Reviewer #1 (Remarks to the Author):

The authors have adequately addressed my comments. I appreciate the very thorough response and I endorse the manuscript for acceptance. The comparisons to BERT and word-level Jaccard similarity, and the accompanying explanations, were also very helpful for understanding what insights their approach provides.

We thank the reviewer for their positive assessment of our revised manuscript.

Reviewer #2 (Remarks to the Author):

I am happy with the responses and edits from the authors in the updated version of the manuscript. In particular, I find the across-lecture knowledge prediction and the "smoothness of knowledge" analyses to be important and interesting, and make a satisfying contribution to the full set of empirical results.

There a few new sentences that are hard to parse, along with some typos, so the authors should take care to make those fixes. I don't need to see an updated version, but if the authors find it useful I can give a few examples:

p. 14 - line 339 - "and and"

p. 23 - line 470 - "The 'spatial smoothness'... is being captured by the knowledge maps we are inferring from their quiz questions."

p. 36 - line 803 - "...from our dataset full dataset."

I should also note that new edits on lines 72 and 738 are cut off on the provided pdf and so I cannot speak to their content.

We thank the reviewer for their positive assessment of our revised manuscript. We have fixed each of the typos and wording issues identified above.

Reviewer #3 (Remarks to the Author):

The authors engaged very thoughtfully, and in great detail, with the comments of all reviewers, including my own. I sincerely thank them for taking the comments seriously, for clarifying misunderstandings on my end, for adding several very helpful analyses, and for expanding the manuscript. It is important to acknowledge the significant amount of work that went into this revision.

We thank the reviewer for their positive assessment of our revised manuscript and responses to the reviewers' comments.

My two main remaining concerns were expressed in my previous comments: the model's performance might not be as good as the manuscript claims (or it lacks a quantitative analysis demonstrating it), and it might not advance us beyond existing methods. Here, I describe how these concerns relate to the revisions that the authors made:

First, the central insights that the authors' method provides about learners' knowledge are about predicting which questions will be answered correctly / incorrectly, based on other questions that have been answered (Figure 6). In my opinion, this could be the most compelling part of the paper. However, it is currently analyzed incorrectly: the Mann-Whitney test requires that observations are independent, whereas the current data are grouped by (1) participant, and (2) question. Any tests that do not account for these groupings are anti-conservative (e.g., do not properly take outliers into account). The appropriate way to analyze grouped data is with linear mixed-effects models (LMEs; in this case, logistic). The authors mention that they chose to avoid such analyses because the overall approach is already relatively complicated. However, this is not a good reason to analyze data incorrectly. LMEs are the standard approach in several fields (e.g., psycholinguistics), and have been so for a while. They are straightforward to implement (1-2 lines of code in R) and interpret. I strongly believe that this analysis should be the one adopted.

As a point of clarification, our primary goal is *not* to predict which specific questions participants will answer correctly or incorrectly. Rather, our main insight is that fine-grained aspects of knowledge and learning can be captured automatically using our framework. The ability to predict performance on specific questions is a *test* of those knowledge and learning predictions. In other words, those tests in Figure 6 are intended as a "validation" of our framework (i.e., showing that the knowledge estimates are informative), as opposed to the primary "goal" of our approach.

The above notwithstanding, we have updated our manuscript as suggested. For each subset of questions and predictions tested in our previous Figure 6 (i.e., for each combination of (a) all questions, across-lecture predictions, and within-lecture predictions for (b) quizzes 1, 2, and 3), we test whether knowledge estimates are informative about participants' responses using generalized linear mixed models (GLMMs) with logistic link functions. Overall, our results are quite similar to what we previously reported (using *t*-tests in our initial submission and Mann-Whitney *U*-tests in our most recent submission), while also helping to clarify our framework's ability to generalize its knowledge predictions across content areas. We have also updated Figure 6 to better reflect the new GLMM model-based tests, and we have added a description of the GLMM approach to our revised *Methods* section.

Once this analysis is adopted, the authors should build two models: one that predicts accuracy (correct vs. incorrect) on each held out question only based on the % correct of that specific participant on questions from the same lecture; and another that predicts accuracy based on this same % correct measure in addition to estimated knowledge. Then, by comparing the two nested models (e.g., with the anova command in R), they could explicitly test whether (or, rather, by how much) the estimated knowledge helps to predict held-out questions over and above a simple % correct measure. This analysis would provide a direct evaluation of the predictivity of "estimated knowledge" against a baseline. It will thus add critical information beyond Figure 7. Currently, Figure 7 does not provide such a direct comparison: it shows that the probability of correct answers in the "neighborhood" of a reference question is starkly different depending on whether that reference question was correctly vs. incorrectly answered, but these stark differences do not translate to differences in "estimated knowledge" for each reference question, shown in Figure 6 – as I describe next.

