Reviewer #2 (Remarks to the Author):

The authors clearly put in a lot of work to address Reviewer 3's concerns. My judgment of the manuscript remains similar to what it was during the last submission, though I do have some comments relating to the new content/analyses and how it is described and interpreted.

We thank the reviewer for their positive assessment of our work, and for taking the time to revisit our manuscript to provide further feedback.

1. The "predicting success on held-out questions using estimated knowledge" section is the one that went through the largest overhaul. I still find the within-lecture and across-lecture analyses to be important, though it is unfortunate that with the new methods the results are harder to interpret. The authors do a good job of trying to explain the pattern of results (sometimes knowledge estimates do predict quiz question accuracy, and other times they do not), though it is not a strong set of results with clear empirical insights. However, since in my view the value of this work is in its methodological rather than empirical contribution, I do not see this as a reason to not approve the manuscript. That being said, I do have a (a) point of confusion, and (b) question/suggestion regarding the analysis.

In general, we view the main contribution of our work similarly to the reviewer (i.e., as primarily methodological, rather than primarily empirical). That said, we appreciate and agree that in the previous version of our manuscript, the results of these analyses were somewhat challenging to interpret. This was due to the effects of two separate issues with our analyses. The first was a statistical bias we noted in our previous response letter (and describe in greater detail in our response to comment 1a) that led to an inverse relationship under our null hypothesis between estimated knowledge for a quiz question and the probability of answering it correctly. The second was a conceptual issue (described in our response to comment 1b) that we identified in the course of the current revision with how we had constructed the "null" models used in our significance tests. As detailed in our response to comment 1b, we have now addressed both of these issues and report an updated set of results in our revised manuscript (pp. 14–19). We believe that our updated results are substantially easier to interpret and provide much clearer empirical insights into whether, when, and to what extent our knowledge estimates predict participants' success on individual quiz questions.

a. In the rebuttal document, the authors explain an issue regarding the models incorporating the percent-correct measure, in which using percent-correct alone results in a negative relationship between estimated knowledge and probability of answering a question correctly (rebuttal starting at page #3). Given this strange relationship, the authors say that they have decided not to include these results in their paper. However, in the paper, the authors do include analyses that relate to this negative null result (starting manuscript page #16). I assume that the authors are referring to slightly different analyses in their manuscript vs. rebuttal, but if they decided to include analyses with this negative relationship in the manuscript, why did they leave out other ones? The fact that I am not clear why certain analyses were included or excluded makes me feel like I am missing an important aspect of these analyses, but perhaps this is a fault of my own.

We apologize for the confusion. To summarize, Reviewer 3 had proposed two separate changes to our manuscript, one of which we adopted, and one we did not. Both changes involved using generalized

linear mixed models (GLMMs) to assess the predictive power of our knowledge estimates for held-out questions, and both were affected by an issue involving a "negative relationship" in the data to which these models were fit. Specifically, holding out individual questions from a participant's quiz responses introduces an inverse (or "negative") relationship between their success on a given held-out question and their proportion-correct score on the remaining ("held-in") questions. The "issue" with this relationship is not that it is negative per se, but that (as we explain below) it is fundamentally not a relationship that can be estimated numerically, and therefore incorporating it into a model of participants' success on held-out questions (i.e., by including their remaining-proportion-correct scores as a predictor) leads to intractable issues with the model's estimation.

The change we declined to adopt was the addition of a new analysis that entailed fitting GLMMs with fixed effects for participants' remaining-proportion-correct scores in addition to their estimated knowledge. As described below, we identified a number of issues with this approach, one of which is that these models cannot be accurately estimated. The other change Reviewer 3 proposed (and which we *did* adopt) was replacing the Mann-Whitney *U*-tests previously reported in Figure 6 with analogous tests based on GLMMs. While the models we fit for these tests didn't directly consider participants' remaining-proportion-correct scores (only their estimated knowledge), under our null hypothesis, our knowledge estimates reduce to remaining-proportion-correct scores and exhibit the same problematic negative relationship with success on held-out questions. (This is the "negative null" we described in our manuscript when introducing this set of results.) In this case, our inability to accurately model this "negative null" relationship led to a conservative bias in the significance tests presented in our previous submission, which we have now corrected for in our revised manuscript.

We describe this "negative relationship" issue and its effects on both proposed analyses in greater detail below. We describe our approach to correcting for it (in the analyses shown in Figure 6) in our response to comment 1b.

First, suppose a participant correctly answers n out of q questions on a given quiz. If we hold out a correctly answered question from their quiz responses, their proportion-correct score on the remaining questions will be $(n - 1)/(q - 1)$, whereas holding out an incorrectly answered question would yield a remaining-proportion-correct score of $n/(q - 1)$. In this way, any given participant's remaining-proportion-correct score will always be *lower* for held-out questions they answered correctly than for those they answered incorrectly—in other words, within participants, remaining-proportion-correct is inversely (or "negatively") related to success on held-out questions.

In both of the GLMM-based analyses Reviewer 3 proposed, they requested we fit models with random effects for participant and question identities to account for potential groupings by those factors in our data. Fitting these models therefore entails (in part) estimating any fixed effects they include based on each individual participant's (and question's) observations, separately. For the analysis we declined to adopt, which involved fitting models with fixed effects for remaining-proportion-correct, this would include estimating the "negative" within-participant relationship described above. However, this isn't a statistical relationship we can estimate through regression, but rather it is simply a mathematical byproduct: each participant's remaining-proportion-correct score can take on only two possible values

($n/(q-1)$ or $(n-1)/(q-1)$, for their particular n), and which of those values it takes on for a given observation is fully determined by the value of the response variable we're trying to predict (i.e., their success on the held-out question). In other words, this relationship is perfectly deterministic and lacks any residual variance that would allow for statistical estimation. Attempting to fit a logistic model to this relationship results in "complete separation" in the response variable, as the combination of participants' identities and their remaining-proportion-correct scores perfectly distinguishes between correctly and incorrectly answered held-out questions. If both of these terms were modeled as fixed-effect predictors (e.g., as in a standard logistic regression), this would lead to infinite parameter estimates, making the likelihood function non-identifiable and causing the model to fail to converge altogether. Modeling participants as random effects masks this problem to a degree, as partial pooling across random-effect levels shrinks these estimates towards zero such that it is *possible* for the model to converge (hence our ability to report *a* result from this analysis for illustrative purposes in our previous response letter). However, this "shrinkage" does not "fix" the underlying circularity in the model's formulation, nor the complete separation it creates. Instead, the model the reviewer had proposed would produce unstable and misleading results, as it assigns extreme values to these otherwise-infinite parameters (and artificially shrinks or inflates other parameter estimates to compensate) in an attempt to accommodate a "relationship" that is really a tautology in disguise.

