Logical Consistency Between Disagreeing Experts And Its Role In AI Safety

Andrés Corrada-Emmanuel
Data Engines
andres.corrada@dataengines.com

***Abstract*—If two experts disagree on a test, we may conclude both cannot be 100% correct. But if they completely agree, no possible evaluation can be excluded. This asymmetry in the utility of agreements versus disagreements is explored here by formalizing a logic of unsupervised evaluation for classifiers. Its core problem is computing the set of group evaluations that are logically consistent with how we observe them agreeing and disagreeing in their decisions. Statistical summaries of their aligned decisions are inputs into a Linear Programming problem in the integer space of possible correct or incorrect responses given true labels. Obvious logical constraints, such as, "The number of correct responses cannot exceed the number of observed responses," are inequalities. But in addition, there are "axioms"—universally applicable linear equalities that apply to all finite tests. The practical and immediate utility of this approach to unsupervised evaluation using only logical consistency is demonstrated by building no-knowledge alarms that can detect when one or more LLMs-as-Judges are violating a minimum grading threshold specified by the user.**

***Index Terms*—unsupervised evaluation, logic, formal verification**

I. Introduction

AI systems did not create the principal/agent problem [1] but they are exacerbating it. We have an economic and epistemological problem when we delegate to agents tasks that we do not want to do, or cannot do because of our ignorance. Thus, we would like to minimize our supervision while still remaining safe when we delegate our decisions. Here we will consider how logical consistency between possibly disagreeing agents can be used to ameliorate this ancient problem (e.g. Plato's Allegory of the Ship of Fools in *The Republic*) that AI systems cannot escape.

LLMs-as-Judges [2] has emerged as one obvious way to handle the problem of monitoring assistant LLMs when we do not have access to a ground truth for evaluating their decisions. This, however, does not resolve the recursive nature of the problem of evaluating experts without answer keys to the tests we give them. If we build a monitoring system, what or who monitors that? This is the problem of *infinite monitoring chains*—who judges the judges?

No intelligence, however advanced, can escape the laws of logic or physics [3]. We can use logical consistency when evaluating experts without a ground truth for their decisions (unsupervised evaluation). Tools built using logical consistency alone will be, by construction, universally applicable to any domain.

Using consistency alone, we can construct a logic that excludes group evaluations that contradict observed statistical summaries of expert responses. The one-bit summary of two experts in the abstract is a simple example of how the logic works. It is also an example of how the exclusion of evaluations can serve as the basis for no-knowledge alarms for classifiers.

The one-bit summary of the responses by the two experts carries no information about the domain of the test. Nor do we have any text or other semantic information to guide our evaluation. Depending on its value, we can exclude a joint evaluation for them (both cannot be 100% correct). If our safety standard had been that both classifiers had to be perfect in their classifications, observing disagreement would have triggered an alarm.

The logic will treat the classifiers as black-boxes and only use summaries of how they decided or responded when asked to classify items into a finite set of R labels to be denoted by $\mathcal{L}$. Such a logic cannot be used in evaluating experts that go beyond answering factual questions with a finite number of answer choices [4]. But it can be used to evaluate the judges of those experts. When we ask LLMs to act as judges of the complex outputs of other LLMs it often happens that we are asking them to be classifiers. Graders are often classifiers.

Consider the MT-Bench [2] dataset created to establish benchmarks for evaluating LLMs-as-Judges when they carry out pair comparisons between the complex outputs of LLMs. Each LLM-as-a-Judge is presented with the outputs of two models that can be generically denoted *model_a* and *model_b*. The judge must decide which output they prefer or if both are equally good. In effect, the LLMs-as-Judges are acting as classifiers by outputting one of three possible grading labels—a, b, or tie.

This concrete example of the appearance of the classification task when we consider the ability of LLM-as-Judges should make clear that any logic of unsupervised evaluation for classifiers could have wide applicability. Not as a way to evaluate agents carrying out the primary task of the system, but as a way to evaluate agents that are grading those primary ones.

Logical consistency is the only thing that can be established in unsupervised settings where no ground truth is available. There are no truth tables to allow us to formally prove that the decisions of experts are correct. But we can prove or disprove the logical consistency of various assertions we may want to

make about tests, their answer keys, and how well experts did answering them.

A. Suspecting the answer key and all test takers

Holistic evaluations of LLMs seek to measure their performance across different dimensions [5]. Aligning these systems to human preferences requires that we define a possible ground truth for what that preference is. But human experts are widely observed to never completely agree on that truth. This occurs in the MT-Bench benchmark where 5K+ human judgments show an agreement rate of 80%. This may be thought to occur because the human experts are judging on properties like "roleplaying" ability. But agreement rates meaningfully below 100% can also be observed across doctors in different countries reviewing medical cases [6].

If we want to design AI systems that are human-centered, what does that mean when human experts cannot agree on what the answer key for a test is? This problem will be considered in section IV where the logic will be used to establish an upper bound on the minimum accuracy of human experts given that we believe there is a correct answer key to the test they took.

II. The basic question in logics of unsupervised evaluation

We are discussing a logic of unsupervised evaluation for classifiers. But it is instructive to consider what properties we would want from anything that claimed to be a logic of unsupervised evaluation. Classification is just one of the many tasks ML systems carry out. Accordingly, we would expect there to be a logic of unsupervised evaluation for regressors, etc.

In each case, a basic question for such a logic must be—what are the group evaluations logically consistent with how we observe test takers agreeing/disagreeing in their decisions? For each case, we can build statistical summaries of the events of the disagreement. For one dimensional regressors we would have the histogram of their pairwise differences, for example. And for classifiers we can summarize the question-aligned responses as the integer counts of the R^N ways that N classifiers can agree/disagree when doing R-label classification.

We can formalize this further for classifiers by considering the trivial ensemble, $M = 1$. In this case, the basic question can be stated as,

> Given the summary of classifier i's test responses as label integer counts, $(R_{\ell_1}, R_{\ell_2}, \ldots, R_{\ell_R})_i$, what are the possible integer points,
>
> $$(Q_{\ell_1}, Q_{\ell_2}, \ldots, Q_{\ell_R}) \otimes \prod_{\ell_{\text{true}} \in \mathcal{L}} (R_{\ell_1;\ell_{\text{true}}}, R_{\ell_2;\ell_{\text{true}}}, \ldots, R_{\ell_R;\ell_{\text{true}}}), \quad (1)$$
>
> logically consistent with those observations?

