanding is broad, and writing on it spans areas as diverse as philosophy, education, religion, and more. It is impossible to discuss all this work, but we provide some highlights.

As was mentioned in the introduction, this paper aims not only to explain how to assess understanding, but to also define the concept. Our approach falls in the camp of *pragmatism* within philosophy—a rich and complex tradition, whose relevant part is summarized in the so-called Pragmatic Maxim of Peirce [1878]:

> Consider what effects, that might conceivably have practical bearings, we conceive the object of our conception to have. Then, our conception of these effects is the whole of our conception of the object.

In other words, for a concept to have meaning, it must have practical consequences, and if two concepts have the same practical consequences, then they are effectively the same concept. According this principle, seeking some deeper notion of what understanding "actually is", such as by probing the internal state of the agent or its other implementation details, provides no useful information beyond the observed agent's competence in answering questions. If all behavioral evidence aligns with the claim that an agent understands, we should accept the claim. Our approach is clearly at odds with that of Potts et al. [2021]. They point to three metaphysical (or ontological) stances: *internalism* (inspecting the internals of the system against some notion of what it means to understand), *referentialism* (mapping objects in the system to other, externally-defined concepts), and pragmatism. They critique the pragmatic approach, arguing that tests (be they the Turing Test or various benchmarks in the AI community) are fallible, susceptible to manipulation and over-fitting, and find the other two approaches to be more definitive. We clearly take the opposite position. We find that although internalism and referentialism capture valid intuitions, they are too ill-defined to be applied rigorously in any setting. We do agree that the empirical tests mentioned and critiqued by Potts et al. [2021] are not the right ones for assessing understanding, and suggest criteria that are. For a broader discussion within philosophy of understanding, with an emphasis on its distinction from 'knowing' and a strong epistemological bent, see Grimm [2021].

We're not aware of previous work that introduced a rigorous theory of understanding—pragmatic or otherwise—that could be applied to answer the question of whether LLMs understand. However, there is certainly prior relevant (if mostly informal) writing that deserves discussion. This includes two pieces that share the primary title of this paper [Wiggins and McTighe, 2005, Wurman, 2017]. Drawing from education theory, Wiggins and McTighe [2005] focus on the educator's goal of achieving in the learner true understanding versus memorized, rote knowledge; or, in their words, "...safeguard the distinction between 'understand' and 'know the right answer when prompted'" (p.11). To them, the distinction hinges on being able to relate different concepts, to apply the knowledge in a broad set of circumstances, and to supply explanations. Three of their quotes from the literature capture these intuitions well.

> To grasp the meaning of a thing [...] is to see it in its relations to other things: [...] what consequences follow from it, what causes it, what uses it can be put to [...] Meanings are general because [they are] applicable in a wide variety of different instances in spite of their difference.[Dewey, 1933]

> Students should not be able to solve the new problems merely by remembering the solution to or the precise method of solving a similar problem in class. [...] Understanding is the ability to marshal skills wisely and appropriately [...] Doing something correctly, therefore, is not, by itself, evidence of understanding. [...] To understand is often being reflected in being able to explain *why*... [Bloom et al., 1956]

> Teaching specific topics or skills without [...] a broader fundamental structure within a field of knowledge is uneconomical.[Bransford et al., 2000]

Our framework can be thought of as concretizing and formalizing some of these intuitions.

Wurman [2017] is different—an epic, associative collection of evocative short essays on the nature of knowledge, understanding, learning, and other aspects of the human experience. It's not directly relevant to the current discussion, and we reference it here primarily because of the common title. But it does contain some relevant intuitions, for example in this quote:

> The most essential prerequisite to understanding is to be able to admit when you don't understand something. (p.15)

As we'll see, these thoughts are directly related to our insisting on defining the scope of understanding and the inclusion of the "I don't know" answer in our framework.

Within AI proper there has also been much writing relevant to the topic. Some researchers lean towards ascribing a measure of understanding even to current LLMs. This includes Bubeck et al. [2023], who argue on experimental grounds that GPT-4 shows "sparks of artificial general intelligence (AGI)" (ill-defined as the term is). Similarly, in a dialogue between two prominent researchers [Hinton and Ng, 2023], both seem to agree that modern LLMs exhibit a measure of understanding, encoded in neural networks' internal structures. Arcas [2022] explicitly argues that LLMs should cause us to reevaluate our notion of understanding, and that "statistics do amount to understanding, in any falsifiable sense".

Most researchers are more circumspect on the topic, offering varying degrees of skepticism. Some, such as Bender and Koller [2020], are completely dismissive of the idea that LLMs can exhibit any understanding at all. Somewhat less opinionated but still quite skeptical, Marcus and Davis [2020, 2023] meticulously chronicle telltale mistakes of LLMs that, to the authors, indicate lack of understanding. Some authors point to specific technical elements that in their view are missing from LLMs. To pick two examples, Russin et al. [2020] point to the higher-level functionality of the prefrontal cortex and suggests that AI systems should learn from its structure by way of achieving out-of-distribution robustness; and Xia et al. [2021] argue that learning in neural networks cannot reconstruct causal structures of the kind defined by Pearl [2000].

Some of the most interesting writing on the topic presents new questions or perspectives, rather than offering highly opinionated answers. For example, West et al. [2023] distinguishes between the ability to produce expert output and the ability to understand the output, and creatively points to how LLMs and humans may diverge in this regard.

Of particular note is a comprehensive survey of the AI literature on understanding by Mitchell and Krakauer [2023], which includes some of the above references and many others. This survey features clearly written commentary on the literature and the diverging views within it. The survey does not offer its own definition of understanding or how it is to be evaluated, but rather concludes that there is a need for defining "different modes of understanding", with the possibility that understanding by machines will be different from that of humans. In this paper we propose a uniform framework for understanding understanding, which can be instantiated in different ways by varying the questions, answers, and explanations. As Mitchell and Krakauer [2023] note, the topic of whether machines can understand is still being actively debated. Our paper will not settle the debate, but hopefully will help move the conversation forward and put it on more solid ground.

3 Defining Understanding

To define understanding, we begin by observing that it is never said simply that someone understands; there is always an explicit or implied subject matter that they do or don't understand. You understand arithmetic, you understand magnetism, you understand Buddhism, you understand human emotions. Let's call that which you understand the *scope of understanding*, or simply the scope. We define a scope by a set of questions with a distribution over them, and understanding by how well one is able to answer those questions ("well" being defined below).[1] Of course, a scope cannot be inferred by a term such as "arithmetic"; we would judge whether a 10th grader understands arithmetic differently than whether a Field Medal winner understands it. So the scope will be defined by a process that generates the questions that comprise the scope. This process could be simply to sample from a distribution over an explicit set of questions, or a more involved procedure (e.g., draw two real numbers from [0, 1,000,000], draw an arithmetic operation from $\{+, -, *, /\}$) defining a large or even infinite set of questions. While we could define any set of questions as a scope, our framework is useful for sets that correspond to a conceptually natural domain, and natural domains are never small. Another issue is robustness to surface perturbations. If a question is in the scope, so should be its natural rephrasings, so that answers aren't overfit to specific surface formulations.

[1] To anticipate a possible objection: One could imagine an argument that a person might understand a question perfectly but still not be able to answer it. But this is being unclear about the scope of understanding; there is no understanding without knowledge. If you can't solve any addition problem then you don't understand arithmetic, and if you can't answer any question about WWII then you don't understand world history.

There is the practical question of how to specify such a large set in practice, especially a set whose boundaries may be fuzzy. But here we take the scope as given.