These issues also affected the GLMM-based analyses we adopted for Figure 6, but in a slightly less direct way. For each of these analyses (i.e., each panel of Fig. 6), we performed a likelihood-ratio test comparing two different GLMMs: (1) a "full" model we fit to explain participants' success on held-out questions given their estimated knowledge at those questions' embedding coordinates, with random effects for participants and questions, and (2) a "null" model we fit to explain the same data given only the full model's random effects. Formally, these tests assess whether the likelihood of our observed data is significantly greater under the full model than under the null model—in other words, whether a model that considers participants' estimated knowledge explains their success on held-out questions significantly better than one that does not (or equivalently, whether knowledge estimates provide significant explanatory power). However, because the models we're comparing differ by only a single fixed effect, these tests are also (asymptotically) equivalent to testing whether the relationship between estimated knowledge and (the log-odds of) success on held-out questions is significantly different from *zero*—which does not necessarily reflect the expected relationship under our null hypothesis. This is because our estimates of participants' knowledge for held-out questions are simply weighted versions of their remaining-proportion-correct scores, where the weight we assign to each held-in question reflects its embedding-space distance from the held-out question (with smaller distances corresponding to larger weights).

Our core assumption in constructing these estimates—which we are testing in these analyses—is that a participant's ability answer a particular question explains more about their ability to answer other questions at nearby embedding coordinates than about their ability to answer questions at far-away coordinates (i.e., that knowledge of the concepts tested by these questions is "smooth" with respect to distance in embedding space). If this is *not* the case, as we assume under our null hypothesis (i.e., if the between-question distances we leverage as weights are randomly distributed with respect to participants' correct vs. incorrect responses), then on average, our estimates of participants' knowledge will approach their *unweighted* remaining-proportion-correct scores and exhibit the same inverse relationship with their success on held-out questions. Note that this applies specifically to the "All questions" and "Within-lecture" versions of these analyses (top row and middle two rows of Fig. 6), wherein we estimate knowledge for each held-out question using the remaining questions from the *same* pool as the one we held out (i.e., all remaining questions from the same quiz, or all remaining questions about the same lecture, from the same quiz). By contrast, in the "Across-lecture" analyses (Fig. 6, bottom two rows), we estimate knowledge for each question about one lecture using only questions about the *other* lecture. In this case, "holding out" a particular question does *not* alter a participant's proportion-correct score on the questions we use to estimate their knowledge.

The above means that our significance tests for the "All questions" and "Within-lecture" analyses were overly conservative, because they assumed a "more positive" null relationship (i.e., zero) than we would actually expect given truly uninformative knowledge estimates. Any predictive power our knowledge estimates *do* afford would therefore have to surpass a relatively high threshold to overcome this "negative baseline" and yield an apparent positive relationship. Although assessing our results against the true "negative null" (as the reviewer suggests in comment 1b) *would* in theory have been a less conservative test, it was unfortunately not clear to us how such a comparison could be operationalized since this null relationship (between unweighted remaining-proportion-correct scores and success on held-out questions) is not a "real" statistical relationship we can represent numerically in order to test hypotheses against. At the same time, we felt that Reviewer 3's proposed GLMM-based approach *was* worth adopting for these analyses, as we agreed with their assessment that it accounted for important effects in our data that our prior *U*-tests had not (e.g., baseline differences in participants' performance and questions' difficulties). In weighing these factors, we ultimately concluded that a highly conservative test would be preferable to a potentially anti-conservative one, and opted to include these analyses in our previous submission along with a brief explanation of why an apparent negative relationship actually reflects a null result. However, we acknowledge this was not an ideal solution. As we describe in response to comment 1b, in our newly revised manuscript, we have now devised a correction for this conservative bias that does not require explicitly modeling this negative null relationship and instead addresses the underlying problem that led to it by ensuring that participants' remaining-proportion-correct scores do *not* vary with their success on held-out questions (as was already the case in the "Across-lecture" analyses). We have also added text to our revised *Methods* section describing this "negative relationship" issue and motivating our approach to correcting for it (pp. 36–38).

Finally, we note that despite having now corrected for this issue, we have still elected not to adopt the additional GLMM-based analysis Reviewer 3 had proposed, as we feel there are remaining problems with its formulation. For this analysis, Reviewer 3 had requested we use a likelihood-ratio test to compare two models (both fit with the appropriate random effects): one with fixed effects for both remaining-proportion-correct *and* estimated knowledge, and one with a fixed effect for remaining-proportion-correct but *not* for estimated knowledge. The goal of this proposed analysis was to "*provide a direct evaluation of the predictivity of 'estimated knowledge' against a baseline*"—i.e., to

isolate the predictive information uniquely contributed by our knowledge estimates from that already afforded by traditional proportion-correct scores. We agree that this is an appropriate baseline against which to evaluate our knowledge estimates, but do not believe that including these scores as a separate predictor is the right way to accomplish this. Rather, a likelihood-ratio test comparing these two models would assess whether considering participants' estimated knowledge *in addition to* their proportion-correct scores better explains their success on held-out questions than considering their proportion-correct scores alone. In other words, this analysis treats estimated knowledge as a measure intended to *supplement* traditional proportion-correct scores, whereas we propose estimating a learner's knowledge as an *alternative* to computing their simple proportion-correct score. Since knowledge estimates are simply weighted proportion-correct scores, these two measures contain partially redundant "information" (and in fact are perfectly collinear under our null hypothesis). Instead, the aspect of our knowledge estimates that *can* be thought of as "supplementing" traditional proportion-correct scores is the particular *weights* we use in constructing them. Thus an analysis that achieves the intended goal of this proposed analysis would be one that isolates the predictive information uniquely contributed by these weights from that of the unweighted proportion-correct scores we apply them to. Since our knowledge estimates reduce to unweighted proportion-correct scores under our null hypothesis, and the bias correction we now employ ensures these scores do not vary with a participant's success on held-out questions, this is exactly what is now accomplished by the likelihood-ratio tests we report in Figure 6. In fact, if we were to apply the same bias correction to this proposed analyses, the estimated effect of remaining-proportion-correct would be exactly *zero*, since there would be no within-participant variation for it to explain, and its between-participant variation would be the same information already captured by the model's per-participant random intercepts—making this analysis mathematically equivalent to the analyses shown in Figure 6. In our view, this reflects what information an individual's proportion-correct score fundamentally *can* provide about their ability to answer a particular quiz question (that is, none beyond their baseline probability of correctly answering any question), and why the question-specific knowledge estimates we compute therefore yield "higher-resolution" insights.

b. In any event, the situation in which the null result is a negative relationship causes confusions and complications (regarding the results displayed in Fig. 6). If the null result is negative (and not zero), then shouldn't the significance of the results be assessed relative to that negative value (rather than zero)? For example, might there be some observed relationship that may not itself be significantly positive, but still different from the negative null? If so, and if I understand the author's analysis correctly, then might they not be using a significance test that is actually too conservative? This might be a situation in which generating a null distribution through permutation testing might be the best solution; then the significance of the observed relationship can be compared to the null distribution to determine significance.