This integer space for enumerating the possible evaluations of a single classifier is composed of two types of subspaces. First we have the space for unknown statistics of the answer key for the test. For R-label classification, that space is of dimension R. But note that not all points in this space are consistent with observations in a test of size Q. In particular, any properly constructed answer key projects to non-negative integers that satisfy the equality,

$$\sum_{\ell_r \in \mathcal{L}} Q_{\ell_r} = Q \quad (2)$$

If there is an answer key to the test, its label counts project to a point, $(Q_{\ell_1}, Q_{\ell_2}, \ldots, Q_{\ell_R})$, in what we will call the Q-complex (QC). And once we are at that point in the QC, all the possible number of responses , $R_{\ell_r;\ell_{\text{true}}}$, of ℓ_r given true label, ℓ_{true}, must be on the plane defined by,

$$\sum_{\ell_r \in \ell_{\text{true}}} R_{\ell_r,\ell_{\text{true}}} = Q_{\ell_{\text{true}}}. \quad (3)$$

All logical statements about evaluations can only occur at a single fixed point in the Q-complex. When we are using how test takers agree and disagree on a given test, there is no other possibility. This will be used when we discuss how to construct the alarm comparing classifier performance at the same assumed point in the QC.

Having picked a point in the QC, we can also lower the dimensionality for the subspaces associated with possible correct responses given true label. These are the spaces associated with tuples of the form,

$$\{(R_{\ell_1;\ell_{\text{true}}}, R_{\ell_2;\ell_{\text{true}}}, \ldots, R_{\ell_R;\ell_{\text{true}}})\}_{\ell_{\text{true}} \in \mathcal{L}}. \quad (4)$$

For these spaces we require,

$$\{\sum_{\ell_r \in \mathcal{L}} R_{\ell_r,\ell_{\text{true}}} == Q_{\ell_{\ell_{\text{true}}}}\}_{\ell_{\text{true}} \in \mathcal{L}} \quad (5)$$

As a consequence of these plane equalities, possible evaluations exist within an $(R-1)(R+1)$ dimensional space within the $R(R+1)$ space in (1). For binary classification ($R = 2$) this is a 3-dimensional space shown in Fig. 1. Note that the possible set does not fill the space and has non-trivial geometry.

We could do all of the logic in a space more familiar to the ML community. We could represent a binary classifiers evaluations as points in the space defined by,

$$(P_{a_i,a} = \frac{R_{a_i;a}}{Q_a}, P_{b_i,b} = \frac{R_{b_i;b}}{Q - Q_a}, P_a = \frac{Q_a}{Q}). \quad (6)$$

The possible set of evaluations in that space for a test of size $Q = 10$, the same as Figure 1, is shown in Figure 2. It is clear that what we will call the R-space representation is easier to visualize than the P-space typically used in ML metrics. The two representations have their own use, however. For example, completeness of the axioms for classifiers is easier to prove in P-space as discussed in the Appendix. The rest of the paper will continue in the R-space representation but equivalent constructions are possible in P-space.

The set of all possible evaluations before we observe a test taker i's responses can be reduced further once we observe

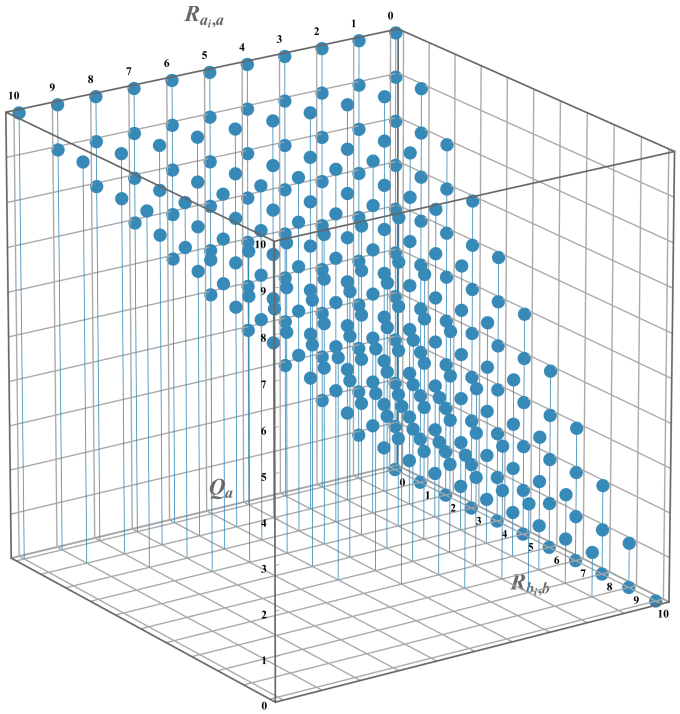

Fig. 1. All possible evaluations for a binary classifier labeling $Q = 10$ items. $R_{a_i,a}$ and $R_{b_i,b}$ are the number of correct responses by classifier i for labels 'a' and 'b' respectively.

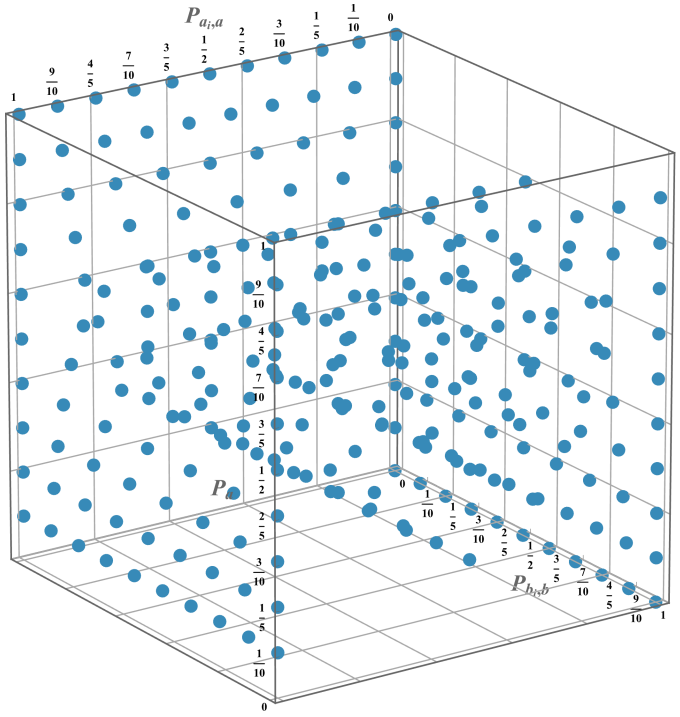

Fig. 2. All possible evaluations for a $Q = 10$ test but now in the space of prevalence,$P_a = Q_a/Q$, and label accuracies, $P_{a_i,a} = R_{a_i,a}/Q_a$ and $P_{b_i,b} = R_{b_i,b}/(Q - Q_a)$. This is visual proof that the geometry of possible evaluations is easier in the integer response space shown in Fig. 1.