Definition 3.1. *A* domain *is a tuple* (Q, Δ_Q, A, S) *where*

- Q *is a set of questions.*
- Δ_Q *is a probability distribution over* Q*.*
- A *is a set of answers with* $idk \in A$ *(*idk *standing for "I don't know").*
- $S : Q \times A \rightarrow [0, 1]$ *is a scoring function for answers to given questions.*

When $S(q, a) = 0$ we say that a is a ridiculous answer to q.[2] We do not constrain the values of $S(q, idk)$. Often such scores will be low; if you profess ignorance about the value of $1 + 1$ in the scope of arithmetic you get a very low score, indeed perhaps even a score of 0. But answering idk to the question of whether $P = NP$ in the scope of computational complexity gives you a high score. That said, domains in which idk gets high scores for all or most questions will rarely be interesting.[3]

Definition 3.2. *An* understanding criterion *for a domain is a pair* (PG, RID) *where*

- $PG \in (0, 1)$ *is the required level of overall competence, or the "passing grade".*
- $RID \in (0, 1)$ *where* $RID << PG$ *is a global ridiculousness threshold, the maximum allowable probability of obtaining a ridiculous answer when sampling a question from* Δ_Q*.*

The intuition should be that PG is relatively high (for example, .7) and RID is so low as to be negligible. One way of choosing a concrete value for RID is to specify a test length n_t on which the agent should be guaranteed to avoid making a ridiculous answer with high probability $1 - \delta$. Then we can write $(1 - RID)^{n_t} \geq 1 - \delta$. Solving for RID, we obtain $RID = 1 - \exp\left(\frac{\ln(1-\delta)}{n_t}\right)$. For example, if $\delta = 0.05$ and $n_t = 100$, $RID = 0.00052$; if $n_t = 1000$, $RID = 0.000052$.

Definition 3.3. *Given a domain and an understanding criterion as above, a question-answering (QA) system* $ans : Q \rightarrow A$ *is said to exhibit understanding within scope* Q *if and only if:*

(Overall passing grade) $\mathbb{E}_{q \sim \Delta_Q}\ S(q, ans(q)) \geq PG$.

(Ridiculous answers are rare) $P_{q \sim \Delta_Q}(S(q, ans(q)) = 0) \leq RID$.

That is, the mean score according to Δ_Q must clear the threshold PG, and the probability with which the agent generates ridiculous answers must not exceed the (very low) limit RID. Relative to this definition, and given reasonable values of PG and RID, modern LLMs fall short. In very circumscribed scopes they do pass the criteria; for example, most do well when asked to identify the capital cities of countries. But they invariably fail in larger or more demanding scopes, witness the example in Appendix C, which, empirically speaking, is not a rare isolated case (the example will also prove useful when we discuss explanations later in the paper).

We make a few remarks on this last definition. First, observe that Δ_Q need not weight every question equally. This non-uniform distribution can reflect e.g., relative prevalence, importance, or difficulty of questions. Second, by their probabilistic nature, LLMs can provide a probability distribution over answers (i.e., when we do not sample with temperature 0); one might therefore want to average over system executions, over top-k answers, or something else. The function S can easily be modified to return an expected score and all of our results that follow carry through directly. Finally, it is possible to imagine a more draconian version of our ridiculousness condition which would set $RID = 0$, that is, disallow any ridiculous answer, but there are several problems with this. First, no finite number of samples would suffice to distinguish between the cases where an agent never gives ridiculous answers and where it gives them with some infinitesimally small probability; thus,

[2] Observe that assigning the same score to all ridiculous answers prevents us from assessing degrees of ridiculousness. We made this modeling choice because our framework treats all ridiculous answers in the same way, and so we have no need for this additional flexibility. It would be straightforward to extend the framework to allow a range of ridiculous scores (e.g., all scores below 0.01) at the expense of additional complexity.

[3] Rarely, but not never. For example, it could be interesting to test an agent's understanding of the limits of its own knowledge via a scope consisting entirely of facts that the agent has no way of knowing, such as current events that occurred after the training data cutoff date.

the draconian definition would preclude high-probability confidence bounds. Second, that definition is counterintuitive: we do conclude that other people understand when they perform well but very occasionally have idiosyncratic lapses. Finally, existing LLMs that exhibit ridiculous behavior do so more than a negligible fraction of the time, making the less stringent condition sufficient.

4 Assessing Understanding via Independently Sampled Questions

Given the framework just laid out, the only way to verify with certainty that an agent understands within a scope is to test it on all questions and see if the answers satisfy the conditions in Definition 3.3. Even if the agent answers a large number of questions well, any questions we omit may reveal a lack of understanding; e.g., they might all yield ridiculous answers. Testing all questions is impractical for any scope of interesting size and impossible for infinite-sized scopes. In such cases, however, it is possible to gain high *confidence* that an agent understands by assessing its performance on many independent samples taken randomly from the distribution of questions. Intuitively, when we choose questions at random, the likelihood that the agent's performance is qualitatively different on the unasked questions as on the asked questions falls with the number of asked questions.

Our goal is to identify a test for understanding based on the agent's answers to independently sampled questions, and to prove that the probability that this test gives incorrect answers falls (ideally, quickly) with the number of questions asked. We can then investigate how many randomly chosen questions we need in order to obtain high confidence in the test's results. As we shall see, the number of questions required is large but not completely implausible. In the next section we will see how the notion of explanations can help to further reduce the required number of questions.

Formally, let n be the number of questions we have asked, each sampled independently from Δ_Q. Each question q_i receives score $s_i = S(q_i, ans(q_i)) \in [0, 1]$. Let $\hat{s}$ denote our sample estimate of the score $\frac{1}{n}\sum_{i=1}^{n} s_i$ and let s denote the true score $\mathbb{E}_{q\sim\Delta_Q}\ S(q, ans(q))$. Let $r_i = 1$ if $s_i = 0$ and 0 otherwise. Let $\hat{r}$ denote the fraction of the agent's answers that were ridiculous: $\frac{1}{n}\sum_i r_i$, and let r denote the agent's true probability of giving a ridiculous answer, $P_{q\sim\Delta_q}(S(q, ans(q)) = 0)$.

We need a few mathematical preliminaries. First, define $d(x, y) = x \ln\left(\frac{x}{y}\right) + (1-x)\ln\left(\frac{1-x}{1-y}\right)$, replacing singularities ($\ln(0)$; division by zero) with limit values; see Definition A.1 in Appendix A. This function can be interpreted as the relative entropy (or KL divergence) between two Bernoulli distributions[4] having parameters $x \in [0, 1]$ and $y \in [0, 1]$.

Our testing procedure will appeal to two different functions. These do not have closed-form expressions, because d does not have a closed-form inverse. Nevertheless, they are easy to approximate numerically because $d(x, \cdot)$ is convex.

$$U(x, n, \delta) = \max\{y \in [0, 1] : d(x, y) \leq \ln(1/\delta)/n\} \tag{4.1}$$

$$L(x, n, \delta) = \min\{y \in [0, 1] : d(x, y) \leq \ln(1/\delta)/n\} \tag{4.2}$$

We can now state our procedure for testing whether an agent understands a given scope.

Procedure 4.1 (Testing Procedure Based on Independent Samples)**.** *This procedure takes two arguments: a number of samples n and a desired maximum failure probability δ.*

1. *Determine that the agent understands the scope of Q if:*

   (Good-Grade) $L(\hat{s}, n, \delta) \geq PG$*; and*
   (Good-Rid) $U(\hat{r}, n, \delta) \leq RID$.