In brief, the reviewer is largely correct in their assessment. This "negative null" phenomenon did indeed render our significance tests highly conservative (specifically for the "All questions" and "Within-lecture" analyses) since these tests effectively assume a null relationship of zero, rather than the "negative" one we would expect. Unfortunately, since this negative null actually reflects a mathematical dependency rather than a "real" statistical relationship with a finite magnitude we can

quantify numerically, any typical approaches we might take to constructing a null hypothesis around it would either fail or not be statistically valid. For example, permutation testing would not work here because this “negative null” arises from the direct correspondence between a specific participant’s success on a specific held-out question and their remaining-proportion-correct score excluding that question. Any permutation of the data that breaks this correspondence would therefore no longer exhibit the negative relationship it was intended to capture. In other words, our observations are inherently not exchangeable under our null hypothesis because this negative relationship is not one that could have occurred by chance (and thus could be approximated through random permutation), but rather one that exists due to the circular dependency *within* each observation.

We also considered a number of other methods of accounting for this negative null relationship in fitting these models (e.g., adjusting for it with a global offset term, using GLMMs that incorporate Bayesian priors, residualizing estimated knowledge against remaining-proportion-correct, permuting the questions’ embedding coordinates rather than correct/incorrect labels, etc.) and ultimately came to a similar conclusion as above in each case. We therefore devised a correction procedure that instead works by *eliminating* this dependency from our data prior to fitting the models. Specifically, this correction ensures that the knowledge estimates we compute for questions a given participant answered correctly and incorrectly are based on the *same* underlying proportion of correctly answered questions, and therefore have the same expected value under our null hypothesis. In our revised manuscript, we refer to this as a “rebalancing procedure” to mirror the term used to describe analogous approaches to correcting for an analogous problem in leave-one-out cross-validation (e.g., Austin et al., 2024: https://arxiv.org/abs/2406.01652).

To summarize, suppose again that a participant correctly answers n out of q questions on a given quiz: their remaining-proportion-correct score will be $(n - 1)/(q - 1)$ when we hold out a correctly answered question and $n/(q - 1)$ when we hold out an incorrectly answered question. To correct for this difference, when we hold out a given question, we identify from the set of $q - 1$ remaining questions all those with the opposite “correctness” label of the held-out question (i.e., when holding out a correctly answered question, we identify all remaining incorrectly answered questions, and vice versa). We then additionally exclude each of these opposite-label questions, in turn, and estimate the participant’s knowledge for the held-out question using the remaining $q - 2$ questions. This yields a set of knowledge estimates for the held-out question that are each derived from an underlying proportion-correct score of $(n - 1)/(q - 2)$, regardless of whether the held-out question was answered correctly or incorrectly. We then average over these estimates to obtain a “rebalanced” estimate of the participant’s knowledge for the held-out question. This rebalanced estimate also has an expected value of $(n - 1)/(q - 2)$ under our null hypothesis (since it’s an average over estimates with that expected null value), but effectively “spreads” the shift in its underlying remaining-proportion-correct score equally across all opposite-label questions’ contributions.

In our revised manuscript, we use these rebalanced knowledge estimates to fit GLMMs for the “All questions” and “Within-lecture” analyses shown in Figure 6. Importantly, this means that the null-hypothesized relationship between estimated knowledge and the log-odds of success on a

held-out question is now zero (as was already the case in the "Across-lecture" analyses), and is therefore accurately reflected by the null models used in our likelihood-ratio tests. This also means that any non-zero relationship we observe in these analyses now reflects predictive information specifically afforded by the embedding-space distances we use to weight individual questions in constructing our knowledge estimates.

Finally, in developing this rebalancing procedure, we identified a separate issue with how we had constructed the "null" models used in these tests. Previously, in instances where the data supported estimating group-specific deviations in the effect of estimated knowledge (i.e., fitting a "full" model with "random slopes" for participants and/or questions), we had excluded these random slope terms from our null models in addition to the fixed effect. This is common practice when performing likelihood-ratio tests for the purpose of variable selection, where the goal is typically to determine whether considering a variable "in any capacity" improves the model's fit to the data. However, since we are using likelihood-ratio tests for the purpose of hypothesis testing, this effectively conflated tests for two different null hypotheses: that there is no relationship between estimated knowledge and the log-odds of success on held-out questions, and that this relationship does not vary among participants and/or questions. In other words, removing these random slopes from our null models actually made our prior likelihood-ratio tests somewhat *anti*-conservative, since an observed significant result could potentially reflect explanatory power contributed by multiple different terms. This was far overshadowed by the more substantial conservative bias in the "All questions" and "Within-lecture" analyses, but led to false-positive results in the "Across-lecture" analyses for Quiz 1. In our revised manuscript, we have now remedied this issue by retaining these random slope terms in our null models when they are also present in the corresponding full models.

2. The authors discuss their LDA embedding space in relation to more complex (BERT) and simpler (word-matching) models. I have (a) a comment on manuscript organization, and (b) a question about what this means in practice.

a. It seems strange to take up so much space with the comparisons between LDA and BERT while the BERT results are relegated to the supplementary materials. As a style suggestion, it would seem cleaner/more efficient to either move some of these discussions to the supplementary material alongside those results, or to condense the current material and keep in the discussion.