a summary of their responses as $(R_{(\ell_1)_i}, R_{(\ell_2)_i}, \ldots, R_{(\ell_R)_i})$. Then, it must be true for any label, ℓ_{true}, that,

$$R_{(\ell_r)_i;\ell_{\text{true}}} \leq R_{(\ell_r)_i}. \tag{7}$$

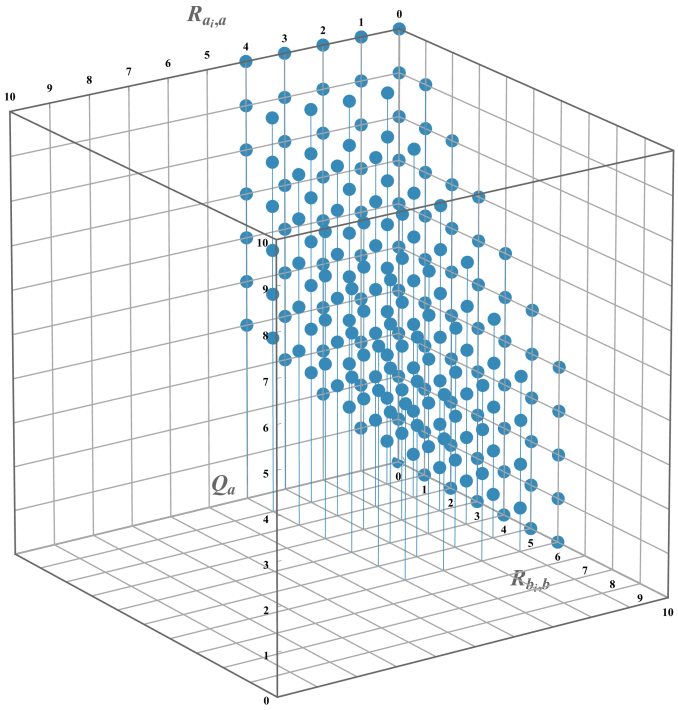

Fig. 3. All possible evaluations for a $Q = 10$ test after we observe the test summary $(R_{a_i} = 4, R_{b_i} = 6)$. The inequalities change the number of possible evaluations but not the dimension of their geometry.

In Fig. 3 we see the effect of these inequality constraints for the binary classification example where we have observed $(R_{a_i} = 4, R_{b_i} = 6)$ and can thus conclude that the only possible evaluations must obey,

$$R_{a_i;a} \leq 4, \; R_{b_i;b} \leq 6. \tag{8}$$

In this $Q = 10$ example, there are 285 possible evaluations and the inequalities after having summarized the test responses reduce this to 235 but do not change the dimension of the possible set given test results. To do that, one must discuss a set of axioms that relate label response counts across the true labels.

A. Axioms for the single classifier

We now discuss the set of equalities that must be obeyed by any classifier i taking a test of size Q and where we observe they respond with summary $(R_{(\ell_1)_i}, R_{(\ell_2)_i}, \ldots, R_{(\ell_R)_i})$. For each label in the set of R labels, called ℓ_{true}, and each classifier

i in an ensemble, the following linear relation is always equal to zero,

$$R_{(\ell_{\text{true}})_i;\ell_{\text{true}}} - Q_{\ell_{\text{true}}} + \sum_{(\ell_r \neq \ell_{\text{true}}) \in \mathcal{L}} R_{(\ell_r)_i} - \sum_{(\ell \neq \ell_{\text{true}}) \in \mathcal{L}} \sum_{(\ell_r \neq \ell_{\text{true}})} R_{(\ell_r)_i;\ell}. \quad (9)$$

There are R of these equations, one for each label, as we would expect from symmetry when no ground truth is available. It cannot be the case that just observing decisions can privilege any one label. Their simple proof is given in the Appendix. They reduce the size of the possible set and also the dimension of its geometry. Fig. 4 shows the effect of the axioms for the $Q = 10$ binary test example.

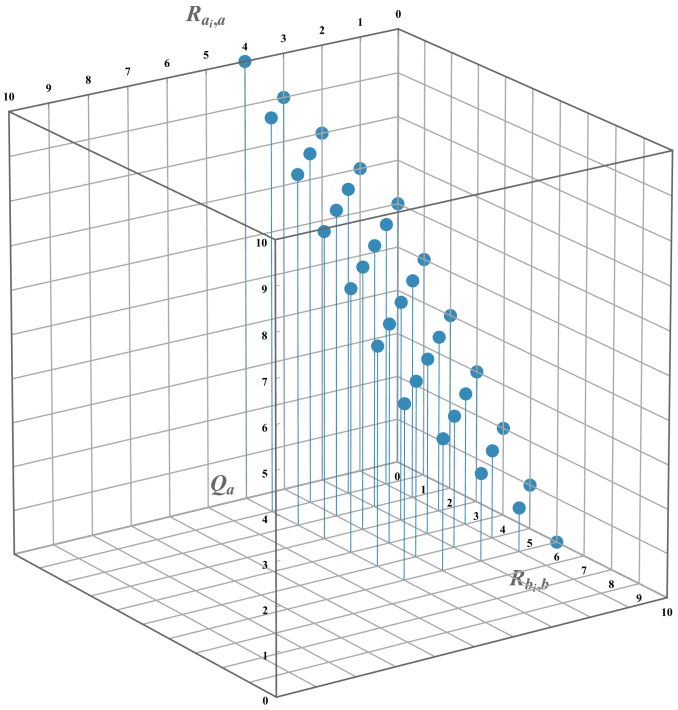

Fig. 4. All possible evaluations for a $Q = 10$ binary test after we observe the test summary $(R_{a_i} = 4, R_{b_i} = 6)$ and pick evaluations consistent with it as expressed by either of the binary axioms: $R_{a_i,a} - Q_a + (R_{b_i} = 6) - R_{b_i,b}$, or $R_{b_i,b} - Q_b + (R_{a_i} = 4) - R_{a_i,a}$.

It is instructive to work through an example of a possible evaluation after the inequality constraints (7) are imposed that is impossible given the observed counts. Consider the running $Q = 10$ example with observed responses $(R_{a_i} = 4, R_{b_i} = 6)$. A possible evaluation is $(R_{a_i,a} = 3, R_{b_i,b} = 1, Q_a = 7)$. It satisfies the inequality constraints (7), but cannot be correct given the observed test responses.

If they have only one label 'b' decision correct $(R_{b_i,b} = 1)$, that means 5 of their 'b' decisions were actually 'a's. But there are only 7 'a's in the answer key, so the classifier cannot be 3 times correct on 'a' decisions $(R_{a_i,a} = 3)$. It can only be $R_{a_i,a} = 2$ everything else being equal. The evaluation $(R_{a_i,a} = 2, R_{b_i,b} = 1, Q_a = 7)$ obeys the axioms (9).

The equations are complete (see Appendix) but not independent, they sum to identically zero. So in binary classification, $R = 2$, we can define the set of possible evaluations as lying on a plane, but in general they reduce the dimension by $R-1$.