2. *Determine that the agent does not understand the scope of Q if:*

   (Bad-Grade) $U(\hat{s}, n, \delta/2) < PG$*; or*
   (Bad-Rid) $L(\hat{r}, n, \delta/2) > RID$.

3. *Otherwise, draw no conclusion; a larger n (or larger δ) is required.*

[4]While $d(x, y)$ has an interpretation as the relative entropy between Bernoulli distributions, we can also simply consider it as an arbitrary function from $[0, 1] \times [0, 1] \to \mathbb{R}^+$. We will use the functions U and L, which depend on d, to upper bound the means of arbitrary (non-Bernoulli) distributions restricted to $[0, 1]$.

Theorem 1. *If Procedure 4.1 determines that an agent does or does not understand a given scope, this conclusion is correct with probability at least* $1 - \delta$*.*

Proof. See Appendix A. $\square$

Remarks. Observe that our result requires no assumptions about the number of questions in the scope or about the probability distribution Δ_Q; all that is needed is that questions are drawn independently. Note that *Good-Grade* and *Good-Rid* depend on δ but that *Bad-Grade* and *Bad-Rid* depend on $\delta/2$. This asymmetry arises because Step 2 asks whether either of two tests is true and so has two ways of failing, thus requiring multiple hypothesis correction. Step 1 asks whether *both* of two tests are true, and hence does not require correction. Finally, in Appendix B we explain why we leveraged the Chernoff bound, despite its lack of a closed-form expression, rather than the Hoeffding bound. In brief, while both bounds apply to our setting, Chernoff is much tighter when the true mean is close to 0 or 1.

Does this result enable the high-probability assessment of understanding within a reasonable number of questions? Let us consider some concrete numbers. Suppose that we are willing to tolerate an overall failure probability of 0.05. First, let us consider average scores that would be required for *Good-Grade* and *Bad-Grade* to hold as we vary the number of questions asked, summarized in Table 1. Observe that our bounds depend on the observed empirical distribution, with confidence intervals narrowing as $|0.5 - \hat{s}|$ grows. If the agent answers 100 questions and achieves a score of 0.9, it will pass if $PG < .811$ or fail if $PG > .963$; if it only achieves a score of 0.5, these thresholds change to passing if $PG < 0.379$ and failing if $PG > 0.634$, a gap of 0.25 vs 0.15. If it answers 10,000 questions and achieves a score of 0.9, it passes if $PG < 0.892$ and fails if $PG > 0.908$.

Now let us consider the tests *Good-Rid* and *Bad-Rid* and the value $RID = 0.00052$ derived in Section 3; see Table 2. Giving no ridiculous answers after answering 1,000 questions is insufficient to satisfy *Good-Rid* with high probability ($U(0, n, 0.05) \approx 0.0029913 > 0.00052$) but after it answers 10,000 questions without giving any ridiculous answers, it does satisfy *Good-Rid*. If the agent first gives 3 ridiculous answers and thereafter always gives non-ridiculous answers, our upper bound starts out much higher but converges (more slowly) towards 0 as $n \to \infty$. In this case *Good-Rid* no longer holds after 10,000 questions but does after 100,000. If the agent consistently produces a 1% empirical fraction of ridiculous answers, *Good-Rid* will never hold ($U(x, n, \delta) \geq x$ for all (n, δ)); *Bad-Rid* will not hold after 100 questions but will after 1,000 questions.

Overall, a good rule of thumb seems to be that 1,000 questions are often enough to assess average score, and that 10,000 questions are often appropriate for assessing the probability of giving ridiculous answers. If we are interested in more tightly bounding the likelihood of observing ridiculous answers, every order-of-magnitude increase in the number of questions for which we would like the agent to be guaranteed with high probability to give no ridiculous answer roughly corresponds to an order-of-magnitude decrease in RID, and hence (referring to the table) roughly to an order-of-magnitude increase in the required n. Are these reasonable numbers in practice? Certainly for human agents they are not; no school test can have this many questions. One could argue that they are more plausible for AI systems, but this too depends on the setting. If one assumes that the tests can be done in a completely self-supervised manner then perhaps they are, but this is a tenuous assumption. When a human assessment (of answer quality in general, and ridiculousness in particular) is needed, which arguably is the common case, then while the number of questions can be greater than when a person is the one answering them, it is still limited. The current cost of human annotation ranges between $5–50 per item; at $10 per assessment this brings the cost to six figures. Given that the process must be repeated as the set of questions evolves and as the model is retrained, this becomes prohibitive.

5 Explanations: Beyond Independent Sampling

We have seen that assessing understanding with acceptable confidence based only on answers to a sample of questions can require very large sample sizes, especially as we push for a small RID. We now show that when explanations[5] accompany answers, we can achieve better sample efficiency.

[5] One could replace the term 'explanation' with 'justification', but we stick to the former since it's the term most commonly used in connection with LLMs.

n	$L(0.9, n, 0.05)$	$U(0.9, n, 0.025)$	$L(0.5, n, 0.05)$	$U(0.5, n, 0.025)$
10	0.545253	0.999023	0.164322	0.861187
100	0.811171	0.962052	0.379423	0.633343
1,000	0.875192	0.923796	0.461356	0.542868
10,000	0.892497	0.907952	0.487763	0.513579
100,000	0.897662	0.902557	0.496130	0.504295
1,000,000	0.899264	0.900813	0.498776	0.501358

Table 1: Example values for bounding s in Procedure 4.1, $\delta = 0.05$. *Good-Grade* requires that the empirical average score exceed PG by at least $L(\hat{s}, n, \delta)$, and conversely *Bad-Grade* requires that PG exceed the empirical average score by at least $U(\hat{s}, n, \delta/2)$. The two pairs of columns thus represent the confidence intervals around empirical averages of $\hat{s} = 0.9$ and $\hat{s} = 0.5$ respectively.

n	$U(0, n, 0.05)$	$U(3/n, n, 0.05)$	$L(0.01, n, 0.025)$
10	0.2588656	0.6783535	0.0000000
100	0.0295130	0.0913315	0.0000933
1,000	0.0029912	0.0094020	0.0036846
10,000	0.0002995	0.0009429	0.0075333
100,000	0.0000300	0.0000943	0.0091693
1,000,000	0.0000030	0.0000094	0.0097321

Table 2: Example values for bounding r in Procedure 4.1 with $\delta = 0.05$. If the empirical fraction of ridiculous answers is zero, *Good-Rid* holds if $U(0, n, 0.05) \leq RID$. In the case that exactly three ridiculous answers are observed independently of n, *Good-Rid* holds if $U(3/n, n, 0.05) \leq RID$. In the case that $\hat{r} = 0.01$ for every n, *Bad-Rid* holds if $L(0.001, n, 0.025) > RID$.

The idea that assessing understanding is aided by eliciting explanations is quite familiar from educational settings, where homework and exam questions often end with "explain your answer". As noted in the (mostly informal) education literature we surveyed in Section 2, good explanations demonstrate knowledge of general principles and the ability to apply them broadly. Explanations have also received attention in the NLP community; e.g., one line of work quantifies the value of explanations as the performance gain they offer a student model seeking to simulate a teacher [Pruthi et al., 2022].