We appreciate this suggestion. We agree it feels odd to devote a large portion of our *Discussion* section to these comparisons, since (while interesting and useful) they are somewhat tangential to our main results and overall narrative. At the same time, given that both of these comparisons (between LDA and BERT, and LDA and simple word-matching) were motivated by similar questions from multiple reviewers, we suspect that some discussion of their implications will also be of interest to readers more broadly. We have therefore elected to move the bulk of text concerning these comparisons to our *Supplementary Information* file while retaining a higher-level overview of what we see as their most relevant implications in our *Discussion*. In our supplement, we have added and expanded on the text related to lower-level comparisons between specific elements of these figures in a new section titled *Supplementary results* (beginning on p. 22 of our *Supplementary Information* file).

b. The authors claim that their LDA embedding space is a "sweet spot" in between LLMs like BERT and simpler word-matching models. How specific is this to the particular lectures and quizzes used in the current study? Might e.g. LLMs work better than LDA for other researchers wanting to use these methods? If this is the case, then how should other researchers using their model determine which kind of embedding space to use? What kind of test can be run to determine the correct level of granularity in the embedding space? The proposed model will not be that useful of a tool if it takes thorough testing to determine the appropriate kind of embedding space before any actual knowledge analysis is done.

The reviewer poses a number of interesting questions here. We will attempt to address each in turn.

> *The authors claim that their LDA embedding space is a "sweet spot" in between LLMs like BERT and simpler word-matching models. How specific is this to the particular lectures and quizzes used in the current study?*

To clarify, it is not our intention to argue that LDA is universally "better" or "worse" than BERT or other large language models, nor do we intend our manuscript to make any claims or comparisons regarding the general usefulness of different model algorithms or architectures. Rather, our point is that if and when our goal is to obtain text embeddings that distinguish between highly similar content within a relatively narrow conceptual domain (e.g., nearby moments in a brief physics lecture, as in our current paper), we suggest that it is beneficial to use a text embedding model trained specifically on the content we want to be able to "explain," rather than one that has been trained to explain a far broader and more diverse range of content (like BERT and other LLMs).

This reflects a fundamental tradeoff in constructing text embedding models (and language models more broadly), between "generality" and "specificity." One key factor that determines where a particular model falls along the spectrum of this tradeoff is the breadth of conceptual content included in its training corpus. Essentially, the enormous and highly diverse text corpora used to train modern LLMs (in the case of BERT, roughly 3.3 billion words across thousands of books and all of English Wikipedia) enables them to represent content from a vast array of conceptual domains in a way that models trained on smaller, domain-specific corpora (like our LDA model) fundamentally cannot. However, since these models' embedding spaces must be able to represent any content within the enormous scope of their training materials, this necessarily limits the level of specificity (or "resolution") with which they can characterize subtle differences between content within a more narrow scope. In other words, while a model trained exclusively on content from a given domain can dedicate its entire feature space to capturing subtle similarities and distinctions that are meaningful within that domain (at the cost of generalizing poorly beyond that domain), a general-purpose model must apportion its representational capacity across a much broader conceptual landscape.

One way to think about this is to substitute the semantic features we want to capture using text embeddings for physical features we might instead want to capture through photographs. Suppose that rather than relating and distinguishing lectures and quiz questions based on their conceptual (semantic) features, we wanted to (either qualitatively or algorithmically) relate and distinguish different individuals based on their facial features. If we used photographs containing only those

individuals' faces, we might be able to do this fairly well. But if we were to "zoom out" too far when taking these pictures, to the point where the vast majority of their "space" was devoted to representing content other than the subject's face (e.g., the surrounding environment), and the subject's face was represented using only a small number of pixels, then the particular features we care about capturing would be blurred and lost. We might still be able to detect that each image contains *a* face, but we would fail to capture the more subtle and specific features that identify and distinguish *different* faces.

This is what we see as the key take-away from the comparisons between our LDA model, which we trained exclusively on the lectures used in our study, and BERT, whose embedding space must "span" its far larger and more diverse training corpus (Supp. Fig. 10). The embeddings afforded by our LDA model capture the lecture and question content at a "resolution" that enables us to distinguish between questions about broadly similar but subtly (and meaningfully) different content areas, and to "match" individual questions to specific sections of lecture content. By contrast, BERT assigns both lectures and all quiz questions to the same tiny region of its embedding space, indicating that they are all broadly "about" the same domain of physics concepts, but blurring the more subtle distinctions between the different concepts they cover within that domain.

To extend this facial-feature analogy, we could also imagine taking pictures that are "zoomed *in*" too far, to the point where each image captures only a small patch of skin or hair. In this case, we would similarly fail to capture the features relevant to our goal, as we would be considering only fragments of them—for example, a given patch of skin might be part of a nose, ear, chin, or something else, but we lack the surrounding context needed to determine this. This is what we see as the key take-away from our comparisons to simple word-matching "models" (Supp. Fig. 11). If a given question contains the word "force," that question could be asking about concepts related to the gravitational force, electromagnetic force, strong or weak nuclear forces, or something else. But the distinction between these possibilities is blurred and lost because the abstract concepts discussed in the lectures and referenced by the quiz questions are not defined or distinguishable at the level of individual words.

This is the "sweet spot" notion we describe in our manuscript, between simpler and more complex methods we could use to capture and relate the lectures' and questions' contents. While larger, more complex LLMs "take too *broad* a view" in characterizing their semantic contents, simpler word-matching approaches "take too *narrow* a view," and as a result, both alternatives fail to capture conceptual similarities and differences at a level of granularity that would allow us to meaningfully estimate knowledge for the different concepts covered within the space of a single course lecture. But by training a model specifically on the course materials for which we want to estimate participants' knowledge, we are essentially defining our "level of zoom" directly based on the content of interest itself, enabling the model to devote the full extent of its embedding space to characterizing subtleties within that content, so that the relevant conceptual features "come into focus."

In this way, the reviewer's question of whether this "sweet spot" is specific to the particular course materials used in the current study has two answers. Given that we trained our text embedding model on the specific lectures our study's participants viewed and answered questions about, the precise embedding space the model identified from that content is, naturally, optimized for characterizing that

content, and we would not expect it to be equally useful for characterizing other very different content. However, the ability to identify an analogously useful embedding space defined by the contents of an arbitrary set of course lectures is a central component of our modeling framework. In other words, the distinction we are drawing here is that the "sweet spot" we refer to in our manuscript is not the exact embedding space our model identified for *these* course lectures, but rather the *approach* we take to defining an embedding space for any given set of lectures (i.e., training the model on overlapping sliding windows of their transcripts). To reiterate from our previous response letter, we by no means believe LDA is the *only* model capable of identifying this sort of "sweet spot" embedding space—in fact, if a different "LDA-like" model were found to characterize the lectures' and questions' contents with similar (or possibly even greater) fidelity, that model could easily be "dropped in" in lieu of LDA without changing any of the core assumptions of the broader framework we are proposing. However, we suggest that an important characteristic of any alternative model that *could* be comparably well-suited for this task (i.e., what we mean by "LDA-like" in this context) is that it can be adequately trained on the specific course materials at hand.