The filtering effect of the axioms is stronger than that of the response inequalities since they define reduce the dimension of the possible set. In the running example of a single binary classifier labeling $Q = 10$ items, there are 286 possible evaluations. In general, there would be evaluations $1/6(Q+1)(Q+2)(Q+3)$ for a single binary classifiers (Appendix). The inequality constraints (7) reduce this to 210 and the axioms (9) to 35. Roughly, the single classifier choosing between R labels is going to have $Q^{R(R-1)}$ evaluations and this is reduced by $(R-1)$ dimensions by these axioms. Thus, the percentage reduction in possible evaluations goes as,

$$\frac{1}{Q^{(R-1)}}. \quad (10)$$

B. Group evaluations logically consistent with $M = 1$ summaries

If we have the question-by-question responses of each test taker, sets of the form $\{(q, R_{(\ell_r)_i})\}$, we could proceed to consider summaries of the test for any subset of size M for the N classifiers. So far, we have been considering the $M = 1$ summaries, $(R_{(\ell_1)_i}, R_{(\ell_2)_i}, \ldots, R_{(\ell_R)_i})$. And we will continue to do so since this allows us to avoid the problem of considering how classifiers are correlated in their decisions.

Note however, that the algebra of the logic allows us to conclude that, whatever the set is that fully takes into account $M = m > 1$ summaries, it must be contained within the product of the m $M = 1$ sets. It cannot be the case that evaluations inconsistent with the $M = 1$ summary of a classifier, become consistent given pair summaries with any other classifier. Whatever that pair summary is, it has to project to the single summary that defines the $M = 1$ possible set.

One way to view this is as a ladder of logical inference. We first have to assume a location point in the QC for the summary of the answer key. The previous section then discussed the evaluations possible given $M = 1$ summaries. To obtain the evaluations logically consistent with the $M = 2$ summaries, we start from the product space of the $M = 1$ summaries and then consider what points are excluded from that product space given the $M = 2$ axioms.

We can visualize this product space at each point in the QC for a pair of binary classifiers as two square spaces, one for each label as in, for example, Fig. 5. These are their possible evaluations at QC point $(Q_a = 6, Q_b = 4)$ consistent with $M = 1$ axioms given that we observed individual test responses $(R_{a_i,a} = 4, R_{b_i,b} = 6)$ and $(R_{a_j,a} = 7, R_{b_j,b} = 3)$. Beyond binary classification, this becomes harder or impossible to visualize so we will use *correctness squares or cubes*, where we plot the number of correct responses for each true label in separate figures, and each label figure has axes variables corresponding to the different classifiers. By construction, we are throwing away the count of error decisions. So for $R > 2$ classification, we can include the

multiplicity of evaluations for each possible correct point. The next section will discuss this as we proceed to explain the atomic logic operation for no-knowledge alarms of misaligned classifiers.

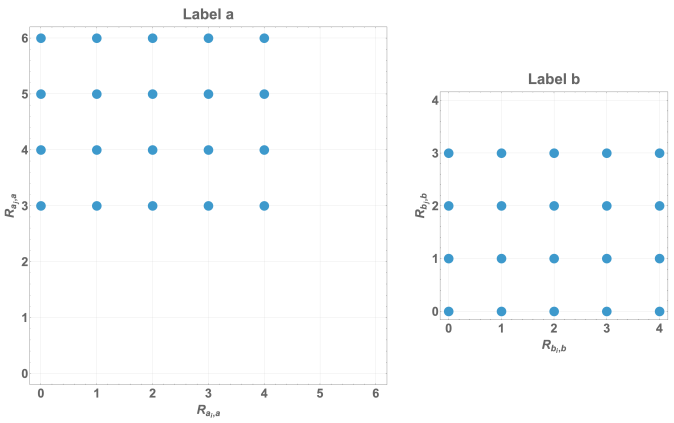

Fig. 5. All possible evaluations for a $Q = 10$ binary test assuming the answer key summary is $(Q_a = 6, Q_b = 4)$ and we have observed the test summaries $(R_{a_i} = 4, R_{b_i} = 6)$ and $(R_{a_j} = 7, R_{b_j} = 3)$ for classifiers i and j.

III. No-knowledge alarms for misaligned classifiers

The MT-Bench benchmark [2] was created to evaluate the performance of LLMs-as-Judges. The concept of using other agents to monitor the work of primary agents is well-known as a safety design pattern through-out history. But as work on the principal/agent problem in economics shows [7], using supervisors to manage the work of others, does not resolve all monitoring issues.

In particular, using AI agents to monitor others in unsupervised settings, as LLMs-as-Judges are intended to do, still leaves unresolved who monitors the judges. The logic being presented here based on consistency between agreement/disagreement counts of test-takers can alleviate this unresolvable problem. Absent ground truth for the decisions of experts and any other side information, using other experts to grade them leads to *infinite monitoring chains*.

But note that as the monitoring chain extends, the evaluation of the judges becomes easier and simpler. This is the case with the core evaluation task in the MT-Bench benchmark – pair comparisons of the complex output of two LLMs. The use of pair comparisons means that, in effect, the expert comparing their outputs is acting as a 3-label classifier. Presented with "model_a" and "model_b" outputs, a judge can pick either one (two labels) or the third label ("models tied"). So we can use the logic developed here to build alarms for such pair comparison agents.

The basic idea of the alarms is that the possible set of classifier evaluations at any point in the QC excludes some evaluations. This means we can build no-knowledge alarms for evaluation conditions that are label and classifier symmetric