Explanations have many facets and come in different forms. A common form is the causal explanation, for example as defined by Pearl [2018], who defines three increasingly sophisticated explanatory levels (the "ladder of causation"). Another facet of causal explanations is that they can come at multiple granularities, as demonstrated in Feynman's [1983] beautiful response to the lay question of "why magnets attract". But causal explanations, while common and important, are a special case. For example, when we ask a student to explain their answer to a mathematical problem, we don't expect a causal explanation, but rather the steps in the derivation, or at least mention of the mathematical principles applied. In other cases, such as in response to factual questions (for example, "Is John an employee of Acme Corp?"), a good explanation is a reference to a definitive source of relevant information (for example, Acme's database of employees). Another source of authoritative information can be a a trusted person.[6] What is common to all of types of explanation is that they involve a *procedure* that is deemed authoritative and that could similarly be applied to many other questions. As a result, the agent can be given some amount of credit for those questions without these being asked explicitly; "an explanation is worth a thousand questions".

Formally, we augment our existing model so that, when we ask the agent a question, the agent sometimes couples its answer with a particular procedure that it used to derive the answer. When it does, we get evidence about how the agent would have answered similar questions. It's important that this be an unambiguous procedure.For example, we might prompt the LLM to use *chain-of-code* reasoning [Li et al., 2023], explicitly asking the LLM to produce code to answer the question. The

[6]One is reminded of Angluin's [1983] tongue-in-cheek list of bogus proofs techniques, which includes the category of 'proof by eminent authority' ("I saw Karp in the elevator and he said it was probably NP-complete").

explanation could also be a mathematical equation, a precise table being looked up, and so on. But it cannot be any string of words that purport to constitute an explanation; Appendix C provides an example of an LLM offering convincing-sounding explanations that aren't worth the metaphorical paper they're written on.

Formally, let $\mathcal{A} = \{\alpha_1, \ldots, \alpha_k\}$ be a set of procedures offered by the agent to explain how it derived answers to a corresponding set of questions. Each procedure α is potentially applicable to more than just the single question that prompted it. We assume that set of questions to which each α_i is applicable is known, denoting this set $Q_{\alpha_i} \subseteq Q$, and that $\forall i, j\ Q_{\alpha_i} \cap Q_{\alpha_j} = \emptyset$. We further assume that the answer generated by each explanation α_i achieves the same score on all questions to which it applies;[7] denote this score s_{α_i}. If $s_{\alpha_i} = 0$, the answers generated by the explanation are ridiculous for every $q_i \in Q_{\alpha_i}$, which we denote $r_{\alpha_i} = 1$; otherwise $r_{\alpha_i} = 0$. We define the set of "remaining" questions to which no procedure applies as $\bar{Q}$ and thus write $Q = \left(\bigcup_{\alpha_i \in \mathcal{A}} Q_{\alpha_i}\right) \cup \bar{Q}$. Let $p_{\alpha_i} = P_{q \sim \Delta_Q}(q \in Q_{\alpha_i})$, the probability that a random question from the original distribution Δ_Q will belong the set of questions covered by explanation α_i. Let $\Delta_{Q_{\alpha_i}}$ denote the renormalized restriction of probability distribution Δ_Q to domain Q_{α_i}: for all $q_i \in Q_{\alpha_i}$, $\Delta_{Q_{\alpha_i}}(q_i) = \frac{1}{\sum_{q_j \in Q_{\alpha_i}} \Delta_Q(q_j)} \cdot \Delta_Q(q_i)$.

Given $\mathcal{A}$, what should we believe about the agent's performance on our original test Q? Intuitively, while a procedure α can tell us more than an answer to a single question, we do not obtain the set Q_α via independent sampling, so we cannot directly incorporate it into our previous analysis. Imagine that in addition to $\mathcal{A}$ we have a set $\bar{X}$ of additional answers to questions sampled independently from $\bar{Q}$, denoting $|\bar{X}|$ as $\bar{n}$, and that the agent does not give an explanation for any question in $\bar{X}$. In this section our key question will be: How much better off are we having received explanations $\mathcal{A}$ than in the situation where we did not receive these explanations and took $\bar{n}$ samples from Q?

In our analysis below we make two substantive assumptions: that we believe that the agent claiming to have used explanation α in fact did, and (even more optimistically) that the agent would apply it reliably to every question in Q_α. Procedure 5.1 tests an agent's understanding in this setting.

Procedure 5.1 (Testing Procedure Based on Samples and Explanations)**.** *This procedure takes two arguments: a number of samples n and a desired maximum failure probability δ.*

1. *Determine that the agent understands the scope of Q if:*

   (Good-Grade-Exp) $\left(\sum_{\alpha_i \in \mathcal{A}} p_{\alpha_i} \cdot s_{\alpha_i}\right) + \bar{p} \cdot L(\hat{\bar{s}}, \bar{n}, \delta) \geq PG$*; and*

   (Good-Rid-Exp) $\left(\sum_{\alpha_i \in \mathcal{A}} p_{\alpha_i} \cdot r_{\alpha_i}\right) + \bar{p} \cdot U(\hat{\bar{r}}, \bar{n}, \delta) \leq RID.$

2. *Determine that the agent does not understand the scope of Q if:*

   (Bad-Grade-Exp) $\left(\sum_{\alpha_i \in \mathcal{A}} p_{\alpha_i} \cdot s_{\alpha_i}\right) + \bar{p} \cdot U(\hat{\bar{s}}, \bar{n}, \delta/2) < PG$*; or*

   (Bad-Rid-Exp) $\left(\sum_{\alpha_i \in \mathcal{A}} p_{\alpha_i} \cdot s_{\alpha_i}\right) + \bar{p} \cdot L(\hat{\bar{r}}, \bar{n}, \delta/2) > RID.$

3. *Otherwise, draw no conclusion; a larger n (or larger δ) is required.*

Theorem 2. *If Procedure 5.1 determines that an agent does or does not understand a given scope, this conclusion is correct with probability at least $1 - \delta$.*

Proof. See Appendix A. □

Respectively comparing each of our conditions (*Good-Grade-Exp* vs *Good-Grade*, etc) answers the question we asked earlier about the quantitative power of explanations: our confidence bounds shrink in proportion to the fraction of Q covered by procedures in $\mathcal{A}$. Stated another way, we would need to take between a factor of $1/\left(1 - \sum_{\alpha_i \in \mathcal{A}} p_{\alpha_i}\right)$ and $1/\left(1 - \sum_{\alpha_i \in \mathcal{A}} p_{\alpha_i}\right)^2$ more samples to achieve the same tightening of our confidence bounds, with the rate of tightening depending on the true distributions $\hat{\bar{s}}$ and $\hat{\bar{r}}$. For example, imagine that Q contains an equal number of problems about multiplication and about literature, that the agent gives us its procedure α for performing multiplication, and that the agent is 33% accurate at answering questions about literature. Thereafter it suffices to sample only literature questions and each counts in our bounds for average grade as much as $1/(1 - 0.5)^2 = 4$ questions from the original distribution and in our bounds for the probability

[7] Our results can be generalized beyond this case at the cost of notational complexity.

of giving ridiculous answers about as much as $1/1 - 0.5 = 2$ questions, in both cases yielding a bound exactly half as wide. This dramatic improvement in the bound combines two effects. First, our knowledge of α eliminates all uncertainty about s_α and r_α; second, we devote all samples to $\bar{Q}$ rather than devoting a p_a fraction of samples to Q_α.

Several extensions and refinements are possible. We have so far described a setting in which all explanations and answers are offered before our analysis occurs. We can straightforwardly construct an online procedure, with the agent offering either individual answers or explanations as it chooses and our bounds updating as we go. Before we have asked any questions, we begin with $Q_\alpha = \emptyset$, $\bar{Q} = Q$, and $\mathcal{A} = \emptyset$. We always sample only from $\bar{Q}$, storing these samples in $\bar{X}$. When we receive an explanation α_i, we add it to $\mathcal{A}$, subtract the set covered by Q_{α_i} from $\bar{Q}$, and discard any previous samples from $\bar{X}$ that do not pertain to the newly restricted $\bar{Q}$ (allowing for the possibility that the agent may later offer an explanation that covers previous individual answers). Theorem 2 then applies directly at every point in this process. More demanding is relaxing the assumption that the agent would apply its supplied procedure to all other applicable questions. We discuss this in the next section.