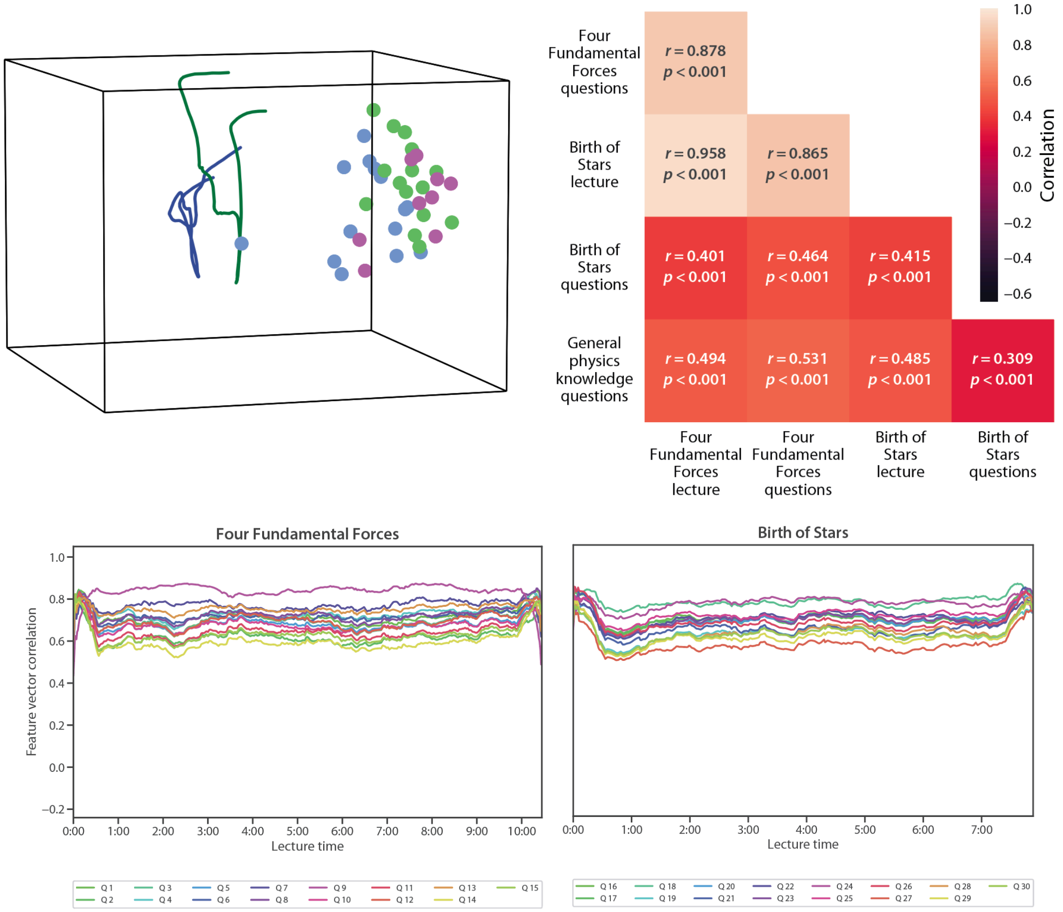
We have refined the language in our updated *Discussion* section to more clearly frame the comparisons between our LDA model and BERT in terms of the conceptual breadth of their training corpora (as opposed to these specific models themselves), and to more clearly present this "sweet spot" idea in terms of the different "semantic scales" at which our approach and other alternatives characterize the lectures' and quiz questions' contents (pp. 24–26).

> *Might e.g. LLMs work better than LDA for other researchers wanting to use these methods? If this is the case, then how should other researchers using their model determine which kind of embedding space to use?*

Following from our response above, we do not believe that LLMs are likely to "work better" for other researchers interested in adopting our framework, given the misalignment between their "generalist" embedding spaces and the level of specificity needed to successfully identify and distinguish concepts on the scale at which our framework applies (i.e., within individual course lectures).

That said, a common way of improving generalist LLMs' performance on domain-specific tasks is to "fine-tune" them on a domain-specific text corpus. In our previous response letter, we had speculated that a fine-tuned LLM might capture nuances in the lecture and question content comparably well to—or perhaps even better than—our LDA model. In the interest of answering the reviewer's question here, we decided to explore this possibility further, and in doing so realized there are some additional reasons why LLMs would likely be poorly suited for our particular application that wouldn't be solved through fine-tuning. We feel strongly that that a discussion of these more technical considerations around different model architectures, fine-tuning techniques, linguistic structures, and so on is outside the scope of our current paper and would distract from our main focus of presenting our more general knowledge estimation framework (similarly to how the extended comparisons with BERT felt out of place in our *Discussion*). However, since the reviewer is specifically asking about the possibility of incorporating LLMs into our framework, we do think these considerations bear mentioning in response. To help illustrate, we fine-tuned BERT using the same text corpus to which

we fit our LDA model (i.e., sliding windows of the two lectures' transcripts) and reassessed its ability to yield "useful" embeddings in the context of our framework (i.e., embeddings that enable us to distinguish lectures and questions about different content areas, and "match" individual questions to specific sections of lecture content). The plots shown below are analogous to those displayed for our LDA model and "base" (i.e., not fine-tuned) BERT in Supplementary Figure 10:



In summary, while fine-tuning leads to a marginal improvement compared to "base" BERT, the model's performance still falls far short of our domain-specific LDA model's. For example, the lectures and questions related to different content areas are now represented more distinctly than they were prior to fine-tuning (e.g., the range of correlation values in the heatmap above is ~16 times greater than that shown in Supp. Fig. 10), but we still fail to recover the temporally specific "matches" between questions and periods of lecture content afforded by the LDA embeddings (i.e., compare the correlation time series plots above to Figs. 4A & 4B). This highlights an important difference between fine-tuning a generalist LLM and training a domain-specific model from scratch: even after fine-tuning, LLMs continue to allocate a portion of their representational capacities to content outside the tuning corpus. (This is very much by design, with failure to do so termed "catastrophic forgetting.") This implies that even for two hypothetical models with identical architectures, parameters, inference procedures, and so on, one that is pre-trained on a domain-general corpus and then fine-tuned to a domain-specific corpus will necessarily afford lower "specificity" for the semantic (conceptual) content within that domain than one that was trained on that domain from scratch.