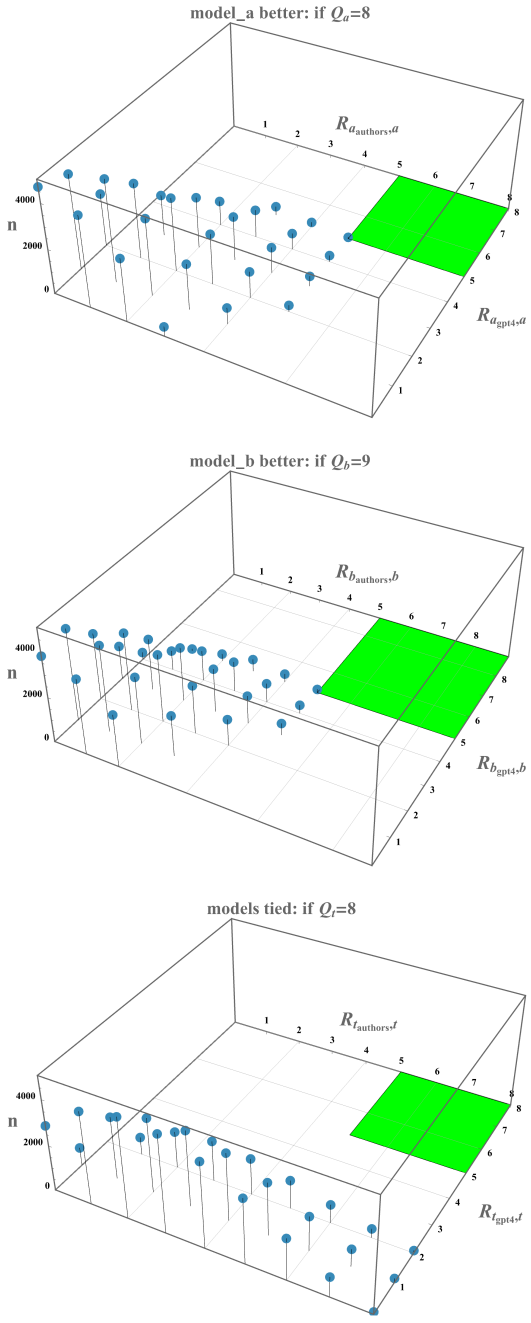

Fig. 6. The multiplicity count of possible number of correct responses by two experts doing pair comparisons of LLM outputs (model_a and model_b). They graded 25 pairs from the MT-Bench dataset. Their grade summaries are, $(R_{a_i} = 5, R_{b_i} = 10, R_{t_i} = 10)$ and $(R_{a_j} = 4, R_{b_j} = 18, R_{t_j} = 3)$ for `authors` and `gpt4`, respectively. There are no possible evaluations, given these summaries, and the assumption that the correct answer key for the test has summary $(Q_a = 8, Q_b = 9, Q_t = 8)$, that has both classifiers grading the tied pair comparisons better than 25%. The green squares represent points were both classifiers are more than 50% correct.

in wholly unsupervised settings. For example, in Fig. 6, the product space of $M = 1$ evaluations for two experts grading 25 pair-comparisons from the MT-Bench benchmark are shown. These are the possible number of correct label responses for two graders (the `authors` of the MT-Bench dataset and `gpt4`) that we have observed to respond with summaries $(R_{a_i} = 5, R_{b_i} = 10, R_{t_i} = 10)$ and $(R_{a_j} = 4, R_{b_j} = 18, R_{t_j} = 3)$ for `authors` and `gpt4`, respectively.

The `authors` decision is the weighted vote of the authors of the MT-Bench dataset [2]. `gpt4` is the LLM-as-Judge grader of the pair comparisons of two LLM outputs. Details about the construction of the voted decisions for the `authors` and the construction of a ground truth decision using human expert judgments are in Appendix B. We see in Fig. 6 that there is no evaluation for `gpt4` where it is doing better than about 25% correct on the pair comparisons that human experts deemed a tie.

If we had set the alarm to trigger on any label performance being less than, say, 50% correct, the observed test summaries for `authors` and `gpt4` would have triggered it for this particular $Q = 25$ test. We have verified that at this QC point, there is at least one classifier violating the threshold for one of the labels. But this observed failure occurred at a single point in the QC. For a test of size $Q = 25$ with $R = 3$ possible responses, there are 351 points of the form (Q_a, Q_b, Q_t) in the QC. It may be that at some other assumed summary of the answer key, both graders are better than 50% or some other value for both labels. It so happens that for this particular $Q = 25$ test a trigger set to any value greater 46% would be able to detect one of the classifiers is misaligned with human experts on one of the labels.

We can determine, universally, what minimum threshold value for label accuracy would be triggered by the observed single classifier summaries. This can be visualized as a plot of the minimum of the maximum label accuracies for both classifiers as shown in Fig. 7. The max of this min-max plot is at about 45%. Thus any setting greater than this could be used to trigger a monitoring alarm that would have gone off with these summaries. It is worthwhile noting that the worst performing accuracy is actually about 14% on the tied judgments of gpt4. Logical consistency can only upper bound the accuracy that would be triggered given the observed $M = 1$ summaries, in this case.

Note that this alarm is predicated on an accuracy threshold that is label and classifier independent. There is no privileged label or classifier when logical consistency is all we are using. If we plot the min-max by label – the mininmum value of the maximum possible accuracy for both classifiers – as shown in Fig. 8, we see that no threshold is possible by label. For any label, there will always be answer key summaries that make it impossible to exclude any possible evaluation from 0% to 100%. We do not discuss the setting of alarms that set different thresholds per label but require the conditions to hold simultaneously for all labels at any point in the QC.

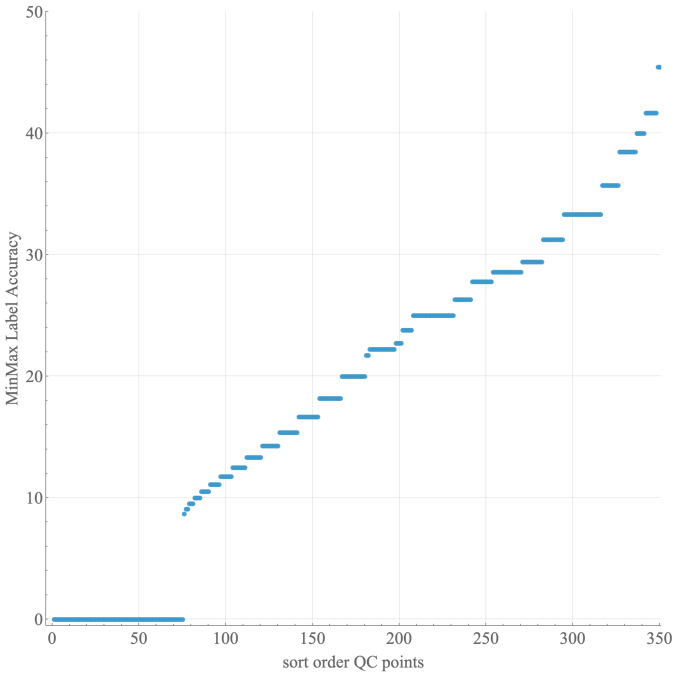

Fig. 7. Plot of the minimum value of the max label accuracy for both graders of the $Q = 25$ MT-Bench pair comparison test. For the observed grade summaries, $(R_{a_i} = 5, R_{b_i} = 10, R_{t_i} = 10)$ and $(R_{a_j} = 4, R_{b_j} = 18, R_{t_j} = 3)$, for `authors` and `gpt4` respectively, any alarm condition set at more than 46% would trigger. The 351 points in the QC of a 3-label $Q = 25$ test have been ordered by the min-max value.