6 Discussion

This paper addressed the question of how one is to judge whether an agent such as an LLM understands a subject matter. We note that our focus on question answering excludes other behaviors that could commonly be viewed as demonstrating understanding. For example, we could have said that the agent (or our dog) understands our commands based on whether it follows them. Or, we could have said that you show that you understand a person's sorrow with your kind look and comforting embrace. In general, people use language in flexible ways, and terms such as 'understanding' are used in multiple senses. We focus on question answering since it is the closest in spirit to the discussion of whether computer systems such as LLMs understand, which has been the impetus for this paper.

A note about the potential danger of omitting "obvious" questions. People have vast background—sometimes called "common sense"—knowledge, which the person formulating the list of questions would not think of including since they are so obvious. There could be the concern that the AI agent might answer the more sophisticated questions well, but would have failed on these obvious questions had they been asked. We're not sure how serious this concern is, since arguably without this background knowledge the agent would eventually fail on the more sophisticated questions. But it is something to watch out for.

The framework discussed here can be expanded in various ways. Some of these were discussed in previous sections, but there are others. For example, rather than being compared to the expected score, the passing grade could be applied to another norm, such as the max. This is admittedly an edge case, but is applicable, for example, if all questions are open questions in the field; if you solved P/NP no further evidence is required to show that you understand the theory of computing. Or, as another example, the framework as discussed is a "static" one, in two senses. First, it asks whether an agent in a given state understands something. It does not discuss the dynamics of understanding, for example how an agent can come to understand things, such as through a series of Socratic questions posed to it. Second, it assumes a given scope, but one can imagine realistic situations in which an initially specified scope is seemed insufficient and is expanded as a result of interaction with the agent that uncovered new relevant questions that had not been anticipated. Both are interesting topics for further research.

Still, already as laid out here, the framework is a good basis for discussing the relationship between LLMs and understanding, and the framework can be applied to both the analysis and design of AI agents. From the analysis point of view, the framework suggests that beside testing the AI agent against a fixed test suite such as HELM [Bommasani et al., 2022] or MMLU [Hendrycks et al., 2021], the user of the agent should require that the agent virtually never provide ridiculous responses. Furthermore, the ridiculous answers need not be entirely specified in advance, which could both be infeasible, and could invite overfitting, but rather can be presented or expanded post-hoc on an ongoing basis. To be sure, this places a burden on the agent builder, who is held accountable to unforeseen future tests, but that does not seem unfair if one is claiming understanding by the agent one is offering. From the design as well as the analysis point of view, the framework suggests two additional things. First, that the agent admit ignorance when it does not have the answer, provided pleading ignorance is not ridiculous. And second, when it does not plead ignorance, that answers be complemented by explanations, which consist of rigorous, accepted procedures by which the answers were derived (be they a simple database lookup or a more complicated program).

Clearly, according to the account of understanding laid out in this paper, current AI systems—in particular, LLMs—can't be said to understand nontrivial domains. The systems are unreliable, occasionally downright ridiculous, unpredictable, opaque, and non-explainable. This is not only of intellectual interest but is intimately tied to the practical limitations industry is experiencing in deploying them broadly.

As we've mentioned, our approach is similar to that of Turing [1950], though the details are different; in this connection, see Grosz [2012]. But Turing's seminal article is relevant in another way as well. The jury is out on which set of methods holds the greatest promise in meeting the criteria for understanding laid out here. Some will surely believe all it would take is scale—bigger models, larger and better training data. Others will believe that the path forward involves the so-called neuro-symbolic approach, which combines the epically knowledgeable and creative yet stochastic LLMs with classical methods based on semantically clear data structures and human-interpretable algorithms. Time will tell, not only which approach triumphs, but also whether the goal is attainable at all. Can AI achieve human-level understanding, or will that be forever out of reach? Here we think it's instructive to recall the words of Turing in the same article:

> It is customary to offer a grain of comfort, in the form of a statement that some peculiarly human characteristic could never be imitated by a machine. I cannot offer any such comfort, for I believe that no such bounds can be set.

We believe this is true in particular of understanding.

Acknowledgements

We thank Noga Alon for feedback on the probabilistic analyses; Barak Lenz for pointing out Feynman's "why magnets attract" example; Moises Goldszmidt for the observation about the danger of omitting obvious examples; Scott Aaronson for the observation that solving P/NP would be sufficient to demonstrate understanding of the theory of computing; and Chris Ré and attendants of the Stanford HAI colloquium [Shoham, 2023], in which some of the ideas in this paper were discussed, for raising the question of the dynamics of understanding. Many additional thanks to Ernie Davis, Mansi Gupta, Joe Halpern, Percy Liang, Judea Pearl and Danish Pruthi for their useful comments.

Bibliography

Dana Angluin. How to prove it. *ACM SIGACT News*, 15(1):16–17, 1983.

Blaise Aguera y Arcas. Do large language models understand us? *Daedalus*, 151(2), 2022.

Emily M Bender and Alexander Koller. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the Association for Computational Linguistics*, pages 5185–5198, 2020.

Benjamin S. Bloom, Max D. Engelhart, Edward J. Furst, Walker H. Hill, and David R. Krathwohl. *Taxonomy of educational objectives: The classification of educational goals. Vol.1 of Handbook: Cognitive domain.* David McKay Company, New York, 1956.

Rishi Bommasani, Percy Liang, and Tony Lee. Language models are changing AI: The need for holistic evaluation, 2022. URL `https://crfm.stanford.edu/2022/11/17/helm.html`.

John D. Bransford, Anne L. Brown, and Rodney R. Cocking. *How People Learn: Brain, Mind, Experience, and School.* National Academy Press, Washington DC, 2000.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. Sparks of artificial general intelligence: Early experiments with GPT-4, 2023.

Noam Chomsky. The false promise of ChatGPT. New York Times, March 2023. URL `https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html`.

John Dewey. *How We Think.* Dover Publications, revised edition, 1933.

Richard Feynman. Magnets. Fun To Imagine, BBC, 1983. URL `https://www.youtube.com/watch?v=Dp4dpeJVDxs`.

Aurélien Garivier and Olivier Cappé. The KL-UCB algorithm for bounded stochastic bandits and beyond. In *Proceedings of the 24th annual conference on learning theory*, pages 359–376, 2011.

Stephen Grimm. Understanding. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Summer 2021 edition, 2021.

Barbara Grosz. What question would Turing pose today? *AI Magazine*, 33(4):73–73, 2012.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding, January 2021. URL `https://arxiv.org/abs/2009.03300`.

Geophrey Hinton and Andrew Ng, 2023. URL `https://www.linkedin.com/posts/andrewyng_had-an-insightful-co`

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms.* Cambridge University Press, 2020.

Chengshu Li, Jacky Liang, Andy Zeng, Xinyun Chen, Karol Hausman, Dorsa Sadigh, Sergey Levine, Li Fei-Fei, Fei Xia, and Brian Ichter. Chain of code: Reasoning with a language model-augmented code emulator, 2023.

Gary Marcus and Ernest Davis. GPT-3, bloviator: OpenAI's language generator has no idea what it's talking about. *Technology Review*, 2020.