Despite this, in many cases this retention of domain-general "knowledge" after fine-tuning is a desirable behavior, and in fact a notable advantage to using a fine-tuned generalist model over a domain-specific model. For example, it can greatly benefit performance when the domain of content the model will need to "handle" is either too large to be fully included in a training corpus, or not fully known or available at the time of training (e.g., if our goal was to characterize scientific papers uploaded to arXiv, it would be intractable to train a model on *every* paper ever uploaded, and impossible to train it on *future* uploads). But this benefit doesn't apply in the context of our framework since the content we (or other researchers and practitioners) might want to characterize (i.e., specific course materials for which we want to estimate learners' knowledge) will necessarily be known and available in advance. And far from being prohibitively time- or resource-intensive, training a domain-specific model on course materials comparable to those used in our study would be fast and easy (e.g., training our LDA model on the lecture transcripts takes less than one second—roughly 1/1000th the time it took to fine-tune BERT to the same data).

Another reason this behavior is often beneficial is that it enables LLMs to retain representations of complex linguistic features (e.g., syntactic features) that are difficult to learn from smaller, domain-specific corpora alone. However, in our particular case, this actually ends up being a further detriment. Transformer-based LLMs are almost universally trained on written text, which tends to follow much more rigid grammatical and syntactic conventions than natural speech. Through their architectures and training paradigms, LLMs are designed to leverage this structure in characterizing a given text sample's meaning. But in our case, the text samples we want to characterize comprise transcripts of content that was originally spoken (i.e., course lectures), and therefore exhibit different linguistic structures than LLMs are optimized to leverage. For example, the transcripts contain various disfluencies, repetitions, and informalities that rarely occur in written text, while simultaneously lacking many of the grammatical elements that usually provide important "guideposts" for transformer models' attention mechanisms, such as function words, discourse markers, and sentence boundaries (e.g., punctuation). To provide a concrete example, here is an excerpt of the first ~25 seconds of the first lecture video used in our study:

> "what I want to do in this video is give a very high-level overview of the four fundamental forces four fundamental forces of the universe and I'm going to start with gravity I'm going to start with gravity and it might surprise some of you that gravity is actually the weakest of the four fundamental forces that's surprising because you say wow that's what keeps us glued not glued but it keeps us from jumping off the planet"

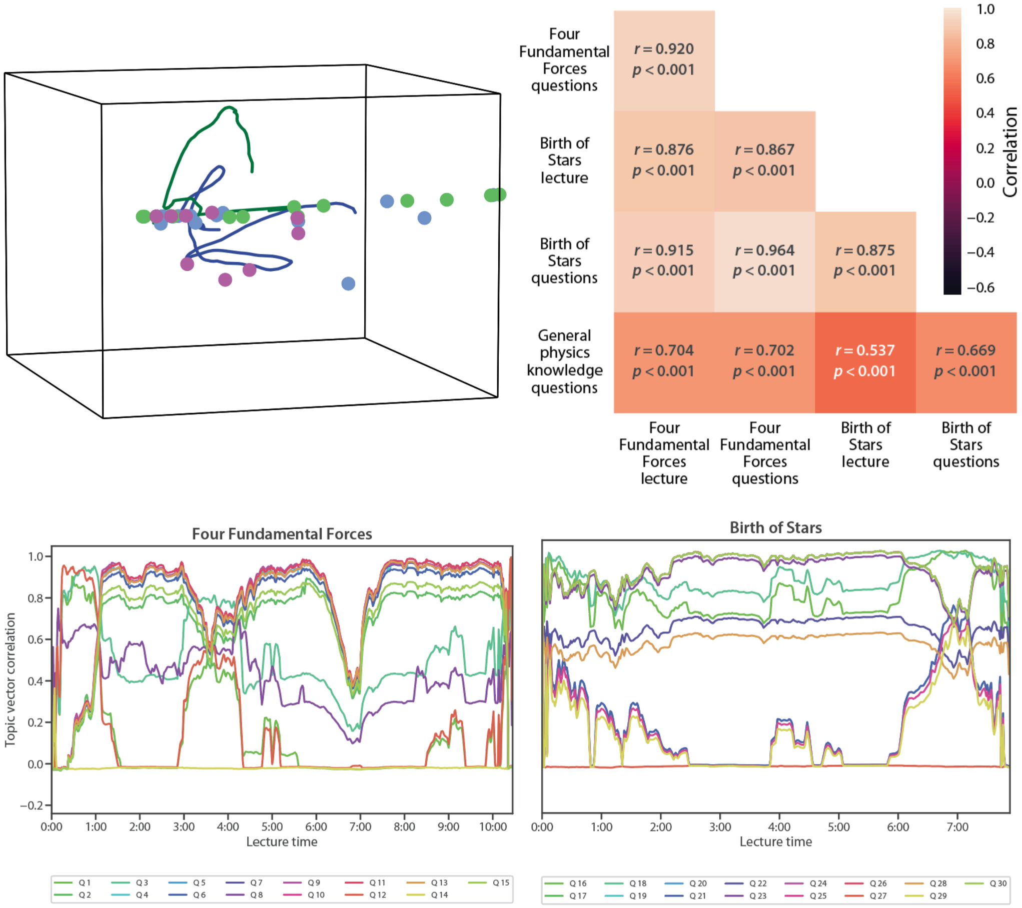
While it might be possible to manually edit the lecture transcripts to exhibit a structure more typical of written text, we view our ability to leverage automatically generated, freely available transcripts (from the YouTube API) as an important part of our framework's potential for scaling up to larger and more diverse sets of course lectures, and being useful to other researchers without requiring significant

upfront effort or investment. Even if the transcripts themselves *were* structured more like typical written text, our approach to constructing "trajectories" representing the lectures' dynamic contents (by projecting sliding windows of their transcripts into text embedding space) means that the individual text samples (i.e., sliding windows) the model "sees" often don't correspond to cleanly delineated sentences. Meanwhile, the quiz questions—which our goal is to "match" to the lecture content to facilitate our knowledge estimates—*did* originally appear in written form, and therefore exhibit linguistic structures that differ from the lecture transcripts, and are closer to what transformer models "expect to see." As we note in our *Supplementary results* section, this likely poses further challenges to "matching" the questions to specific lecture content based on their embedding weights: since BERT's representations of the *semantic* features that relate and distinguish content within this domain are highly homogenous, much of the variation in its embeddings instead reflects the lectures' and questions' *syntactic* differences (e.g., see the 3D PCA projections above and in Supp. Fig. 10).