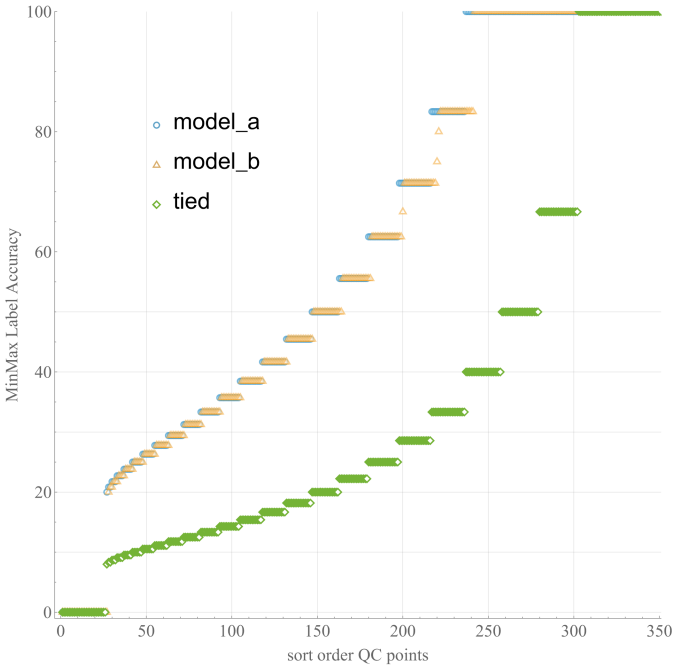

Fig. 8. Plot of the minimum value of the max label accuracy, by label, for both graders of the $Q = 25$ MT-Bench pair comparison test. We cannot set an alarm on an individual label since, for any label, there are answer summaries for any accuracy level.The 351 points in the QC of a 3-label $Q = 25$ test have been order by the min-max value, per label. So points at this plot cannot be compared across label for fixed sort order.

IV. Suspecting the ground truth

Logical consistency is not magical. It cannot sense the presence of unknown answer keys. Strictly speaking, the no-knowledge alarms are not establishing that classifiers are misaligned to an unknown ground truth. They are establishing that there is no answer key, given the count of test responses, that has all classifiers obeying the accuracy specification at the set trigger level.

It could be that there is no answer key for the test we have formulated. Logical consistency cannot answer those scientific questions. Those depend on knowledge of the domain where the logic is applied. The logic is universally applicable to any domain, but it cannot establish whether the tests we use are correct or not.

There is a practical way to use this inability of what assumption in our statements is the false one. Consider the widespread use of agreement rates between human experts to establish the answer key for evaluations. This is the case with the MT-Bench dataset we have been using. It contains about five thousand *human expert* pair comparison judgments across different dimensions of LLM performance.

The agreement rate between the MT-Bench human experts is about 80% [2]. This is about typical with other agreement rates, even in medical contexts where we may expect the truth to be easier to ascertain by experts. So how come we believe that we can align AI systems to human values when humans, even those considered experts in a domain, are observed to disagree at rates of 20% or higher [8]?

The approach presented here suggests that we should "invert" the process of discussing ground truths established by disagreeing experts. We can ask, what is the trigger threshold set by the observed disagreements on a supposedly existing ground truth? Whether human experts disagreeing 20% on a ground truth for evaluating LLM outputs is a cause for concern or not is not resolved by just being told that they agree 80%. That sounds good enough. Is it? This formalism turns the observed counts of disagreements into a statements of the form,

> The ground truth established by these N experts is logically consistent with them being x% or more correct on all the labels.

We argue, in fact, that this should become a minimum standard for scientific work that reports answer keys established by disagreeing experts. The statistics of how they disagree and what this implies for their minimum competency based on the belief of the existence of a correct, but unknown, answer key can be answered by this logic. And can serve as a surrogate for the "error" in the answer key constructed by ensembles of imperfect classifiers.

V. Conclusions

Disagreements between graders of 25 pair comparisons of LLM outputs from the MT-Bench benchmark have been shown to be sufficient to trigger alarms of the form,

$$\begin{aligned}((P_{a_i,a} > x\%) \wedge (P_{a_j,a} > x\%)) \wedge \\ ((P_{b_i,b} > x\%) \wedge (P_{b_j,b} > x\%)) \wedge \\ ((P_{t_i,t} > x\%) \wedge (P_{t_j,t} > x\%)),\end{aligned} \quad (11)$$

for any x greater than 46%. This was done by looking at the set of possible evaluations logically consistent with the $M = 1$ summaries, $(R_{(\ell_1)_i}, R_{(\ell_2)_i}, \ldots, R_{(\ell_R)_i})$, as established by the single axioms (9) as applied to three label classification.

The limitations of logical consistency alone should be clear. It can never detect that classifiers agreeing in their answers are incorrect. Most importantly, these logical considerations cannot establish the scientific validity or usefulness of using these alarms in any context. Those remain scientific and engineering questions that no logic could universally answer.

These no-knowledge alarms should be viewed as simple devices that can form part of a greater monitoring system. Like smoke alarms, they cannot determine what is causing the problem or how to fix it. But knowing that something is definitely wrong, with logical certainty, can be used to increase the safety of actions taken using the decisions of disagreeing experts, whether human or robotic.

VI. Related Work

Unsupervised evaluation has been treated in the ML literature. Most work has focused on making point estimates of possible correctness of experts, not its logic and what consistency implies for the possible set of evaluations given test response summaries. Dawid and Skeene [9] first treated the problem of grading medical doctors without the ground truth of their case decisions. Bayesian approaches were started by Raykar [10] and further developed by Xang, Xi, Zhung, and Jordan [11]. Parisi and subsequently Nadler, have developed a spectral approach [12]–[14].

As noted, all these approach are probabilistic and did not discuss the concept of possible evaluations being restricted by universally applicable axioms as is being claimed here. This paper, however, is not the first to discuss logical aspects of unsupervised evaluation. The first proof of unsupervised axioms was done by Platanios and his *agreement equations* [15], [16]. It should be clear from this paper that agreements are not sufficient nor very useful as ways to keep track of classifier performance. The agreement equations are part of the polynomials we can write for all the possible ways N classifiers can agree/disagree. There are R^N of these polynomials, only R of them are about the agreement events for all of them.

To fully understand the axioms for any $M = m$ summary, one must go to the P-space we studiously avoided in the paper. In that space it becomes possible to quickly write polynomials of the unknown evaluation values (prevalences, label accuracies, and their error correlations). The tools of algebraic geometry can then be used to establish factorizations of those polynomials. The use of algebraic geometry in statistics was pioneered by Pistone [17].

References

[1] Wikipedia contributors, "Principal–agent problem — Wikipedia, the free encyclopedia," 2025, [Online; accessed 3-September-2025]. [Online]. Available: https://en.wikipedia.org

[2] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica, "Judging llm-as-a-judge with mt-bench and chatbot arena," 2023.