Gary Marcus and Ernest Davis. Large language models like ChatGPT say the darnedest things. *The Road to AI We Can Trust*, 2023.

Luigi Federico Menabrea. Sketch of the analytical engine invented by Charles Babbage. with notes by the translator, 1843. Translated by Ada Lovelace.

Melanie Mitchell and David C. Krakauer. The debate over understanding in AI's large language models. *Proceedings of the National Academy of Sciences*, 120(13), 2023.

Judea Pearl. *Causality models, reasoning and inference*. Cambridge University Press, 2000.

Judea Pearl. *The book of Why: The new science of cause and effect*. Basic Books, 2018.

Charles Sanders Peirce. How to make our ideas clear. *Popular Science Monthly*, 12(Jan.):286–302, 1878.

Christopher Potts, Thomas Icard, Eva Portelance, Dallas Card, Caitlyn Zhou, and John Etchemendy. *Philosophy of Understanding*, chapter 2.6. Center for Research on Foundation Models (CRFM), Stanford Institute for Human-Centered Artificial Intelligence, (HAI), Stanford University, 2021. in On the Opportunities and Risks of Foundation Models (P. Liang et al.).

Danish Pruthi, Rachit Bansal, Bhuwan Dhingra, Livio Baldini Soares, Michael Collins, Zachary C Lipton, Graham Neubig, and William W Cohen. Evaluating explanations: How much do explanations from the teacher aid students? *Transactions of the Association for Computational Linguistics*, 10:359–375, 2022.

Jacob Russin, Randall C O'Reilly, and Yoshua Bengio. Deep learning needs a prefrontal cortex. *Work Bridging AI Cogn Sci*, 107(603-616):1, 2020.

John Searle. Minds, brains and programs. *Behavioral and Brain Sciences*, 3:417–57, 1980.

Yoav Shoham. Understanding understanding (HAI seminar), 2023. URL `https://www.youtube.com/watch?v=ZNi69FHe3y8`.

Alan Turing. *Computing machinery and intelligence*. Mind Volume LIX, Issue 236, 1950.

Peter West, Ximing Lu, Nouha Dziri, Faeze Brahman, Linjie Li, Jena D. Hwang, Liwei Jiang, Jillian Fisher, Abhilasha Ravichander, Khyathi Chandu, Benjamin Newman, Pang Wei Koh, Allyson Ettinger, and Yejin Choi. The generative AI paradox: "What it can create, it may not understand". *arXiv preprint arXiv:2311.00059*, 2023.

Grant Wiggins and Jay McTighe. Understanding Understanding. In *Understanding by Design*, chapter 2. Vanderbilt University Center for Teaching, 2005.

Wikipedia. Knowledge Graph, Accessed March 26, 2024. URL `https://en.wikipedia.org/wiki/Knowledge_graph`.

Richard Saul Wurman. *Understanding Understanding*. Richard Saul Wurman and Jack Dangermond, 2017.

Kevin Xia, Kai-Zhan Lee, Yoshua Bengio, and Elias Bareinboim. The causal-neural connection: Expressiveness, learnability, and inference. *Advances in Neural Information Processing Systems*, 34:10823–10836, 2021.

A Deferred Technical Material

Definition A.1 (Relative entropy between Bernoulli distributions)**.** *Let $d(x, y)$ denote the relative entropy between two Bernoulli distributions having parameters $x \in [0, 1]$ and $y \in [0, 1]$:*

$$d(x,y) = \begin{cases} x \ln\left(\frac{x}{y}\right) + (1-x)\ln\left(\frac{1-x}{1-y}\right) & x \in (0,1) \text{ and } y \in (0,1) \\ \ln(1/(1-y)) & x = 0 \text{ and } y \neq 1 \\ \ln(1/y) & x = 1 \text{ and } y \neq 0 \\ \infty & x \neq 0 \text{ and } y \in \{0,1\} \\ \infty & x \in \{0,1\} \text{ and } y = 1-x \end{cases}$$

Our proof of Theorem 1 relies on five lemmas. We begin with a Chernoff bound from the literature, following Lattimore and Szepesvári [2020] but using notation aligned with the rest of our paper and stating a slightly stronger version of the result than is presented in their Corollary 10.4: specifically, restricting only to bounded random variables rather than to Bernoulli random variables. Lattimore and Szepesvári explain why this stronger version of their Corollary 10.4 also holds in their Notes 10.1 and 10.2 on page 140, leveraging a result due to Garivier and Cappé [2011]. Observe that this lemma (counter-intuitively) applies $d(x, y)$ to non-Bernoulli distributions.

Lemma 1 (Chernoff bound following Corollary 10.4 and Note 10.1 of Lattimore and Szepesvári [2020])**.** *Let $X_1, \ldots X_n$ be a sequence of independent and identically distributed random variables for which $X_t \in [0, 1]$. Let μ denote each variable's expected value, and let $\hat{\mu} = \frac{1}{n}\sum_{t=1}^{n} X_t$ be the sample mean. For $v \geq 0$, define $U'(v) = \max\{y \in [0, 1] : d(\hat{\mu}, y) \leq v\}$ and $L'(v) = \min\{y \in [0, 1] : d(\hat{\mu}, y) \leq v\}$. Then $P(\mu \geq U'(v)) \leq \exp(-nv)$ and $P(\mu \leq L'(v)) \leq \exp(-nv)$.*

Lemma 2. *If Good-Grade holds, then with probability at least $1 - \delta$, $s \geq PG$.*

Proof. Each s_i is a independent and identically distributed random variable on the interval $[0, 1]$; s is the mean of the distribution from which each s_i was sampled, and $\hat{s}$ is our sample estimate of s. So, from Lemma 1,

$$P(s \leq L'(v)) \leq \exp(-nv).$$

The event $s > L'(v)$ is complementary to the event $s \leq L'(v)$ so

$$P(s > L'(v)) \geq 1 - \exp(-nv).$$

Substituting in $v = \ln(1/(\delta))/n$ and simplifying,

$$P(s > L'(\ln(1/\delta)/n)) \geq 1 - \delta. \tag{A.1}$$

Appealing to Equation (4.2), we can rewrite $L'(\ln(1/\delta)/n) = L(\hat{s}, n, \delta)$. Then, Inequality (A.1) tells us that with probability no less than $1 - \delta$,

$$s > L(\hat{s}, n, \delta). \tag{A.2}$$

If *Good-Grade* holds,

$$L(\hat{s}, n, \delta) \geq PG. \tag{A.3}$$

Intersecting Inequalities (A.2) and (A.3) we obtain that with probability at least $1 - \delta$, $s > PG$. Thus, with probability at least $1 - \delta$, $s \geq PG$. □

Lemma 3. *If Bad-Grade holds, then with probability at least $1 - \delta/2$, $s < PG$.*

Proof. From Lemma 1,

$$P(s \geq U'(v)) \leq \exp(-nv).$$

The event $s < U'(v)$ is complementary to the event $s \geq U'(v)$ so

$$P(s < U'(v)) \geq 1 - \exp(-nv).$$

Substituting in $v = \ln(1/(\delta/2))/n$ and simplifying,

$$P(s < U'(\ln(1/(\delta/2))/n)) \geq 1 - \delta/2. \tag{A.4}$$

Appealing to Equation (4.1), we can rewrite $U'(\ln(1/(\delta/2))/n) = U(\hat{s}, n, \delta/2)$. Then, Inequality (A.4) tells us that with probability no less than $1 - \delta/2$,

$$s < U(\hat{s}, n, \delta/2). \tag{A.5}$$

If *Bad-Grade* holds,

$$U(\hat{s}, n, \delta/2) < PG \tag{A.6}$$

Intersecting Inequalities (A.5) and (A.6) we obtain that with probability at least $1 - \delta/2$, $s < PG$. ☐