Despite being an arguably "less powerful" model than modern LLMs, LDA is actually fairly robust to these challenges that the lecture transcripts' and quiz questions' linguistic structures pose for transformer models. Like many comparatively simple text embedding models, LDA makes the simplifying assumption that the text samples it encounters are equivalent to "bags of words." In other words, it ignores word order and syntax entirely in characterizing a text sample's semantic meaning and instead considers only individual words' frequencies and co-occurrence patterns. It also naturally emphasizes "content-bearing" words (e.g., nouns, verbs, adjectives, etc.) in its representations while downplaying the influence of "low-information" words like disfluencies and discourse markers (in fact, these "stop words" are typically discarded altogether during preprocessing). This ends up making models like LDA actually quite *well*-suited to capturing conceptual similarities across modalities (e.g., between written content and transcripts of spoken content) since most of the linguistic features that tend to differ between modalities are ones that these models naturally disregard or de-emphasize. However, we want to emphasize that while we think these comparisons are interesting and relevant in the context of the reviewer's particular question here, it is not our intention to make claims, provide commentary, or express preferences related to different model architectures in our manuscript. Importantly, we view these considerations around applying transformer models versus "bag-of-words" models to natural speech as a separate and more specific set of considerations than the general "sweet spot" notion we describe in our *Discussion*, since lecture transcripts are by no means the *only* form of content to which our framework could theoretically be applied (e.g., one could easily adapt our approach to estimate learners' knowledge for concepts expressed in written media like books or scientific papers). By contrast, the idea that training a text embedding model on the specific content under consideration constitutes a "sweet spot" in terms of the "granularity" of its conceptual representations would hold true for any potential application of our framework.

We think using BERT as our example of a "generalist" model that "zooms out too far" is useful given its popularity as an off-the-shelf solution for generating text embeddings, particularly in behavioral research contexts. And seeing as multiple reviewers asked about the possibility of using LLMs in our framework, we suspect this example will be a useful one for readers as well. However, given the additional considerations we've touched on in the context of answering the reviewer's specific

question, we feel it merits further demonstrating here that this example is a reasonable one for the point we use it to illustrate in our *Discussion*—i.e., that the lack of "within-domain specificity" afforded by generalist models does indeed reflect the conceptual breadth of their training corpora, rather than something specific to "LDA versus BERT" or "bag-of-words models versus transformer models." To this end, below we show an analogous set of plots generated using an LDA model we trained on a random sample of 100,000 English Wikipedia articles (a small but computationally tractable fraction of BERT's training corpus). Of note, this generalist LDA model's embeddings of our study's course materials exhibit patterns similar to those we highlight as meaningful in BERT's embeddings—namely, a reduced range of within- and across-content area correlations (with most being very high and all significant), diminished specificity in matching questions to periods of lecture content, and question embeddings that are less "near to" (i.e., often fall outside the convex hull of) the lecture trajectories. With sufficient time and computing resources to train an LDA model on a corpus even closer to the size and scope of BERT's, we might expect these patterns to converge even more closely.



What kind of test can be run to determine the correct level of granularity in the embedding space?

The idea of a specific "test" to determine the optimal level of granularity for the embedding space is an interesting one, and touches on a complex aspect of constructing text embedding models.

As described above, one key factor in determining the granularity of a model's embedding space is the conceptual breadth of content included in its training corpus. For the particular role the text embedding model plays in our framework, the optimal breadth of training content is that which allows the model to devote its full representational capacity to capturing conceptual similarities and distinctions *within* the specific content for which we want to estimate learners' knowledge, without "wasting" any of its capacity on representing other content *outside* the domain of interest. This is achieved by training the model solely on the specific course materials at hand.

The other key factor is the number of features (in the case of LDA, "topics") the model is able to "use" to represent the content we want to embed (i.e., the dimensionality of the embedding space). Identifying an optimal value for this hyperparameter is comparatively less straightforward. Over the past few decades, there have been many attempts to define standardized, quantitative metrics for selecting an appropriate number of topics to characterize a given text corpus, with historically popular choices including held-out document perplexity, normalized pointwise mutual information, and various formalizations of "topic coherence." We discussed and even implemented some of these measures in response to a comment from Reviewer 3 on our initial submission. However, there is growing consensus within this field that such automated evaluation metrics for topic models are fraught, and optimizing a model's embedding space based on them tends to yield representations that are less useful in practice (e.g., Chang et al., 2009; Bhatia et al., 2017; Lipton, 2018; Doogan & Buntine, 2021; Hoyle et al., 2021; Stammbach et al., 2023). We tend to share the perspective described by Boyd-Graber et al. in their 2014 book chapter entitled "Care and feeding of topic models," which emphasizes performing more holistic evaluations designed to assess the model's usefulness for the specific task at hand. In our case, a "useful" model is one whose embeddings enable us to (A) distinguish between lectures and quiz questions about different content areas, and (B) match individual questions to relevant, temporally specific sections of lecture content, so that we can meaningfully infer learners' knowledge of that content from their quiz responses.

In this way, the best "test" of whether a given model's embedding space is appropriately "granular" for use in our framework is the sorts of diagnostic visualizations we show above and in Supplementary Figures 10 and 11. Indeed, this is the exact purpose we intended those figures to serve. In Supplementary Figure 10, we show how our LDA model's embeddings of the lecture and question content differ from those obtained from a model whose embedding space is *not* sufficiently granular (i.e., BERT) in order to highlight how the latter model's representations fail to meet the criteria for "usefulness" in our framework. In Supplementary Figure 11, we show how our LDA model's embeddings differ from semantic representations that are *too* granular in order to highlight similar shortcomings. (While we didn't explicitly fit a second model for Supp. Fig. 11, the Jaccard index of two text samples is equivalent to their similarity in an embedding space where every unique word is a separate binary feature.)

In our new *Supplementary results* section, we discuss the comparisons shown in these figures in greater detail, with a focus on what their specific visual elements indicate about a model's usefulness in the context of our framework. We hope this can serve as a high-level guide that (along with the

publicly available code we provide for fitting our models and generating these figures) can help researchers interested in adopting our approach apply and interpret these holistic "tests" themselves. As with virtually all natural language processing tasks, our text embeddings benefit from some degree of manual oversight and hand-tuning. But fortunately, the speed with which an LDA model can be fit, evaluated, tweaked, and re-fit to tune performance generally makes this process fast and easy.

3.In their explanation of their Fig. 6 results, the authors suggest that one possibility is that participants forgot some material: "If this forgetting happens in a relatively "random" way (with respect to spatial distance within the embedding space), then it could explain why some held-out questions about Four Fundamental Forces were answered incorrectly, even if questions at nearby coordinates (i.e., about similar content) were answered correctly." (p. 18).