[3] M. Tegmark and S. Omohundro, "Provably safe systems: the only path to controllable agi," 2023. [Online]. Available: https://arxiv.org/abs/2309.01933

[4] A. T. Kalai, O. Nachum, S. S. Vempala, and E. Zhang, "Why Language Models Hallucinate," Sep. 2025, arXiv:2509.04664 [cs]. [Online]. Available: http://arxiv.org/abs/2509.04664

[5] P. Liang, R. Bommasani, T. Lee, D. Tsipras, D. Soylu, M. Yasunaga, Y. Zhang, D. Narayanan, Y. Wu, A. Kumar, B. Newman, B. Yuan, B. Yan, C. Zhang, C. Cosgrove, C. D. Manning, C. Ré, D. Acosta-Navas, D. A. Hudson, E. Zelikman, E. Durmus, F. Ladhak, F. Rong, H. Ren, H. Yao, J. Wang, K. Santhanam, L. Orr, L. Zheng, M. Yuksekgonul, M. Suzgun, N. Kim, N. Guha, N. Chatterji, O. Khattab, P. Henderson, Q. Huang, R. Chi, S. M. Xie, S. Santurkar, S. Ganguli, T. Hashimoto, T. Icard, T. Zhang, V. Chaudhary, W. Wang, X. Li, Y. Mai, Y. Zhang, and Y. Koreeda, "Holistic evaluation of language models," 2023. [Online]. Available: https://arxiv.org/abs/2211.09110

[6] I. Kohane. (2025) Doctor group consenses by country. [Online]. Available: https://bit.ly/4gPGSRt

[7] J.-J. Laffont and D. Martimort, *The Theory of Incentives: The Principal-Agent Model*. Princeton, NJ Oxford: Princeton University Press, Jan. 2002.

[8] O. Salaudeen, A. Reuel, A. Ahmed, S. Bedi, Z. Robertson, S. Sundar, B. Domingue, A. Wang, and S. Koyejo, "Measurement to meaning: A validity-centered framework for ai evaluation," 2025. [Online]. Available: https://arxiv.org/abs/2505.10573

[9] P. Dawid and A. M. Skene, "Maximum likelihood estimation of observer error-rates using the em algorithm," *Applied Statistics*, pp. 20–28, 1979.

[10] V. C. Raykar, S. Yu, L. H. Zhao, G. H. Valadez, C. Florin, L. Bogoni, and L. Moy, "Learning from crowds," *Journal of Machine Learning Research*, vol. 11, no. 43, pp. 1297–1322, 2010.

[11] Y. Zhang, X. Chen, D. Zhou, and M. I. Jordan, "Spectral methods meet em: A provably optimal algorithm for crowdsourcing," in *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2014, pp. 1260–1268.

[12] F. Parisi, F. Strino, B. Nadler, and Y. Kluger, "Ranking and combining multiple predictors without labeled data," *Proceedings of the National Academy of Sciences*, vol. 111, no. 4, pp. 1253–1258, 2014.

[13] A. Jaffe, B. Nadler, and Y. Kluger, "Estimating the accuracies of multiple classifiers without labeled data," in *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, G. Lebanon and S. V. N. Vishwanathan, Eds. San Diego, California, USA: PMLR, 2015, pp. 407–415.

[14] A. Jaffe, E. Fetaya, B. Nadler, T. Jiang, and Y. Kluger, "Unsupervised ensemble learning with dependent classifiers," in *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, A. Gretton and C. C. Robert, Eds. Cadiz, Spain: PMLR, 2016, pp. 351–360.

[15] E. A. Platanios, E. A. Blum, and T. Mitchell, "Estimating accuracy from unlabeled data: A bayesian approach," in *Proceedings of The 33rd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 48, New York, New York, USA, 2014, pp. 1416–1425.

[16] E. A. Platanios, A. Dubey, and T. Mitchell, "Estimating accuracy from unlabeled data: A bayesian approach," in *Proceedings of The 33rd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. F. Balcan and K. Q. Weinberger, Eds., vol. 48, New York, New York, USA, 2016, pp. 1416–1425.

[17] G. Pistone, E. Riccomagno, and H. P. Wynn, *Algebraic Statistics: Computational Commutative Algebra in Statistics*. Chapman and Hall/CRC, Dec. 2000.

Appendix A
Proof of the $M = 1$ Axioms

The proof the $M = 1$ axioms for any number of labels is straightforward and proceeds by expressing all variables in the axiom in terms of responses by true label. For example, for any observed count of ℓ_r by classifier i, $R_{(\ell_r)_i}$, it must be true that,

$$R_{(\ell_r)_i} = \sum_{\ell_{\text{true}} \in \mathcal{L}} R_{(\ell_r)_i, \ell_{\text{true}}}. \tag{12}$$

Likewise, for any $Q_{\ell_{\text{true}}}$, we can express it in terms of the classification correct and incorrect decisions as,

$$Q_{\ell_{\text{true}}} = \sum_{\ell_r \in \mathcal{L}} R_{(\ell_r)_i, \ell_{\text{true}}}. \tag{13}$$

The axioms then follow trivially by linear cancellation of all terms.

Appendix B
Constructing the ground truth and expert decisions

The majority vote of human experts in the MT-Bench pair comparisons was done by weighted voting. An expert vote was split into two $1/2$ votes in the case of ties in a pair comparison, but kept as 1 if model a or model b was picked. Thus, for example, if two experts voted (model a, tie) this would be consider a preference for model a. Since we are using the tie grading label, this procedure is unambiguous in assigning one of the three available grades. The same procedure was used to establish the grades by the panel of `authors`. In that case, we allowed pair comparisons that had just grades by one author. The turn 1 pair comparisons used are show in Table B.

As discussed in the paper, logical consistency cannot establish that this construction of *a ground truth* is scientifically valid or useful. This MT-Bench example is meant to illustrate the algebraic operations of the logic and that differences can trigger no-knowledge alarms set to practical thresholds of safety.

Appendix C
The $M = 2$ axioms and correlated experts

Once one has determined the point in the QC and its corresponding product space of $M = 1$-consistent set of evaluations, we can consider the subset of that space that is consistent with the $M = 2$ set of axioms for the $\binom{N}{2}$ pairs in an ensemble of N classifiers. This subset in single response space is a projection of the logically consistent space that now must include variables of the form,

$$R_{(\ell_r)_i, (\ell_s)_j}. \tag{14}$$

These are the statistical summaries of how many times we observed the two classifiers responding $((\ell_r)_i, (\ell_s)_j)$ on the same question.

The unknown statistics of their aligned responses given true label form R copies of R^2 dimensional spaces. These are the counts of how often the pair agreed and disagreed given true label. By construction, they also lie on simplexes of one lower

TABLE I

HUMAN EXPERT WEIGHTED VOTE GROUND TRUTH, EXPERTS,FOR THE 25 LLM PAIR COMPARISONS FROM THE MT-BENCH DATASET. THERE ARE 4 MODEL_A GRADES, 14 MODEL_B, AND 7 TIES.