Lemma 4. *If Good-Rid holds, then with probability at least* $1 - \delta$*,* $r \leq RID$*.*

Proof. Each measurement of whether or not the agent gave a ridiculous answer, denoted r_i, is an identically distributed Bernoulli random variable (each is either 1 or 0 depending on whether or not the agent received score zero for its answer to the sampled question); r is the mean of the Bernoulli distribution from which each r_i was sampled (the probability according to which the agent gives a ridiculous answer), and $\hat{r}$ is our sample estimate of r. We can therefore make an argument just as in Lemma 3, concluding analogously to Inequality (A.5) that with probability no less than $1 - \delta$,

$$r < U(\hat{r}, n, \delta). \tag{A.7}$$

If *Good-Rid* holds,

$$U(\hat{r}, n, \delta) \leq RID. \tag{A.8}$$

Intersecting Inequalities (A.7) and (A.8) we obtain that with probability at least $1 - \delta$, $r < RID$. Thus, with probability at least $1 - \delta$, $r \leq RID$. ☐

Lemma 5. *If Bad-Rid holds, then with probability at least* $1 - \delta/2$*,* $r > RID$*.*

Proof. Likewise, we can make an argument just as in Lemma 2, concluding analogously to Inequality (A.2) that with probability no less than $1 - \delta/2$,

$$r > L(\hat{r}, n, \delta/2). \tag{A.9}$$

If *Bad-Rid* holds,

$$L(\hat{r}, n, \delta/2) > RID. \tag{A.10}$$

Intersecting Inequalities (A.9) and (A.10) we obtain that with probability at least $1 - \delta/2$, $r > RID$. ☐

We are now ready to prove Theorem 1.

Theorem 1. *If Procedure 4.1 determines that an agent does or does not understand a given scope, this conclusion is correct with probability at least* $1 - \delta$*.*

Proof. To show that an agent understands a given scope, we must demonstrate that both $s \geq PG$ and $r \leq RID$. By Lemma 2, if *Good-Grade* holds, $s < PG$ with probability at most δ. Likewise, by Lemma 4, if *Good-Rid* holds then $r > RID$ with probability at most δ. We call the event where one of these conditions holds but the desired inequality on the corresponding true distribution does not hold a failure. The event where *both* failures occur is the intersection of the events in which each failure occurs, and so its probability is no greater than the maximum of the two failure probabilities. Each of these probabilities is bounded by δ, so we can conclude that, if both *Good-Grade* and *Good-Rid* both hold, ($s \geq PG$ and $r \leq RID$) will also hold with probability no less than $1 - \delta$.

To show that the agent does *not* understand a given scope, it is sufficient to demonstrate either that $s < PG$ or that $r > RID$. Our procedure draws this conclusion if at least one of *Bad-Grade* and *Bad-Rid* holds. By Lemma 3, if *Bad-Grade* holds, then with probability at least $1 - \delta/2$, $s < PG$. By Lemma 5, if *Bad-Rid* holds, then with probability at least $1 - \delta/2$, $r > RID$. By the union bound, the probability that at least one of these two conditions is satisfied while the corresponding condition on the underlying true distribution does not (a failure) is no more than the sum of the probabilities that each condition fails individually.[8] Thus, if Procedure 4.1 determines that an agent does not understand a given scope, the condition ($s < PG$ or $r > RID$) holds with probability at least $1 - (\delta/2 + \delta/2) = 1 - \delta$. ☐

[8]Thus, it was not necessary to make $\delta_s = \delta_r = \delta/2$; we could have picked any positive values for δ_s and δ_r that sum to δ or less.

Theorem 2. *If Procedure 5.1 determines that an agent does or does not understand a given scope, this conclusion is correct with probability at least* $1 - \delta$*.*

Proof. When we sample from Q, we obtain a sample from each Q_{α_i} with probability p_{α_i} and from $\bar{Q}$ with probability $\bar{p} = 1 - \sum_i p_{\alpha_i}$. First, let us consider average score. Denote the true score on $\bar{Q}$ as $\bar{s}$. Because the true score s is the expectation over the scores of all questions in Q and expectation is linear, s is a convex combination of each s_{α_i} and of $\bar{s}$,

$$s = \left(\sum_{\alpha_i \in \mathcal{A}} p_{\alpha_i} \cdot s_{\alpha_i} \right) + \bar{p} \cdot \bar{s}. \tag{A.11}$$

Consider each set of questions in turn. For questions from each Q_{α_i}, our beliefs about the average score do not depend on our samples at all: we know that the agent's true average score is s_{α_i}.

The samples $\bar{s}_1, \ldots, \bar{s}_{\bar{n}}$ we have taken from $\bar{Q}$ are independent random variables on $[0, 1]$, so the argument in the proof of Lemma 2 applies directly, substituting $n = \bar{n}$, $\hat{s} = \hat{\bar{s}}$ and $s = \bar{s}$. Thus, with probability at least $1 - \delta$,

$$\bar{s} > L(\hat{\bar{s}}, \bar{n}, \delta). \tag{A.12}$$

Intersecting Inequalities (A.11) and (A.12), with probability at least $1 - \delta$,

$$s > \left(\sum_{\alpha_i \in \mathcal{A}} p_{\alpha_i} \cdot s_{\alpha_i} \right) + \bar{p} \cdot L(\hat{\bar{s}}, \bar{n}, \delta) \tag{A.13}$$

If *Good-Grade-Exp* holds,

$$\left(\sum_{\alpha_i \in \mathcal{A}} p_{\alpha_i} \cdot s_{\alpha_i} \right) + \bar{p} \cdot L(\hat{\bar{s}}, n, \delta) \geq PG \tag{A.14}$$

Intersecting Inequalities (A.13) and (A.14), if *Good-Grade-Exp* holds then with probability at least $1 - \delta$, $s > PG$.

We can make an analogous argument about *Bad-Grade-Exp*. If *Bad-Grade-Exp* holds,

$$\left(\sum_{\alpha_i \in \mathcal{A}} p_{\alpha_i} \cdot s_{\alpha_i} \right) + \bar{p} \cdot U(\hat{\bar{s}}, \bar{n}, \delta/2) < PG \tag{A.15}$$

As above, we can make the same argument as in the proof of Lemma 3 but with $n = \bar{n}$, $\hat{s} = \hat{\bar{s}}$, and $s = \bar{s}$, obtaining $s < U(\hat{s}, n, \delta/2)$. Intersecting with Inequality (A.15), we obtain that with probability at least $1 - \delta/2$, $s < PG$.