This seems to contradict their theory/finding regarding the "smoothness" of knowledge space. If forgetting happens randomly relative to the spatial location of knowledge, this evokes a kind of "Swiss cheese" knowledge space with holes randomly scattered about—in this case, knowledge would NOT be smooth, as the authors seem to claim. I can't reconcile these two claims. Perhaps the smoothness visualized in Fig. 7 is a function of averaging/combining across participants, but then making the general smoothness claim would not be so warranted.

We appreciate the reviewer pointing out this inconsistency. We agree that the speculative explanation we had offered for this result (that within-lecture knowledge estimates do *not* predict success on Quiz 3 questions about *Four Fundamental Forces*) didn't fit with our broader thesis about the "smoothness" of knowledge. While our interpretations of the results shown in Figure 6 have broadly changed (and, in our opinion, become more intuitive) now that we have addressed the issues described in our responses to comments 1a and 1b above, it's worth noting that in every version of these analyses we have run to date (*U*-tests, conservatively biased likelihood-ratio tests, and now bias-corrected likelihood-ratio tests), these particular knowledge estimates have been the only ones we compute for Quiz 3 to *not* predict participants' success on held-out questions. The consistency in this pattern of results across multiple changes to our analysis approach would seem to suggest that while estimated knowledge *does* in general predict success on Quiz 3 questions, there is some genuine effect in participants' performance on the particular subset of questions used in this analysis that impedes our ability to distinguish between correctly and incorrectly answered held-out questions.

We did some deeper digging into the data for this analysis to try to determine what this effect might be. To summarize, we believe this consistent null result reflects a particular ceiling effect in participants' quiz performance (rather than some broader aspect of their behaviors or knowledge). We describe this on pages 16–17 of our revised manuscript:

> **Speculatively, the Quiz 3 results suggest that the within-lecture knowledge estimates may be susceptible to ceiling effects in participants' quiz performance. On Quiz 3, after viewing both lectures, no participant answered more than three *Four Fundamental Forces*-related questions incorrectly, and all but five participants (out of 50) answered two or fewer incorrectly. (This was the only subset of questions about either lecture, across all three quizzes, for which this was true.) Consequently, for 90%**

> of participants, our within-lecture estimates of their knowledge for *Four Fundamental Forces*-related questions that they answered incorrectly leveraged information from at most a single other question they were *not* able to correctly answer. This likely hampered our ability to accurately characterize the specific (and by the time they took Quiz 3, relatively few) aspects of the lecture content these participants did *not* know about, and successfully distinguish them from the far more numerous aspects of the lecture content they now *did* know about.

Essentially, as participants acquire more knowledge over the course of our experiment, a few things tend to happen that have opposing effects on our ability to predict their success on held-out questions.

On one hand, with each successive quiz, the patterns in participants' correct and incorrect responses will tend to become more faithful reflections of their "true" knowledge. Intuitively, this occurs because after participants view each lecture, their correct responses are by and large more likely to reflect "real" knowledge of the lecture content rather than spurious "noise" due to successful guessing based on vague intuitions. Additionally, as we describe on pages 18–19 of our revised manuscript, we suspect that after viewing the lectures, participants' internal representations of their conceptual contents may become more closely aligned with the conceptual representations learned by our text embedding model, since the model was explicitly trained on those lectures' transcripts. Together, if participants' quiz responses come to better reflect their knowledge of the concepts being tested, and the relationships between the quiz questions' embeddings come to better reflect how participants organize their internal representations (and/or knowledge) of those concepts, then our ability to predict success using estimated knowledge should increase with each quiz and be greatest on Quiz 3. Accordingly, one high-level pattern that has persisted across all versions of these analyses (irrespective of various biases that had nudged some results above or below the "*p* = 0.05" threshold in prior versions) is that the predictive strength of estimated knowledge tends to increase over successive quizzes (e.g., compare odds ratios & λ_{LR} values between columns in each row of Fig. 6). This would also explain why our majority-significant results for Quiz 3 have been robust to these fairly substantial changes to our analysis approach.

On the other hand, as a participant gains more knowledge and answers more quiz questions correctly, our estimates of their knowledge become increasingly "saturated" (i.e., relatively high for all content) and the "signal" distinguishing content they do versus do not know about becomes somewhat weaker. This occurs because when a participant's quiz responses contain fewer incorrectly answered questions, those responses provide more limited information about content that participant does *not* know about. The information contributed by these "negative samples" of a participant's knowledge (i.e., questions they answered incorrectly) plays an important role in our knowledge estimates: while correct responses help identify underlying conceptual themes that are *common* among content a participant knows about, incorrect responses help determine which of those themes are *specific* to content they know about, by identifying conceptual themes that are also represented in content they do *not* know about. Within our framework, themes that are "common" among sets of quiz questions appear as similar patterns in those questions' topic activations when we transform them using our topic model.

Themes that are relatively "specific" to a particular set of questions then appear as patterns of topic activations that are more similar to each other than they are to those topics' activations for other questions. When we estimate a participant's knowledge for a given held-out question, that estimate gets "pulled" away from their unweighted remaining-proportion-correct score in the direction of their binary success on held-in questions that exhibit similar conceptual themes, with a "strength" proportionate to how specific those shared themes are (i.e., the correlation between the held-out question's topic activations and a given held-in question's topic activations, rescaled relative to that correlation value for other held-in questions).

That we have consistently observed significant positive relationships in four of the five analyses we perform for Quiz 3 questions indicates that, in general, our knowledge estimates *can* successfully distinguish conceptual themes that are specific to known versus unknown content even when relatively few "negative samples" are available because participants' quiz performance is near ceiling. However, the nature of the "Within-lecture" analyses further compounds the challenges posed by this ceiling effect: since all questions we consider in these analyses pertain to content from the same lecture, they will tend to exhibit largely similar sets of conceptual themes overall (i.e., all questions' topic activations will tend to be highly correlated), and the particular themes that tend to differ between correctly and incorrectly answered questions will present as relatively subtle differences between those questions' overall patterns of topic activations. Our results suggest that for *Birth of Stars*-related questions on Quiz 3, we are still able to leverage these subtle differences to distinguish between known and unknown content. But for *Four Fundamental Forces*-related questions, for which our knowledge estimates could rarely leverage more than a single incorrectly answered question, we lacked sufficient information to reliably characterize these subtle, specific differences. We note that this is not a problem for our knowledge estimates per se, but rather it is a problem specific to *validating* our knowledge estimates. In other words, if a learner demonstrates perfect performance on every question, our framework will predict that they will also likely perform well on *other* questions (which seems to us a reasonable generalization). But since our testing framework requires us to show a contrast between correctly versus incorrectly answered questions, ceiling performers appear to be poorly predicted (since they have very little contrast between what they know most versus least).