Question id	Model A	Model B	experts	authors	gpt4
82	gpt-4	claude-v1	b	a	a
82	llama-13b	gpt-4	b	b	b
85	vicuna-13b-v1.2	gpt-3.5-turbo	tie	tie	b
91	gpt-3.5-turbo	gpt-4	tie	tie	a
91	llama-13b	gpt-3.5-turbo	b	b	a
91	vicuna-13b-v1.2	gpt-3.5-turbo	a	tie	b
98	vicuna-13b-v1.2	gpt-3.5-turbo	tie	a	a
99	vicuna-13b-v1.2	gpt-3.5-turbo	b	b	b
103	gpt-3.5-turbo	gpt-4	b	b	b
104	alpaca-13b	gpt-3.5-turbo	tie	tie	b
105	gpt-3.5-turbo	gpt-4	b	tie	b
111	llama-13b	gpt-4	b	b	b
112	gpt-3.5-turbo	claude-v1	a	tie	tie
115	gpt-3.5-turbo	claude-v1	a	a	a
121	vicuna-13b-v1.2	gpt-4	b	tie	b
124	gpt-3.5-turbo	gpt-4	b	tie	tie
134	gpt-3.5-turbo	gpt-4	b	b	b
140	alpaca-13b	gpt-3.5-turbo	tie	b	b
141	gpt-3.5-turbo	gpt-4	tie	tie	b
142	gpt-3.5-turbo	gpt-4	b	b	b
143	alpaca-13b	vicuna-13b-v1.2	tie	a	b
148	gpt-3.5-turbo	gpt-4	b	a	b
148	llama-13b	claude-v1	b	b	b
151	llama-13b	gpt-3.5-turbo	b	b	b
152	alpaca-13b	vicuna-13b-v1.2	a	tie	b

dimension for each label. And, similar to the $M = 1$ case, we need to impose the inequalities,

$$R_{(\ell_r)_i,(\ell_s)_j,\ell_{\text{true}}} \leq R_{(\ell_r)_i,(\ell_s)_j}. \tag{15}$$

The count of any pair label event given true label cannot be higher than the observed count of that event in their question aligned responses.

The $M = 2$ axioms are also a set of R linear equations but now in the expanded space that includes the QC point, the $M = 1$ product space point for the pair, and the new pair response spaces given true label. We state the axiom outright. For every ℓ_{true}in the R labels $(\ell_1, \ell_1, \ldots, \ell_R)$ and every pair of classifiers i and j the following linear relation is identically zero.

$$\begin{aligned} &R_{(\ell_{\text{true}})_i,(\ell_{\text{true}})_j;\ell_{\text{true}}} - Q_{\ell_{\text{true}}} \\ &\quad + \sum_{c\in\{i,j\}} \sum_{\ell_r\neq\ell_{\text{true}}} R_{(\ell_r)_c} - \sum_{c\in\{i,j\}} \sum_{\ell_r\neq\ell_{\text{true}}} \sum_{\ell_s\neq\ell_{\text{true}}} R_{(\ell_s)_c;\ell_r} \\ &\quad - \sum_{\ell_r\neq\ell_{\text{true}}} R_{(\ell_r)_i,(\ell_r)_j} + \sum_{\ell_r\neq\ell_{\text{true}}} \sum_{\ell_s\neq\ell_{\text{true}}} R_{(\ell_s)_i,(\ell_s)_j;\ell_r} \\ &\quad - \sum_{\ell_r\neq\ell_s\neq\ell_{\text{true}}} R_{(\ell_r)_i,(\ell_s)_j;\ell_{\text{true}}} \end{aligned} \tag{16}$$

Its proof is exactly as the one for $M = 1$ but now we expand all response variables to pair responses given true label,

$$R_{(\ell_r)_i,(\ell_s)_j,\ell_{\text{true}}}. \tag{17}$$

Statistical measures of the correlation between pairs of classifiers can then be defined as,

$$\Gamma_{(\ell_r)_i,(\ell_s)_j;\ell_{\text{true}}} = \frac{R_{(\ell_r)_i,(\ell_s)_j;\ell_{\text{true}}}}{Q_{\ell_{\text{true}}}} - \frac{R_{(\ell_r)_i;\ell_{\text{true}}}}{Q_{\ell_{\text{true}}}} \frac{R_{(\ell_s)_j;\ell_{\text{true}}}}{Q_{\ell_{\text{true}}}} \tag{18}$$

APPENDIX D
CLASSIFIERS AND MULTIPLE-CHOICE TEST TAKERS HAVE THE SAME UNSUPERVISED EVALUATION LOGIC

The lack of semantics in the logic explained here means that its algebra applies equally well to classifiers or any test taker responding to a Multiple Choice (MC) exam. The only difference is how we interpret the response counts.

When we are doing classification, response 'a' in one item/question is pointing to the same thing as response 'a' in another. In classification, an evaluation of the classifiers is also telling us a statistic about the items the classifiers labeled – their prevalence.

In an MC exam, response letters need not have any semantic equality between questions. Saying 'b' in one question may mean something completely different from answering 'b' in another. The letters are a convenience that allows, for example, students to fill out bubble-answer sheets. In the MC exam, the ground truth for the QC point has no meaning outside the test. It is an artifact of how the fixed, finite number of responses were encoded to letter responses.

The algebraic purpose of the label prevalences in an MC exam lies in that they allow you to compute the percentage of test questions correct, g_i, as,

$$g_i = \sum_{\ell_r\in\mathcal{L}} \frac{Q_{\ell_r}}{Q} \frac{R_{(\ell_r)_i,\ell_r}}{Q_{\ell_r}}, \tag{19}$$

for each classifier i given QC point $(Q_{\ell_1}, \ldots, Q_{\ell_R})$. This defines a set of logically consistent grades given observed test responses on an MC exam.

APPENDIX E
IF YOU SOLVE THE LOGIC OF SINGLE QUESTION SUMMARIES, YOU SOLVE THE LOGIC FOR QUESTION SEQUENCES

Yet another advantage of logical consistency of counts of agreements/disagreements being semantic free is that if you have axioms for the question-aligned summaries, you immediately can write down the algebra for any sequence-aligned summaries you may care to do.

Take the case of wanting to know statistics of correctness/errors for consecutive pairs in a test. For a test of size Q there are $Q-1$ such pairs and we can map the R^2 tuples of the form $((\ell_r)_i, (\ell_s)_j)$ to R^2 labels. There is always an isomorphic mapping of whatever sequence of questions we define for the test and R^S labels where S is the size of the sequences.