Now let us consider the probability that the agent will return a ridiculous answer. Denote the true probability that the agent will return a ridiculous answer on $\bar{Q}$ as $\bar{r}$. By the Law of Total Probability, r is a convex combination of each r_{α_i} and of $\bar{r}$,

$$r = \left(\sum_{\alpha_i \in \mathcal{A}} p_{\alpha_i} \cdot r_{\alpha_i} \right) + \bar{p} \cdot \bar{r}. \tag{A.16}$$

As before, we can repeat the argument in the proof of Lemma 4, substituting $n = \bar{n}$, $\hat{r} = \hat{\bar{r}}$, and $r = \bar{r}$, obtaining $\bar{r} < U(\hat{\bar{r}}, \bar{n}, \delta)$. Intersecting with Inequality (A.16), with probability at least $1 - \delta/2$,

$$r \leq \left(\sum_{\alpha_i \in \mathcal{A}} p_{\alpha_i} \cdot r_{\alpha_i} \right) + \bar{p} \cdot U(\hat{\bar{r}}, \bar{n}, \delta/2). \tag{A.17}$$

If *Good-Rid-Exp* holds,

$$\left(\sum_{\alpha_i \in \mathcal{A}} p_{\alpha_i} \cdot r_{\alpha_i} \right) + \bar{p} \cdot U(\hat{\bar{r}}, \bar{n}, \delta/2) \leq RID. \tag{A.18}$$

Intersecting Inequalities (A.17) and (A.18), if *Good-Rid-Exp* holds then with probability at least $1 - \delta$, $r < RID$.

n	$0.9 - \sqrt{\frac{-\ln(0.05)}{n}}$	$0.9 + \sqrt{\frac{-\ln(0.025)}{n}}$	$0.5 - \sqrt{\frac{-\ln(0.05)}{n}}$	$0.5 + \sqrt{\frac{-\ln(0.025)}{n}}$
10	0.352667	1.507361	-0.047333	1.107361
100	0.726918	1.092065	0.326918	0.692065
1,000	0.845267	0.960736	0.445267	0.560736
10,000	0.882692	0.919206	0.482692	0.519206
100,000	0.894527	0.906074	0.494527	0.506074
1,000,000	0.898269	0.901921	0.498269	0.501921
10,000,000	0.899453	0.900607	0.499453	0.500607

Table 3: Example values for bounding s in Procedure 4.1, $\delta = 0.05$, leveraging the Hoeffding bound instead of the Chernoff bound. Compare to corresponding values in Table 1. As in that table, the two pairs of columns represent the confidence intervals around observed empirical averages of $\hat{s} = 0.9$ and $\hat{s} = 0.5$ respectively. Observe that the width of the confidence interval derived from the Hoeffding bound is the same in both cases. When values are greater than 1 or less than 0, the Hoeffding bound is uninformative.

Finally, we can make an analogous argument about *Bad-Rid-Exp*. If *Bad-Rid-Exp* holds,

$$\left(\sum_{\alpha_i \in \mathcal{A}} p_{\alpha_i} \cdot s_{\alpha_i}\right) + \bar{p} \cdot L(\hat{\bar{r}}, \bar{n}, \delta) > RID \tag{A.19}$$

As above, we can make the same argument as in the proof of Lemma 5, substituting $n = \bar{n}$, $\hat{r} = \hat{\bar{r}}$, and $r = \bar{r}$, obtaining $\bar{r} > L(\hat{\bar{r}}, \bar{n}, \delta/2)$. Intersecting with Inequality (A.19), we obtain that with probability at least $1 - \delta/2$, $r > RID$.

We have now proven results exactly analogous to those of Lemmas 2, 3, 4, and 5 for *Good-Grade-Exp*, *Bad-Grade-Exp*, *Good-Rid-Exp*, and *Bad-Grade-Exp*, the four conditions used in Procedure 4.1. We can thus follow the same argument as in the proof of Theorem 1 (substituting *Good-Grade-Exp* for *Good-Grade*, etc., and appealing to these new results rather than each of the corresponding lemmas just mentioned) to show that if Procedure 5.1 determines that an agent does or does not understand a given scope, this conclusion is correct with probability at least $1 - \delta$. $\square$

B On Chernoff vs Hoeffding Bounds

In our testing procedures, we leveraged upper- and lower-bounding functions U and L derived from the Chernoff bound, which lack closed-form expressions and hence require numerical approximation. However, we could alternatively have derived our four tests of average grade and ridiculousness via the Hoeffding bound, which does have a closed-form expression. Specifically, each s_i is an identically distributed random variable restricted to the interval $[0, 1]$ and each r_i is a Bernoulli random variable. We could thus have used Hoeffding's inequality to bound the gap between empirical and true distributions in both cases. Consider the simplest setting of Procedure 4.1. Here we could have concluded e.g., that

$$\hat{s} - \sqrt{\frac{-\ln(\delta_{\mathrm{s}})}{n}} \leq s$$

with probability at least $1 - \delta_{\mathrm{s}}$ and similarly that

$$\hat{r} + \sqrt{\frac{-\ln(\delta_{\mathrm{r}}/2)}{n}} \geq r$$

with probability at least $1 - \delta_{\mathrm{r}}/2$.

However, tests based on these bounds are somewhat worse for tests of average grade and lamentably difficult to satisfy for tests of ridiculousness. See Tables 3 and 4, which are analogues of Tables 1 and 2 using the Hoeffding bound. Consider a setting in which the agent has given no ridiculous answers. How many samples would we need to establish confidence that the true probability of giving a ridiculous answer has fallen below 0.00052? We can infer the answer from Table 4: over

n	$0 + \sqrt{\frac{-\ln(0.05)}{n}}$	$3/n + \sqrt{\frac{-\ln(0.05)}{n}}$	$0.01 - \sqrt{\frac{-\ln(0.025)}{n}}$
10	0.547333	0.847333	-0.597361
100	0.173082	0.203082	-0.182065
1,000	0.054733	0.057733	-0.050736
10,000	0.017308	0.017608	-0.009206
100,000	0.005473	0.005503	0.003926
1,000,000	0.001731	0.001734	0.008079
10,000,000	0.000547	0.000548	0.009393

Table 4: Example values for bounding r in Procedure 4.1 with $\delta = 0.05$, leveraging the Hoeffding bound instead of the Chernoff bound. Compare to corresponding values in Table 2. When values are greater than 1 or less than 0, the Hoeffding bound is uninformative.

ten million samples. In contrast, our test derived from the Chernoff bound required fewer than 10,000 samples.

Why do the two inequalities produce such different bounds, particularly when testing ridiculousness? The issue is that RID is very close to zero and, in the event that our test passes, $\hat{r}$ will be too. This is a problem for the Hoeffding bound when applied to Bernoulli random variables, which turns out to be extremely loose when the true probability is close to 0 or 1. Indeed, the Hoeffding bound always scales with the square root of the number of samples. The Chernoff bound does the same when the true probability is 0.5, but scales almost linearly with the number of samples when the true probability is close to 0 or 1.

C Wrong Answers and Bogus Explanations: An Example

Below is a session with ChatGPT on November 19th, 2023. ChatGPT provides wrong answers, and bogus explanations of the answers. (This is not a critique specifically of ChatGPT, which is used simply as an example.)

```
You
Consider events A, B, C. B happened during A. C started after A did, and
ended before B did.  Which started first - B or C?

ChatGPT
Let's break down the information:
• Event A started.
• Event B happened during A.
• Event C started after A did.
• Event C ended before B did.
Given this information, we can conclude that:
1.  A started first.
2.  C started after A.
3.  C ended before B.
Therefore, in the sequence of events, C started after A, and since C ended
before B, it means that C started after B as well.  The chronological order
would be B, C, and then A.

You
Are you sure?  Remember that events have duration.

ChatGPT
Apologies for any confusion in my previous responses.  Let's consider the
duration of events:
• Event A started.
• Event B happened during A.
```

• Event C started after A did and ended before B did.
Given the duration of events, we can refine the chronological order:
1. A started.
2. C started after A and ended before B.
So, considering the duration, C started after A and ended before B.
The correct chronological order is A, C, and then B. I appreciate your patience, and thank you for pointing out the importance of event duration. If you have any more questions or need further clarification, feel free to ask.
ChatGPT can make mistakes. Consider checking important information.