As requested, we fit two models: one that incorporates a simple percent correct measure, and a second that additionally incorporates "estimated knowledge". We compared these two nested models via a likelihood-ratio test (which the `anova` function the reviewer referenced performs for nested models) and computed a *p*-value using the parametric bootstrap procedure described in our revised Methods section. We found that incorporating estimated knowledge provides explanatory power above and beyond percent correct alone (λ_{LR} = 17.452, p = 0.006).

Although this analysis "works out" (i.e., displays the "effect" the reviewer is looking for), there are a number of issues we have identified with this suggested analysis that led us to decide not to include these results in our paper. Fundamentally, these issues relate to what (in principle) the measured proportion of correctly answered questions can possibly tell us about held-out questions.

One issue is that, mechanically, there is a practical confound in carrying out the proportion correct computations on held-out questions. We describe this on page 16:

> **"...our null hypothesis is that the knowledge estimates we compute based on the quiz questions' embedding coordinates do *not* provide useful information about participants' abilities to answer those questions. What result might we expect to see if this is the case? To gain an intuition for this possibility, consider the expected outcome if we carried out these same analyses using a simple proportion-correct measure in lieu of our knowledge estimates. Suppose a participant correctly answered n out of q questions on a given quiz. If we hold out a single *correctly* answered question, the proportion of remaining questions answered correctly would be (n - 1)/(q - 1). If we hold out a single *incorrectly* answered question, the proportion of remaining questions answered correctly would be n/(q - 1). In this way, the proportion of correctly answered remaining questions is always *lower* when the held-out question was answered correctly than when it was answered incorrectly. Because our knowledge estimates are computed as a weighted version of this same proportion-correct score (where each held-in question's weight reflects its embedding-space distance from the held-out question; see Eqn. 1), if these weights were uninformative (e.g., randomly distributed), then we should expect to see this same inverse relationship between estimated knowledge and performance, on average. On the other hand, if the spatial relationships among the quiz questions' embeddings *are* predictive of participants' knowledge about the questions' content, then we would expect *higher* estimated knowledge for held-out correctly (versus incorrectly) answered questions."**

In other words, raw proportion correct measures (if uncorrected) make exactly the wrong "predictions" about the held-out questions. We observe this in both the "percent correct alone" and "percent correct and knowledge" GLMMs, wherein percent correct in both models is assigned a negative coefficient that (per a Wald Z test) is significantly different from 0 (percent correctly only model: β = -5.935, Z = -4.127, p < 0.001; percent correct and knowledge model: β = -6.7455, Z = -4.535, p < 0.001).

Even if one were to correct for this mechanical bias (e.g., by using some sort of Bayesian and/or bootstrap-based approach), then proportion correct will be exactly *equal* for any held-out question. So

even in this "best case" scenario, where proportion correct could be estimated in an unbiased way, it would then not carry *any* predictive information about the held-out question.

More broadly, our main claim is that estimating knowledge (i.e., by taking question-specific content into account via text embeddings) can provide more nuanced insights into what students know than simply computing the proportion of questions they answer correctly. The across-lecture and within-lecture tests (Fig. 6) also show that those estimates can (a) extend across content areas and (b) enable us to "zoom in" within a given content area– illustrating two additional aspects of our approach that have no analogs in a framework that considers only the proportions of correctly answered questions.

The results the reviewer mentioned from Figure 7 are also relevant. Those analyses show how "far" (in the text embedding space) the influence of "knowing" or "not knowing" about the content at the embedding coordinate of a given question "spreads" through the embedding space. There are two findings worth noting. First, the analyses in Figure 7 provide another means of showing that knowledge space is "smooth"—i.e., that knowledge about a given concept implies knowledge about other concepts that are nearby in the embedding space. (And similarly, *lack* of knowledge about a given concept implies lack of knowledge about other nearby concepts.) Second, the analyses in Figure 7 show that if you travel far enough in the embedding space, eventually the content-specific insights we get about what a participant knows based on how they performed on a given individual question becomes statistically indistinguishable from simply knowing the overall proportion of questions they answered correctly.

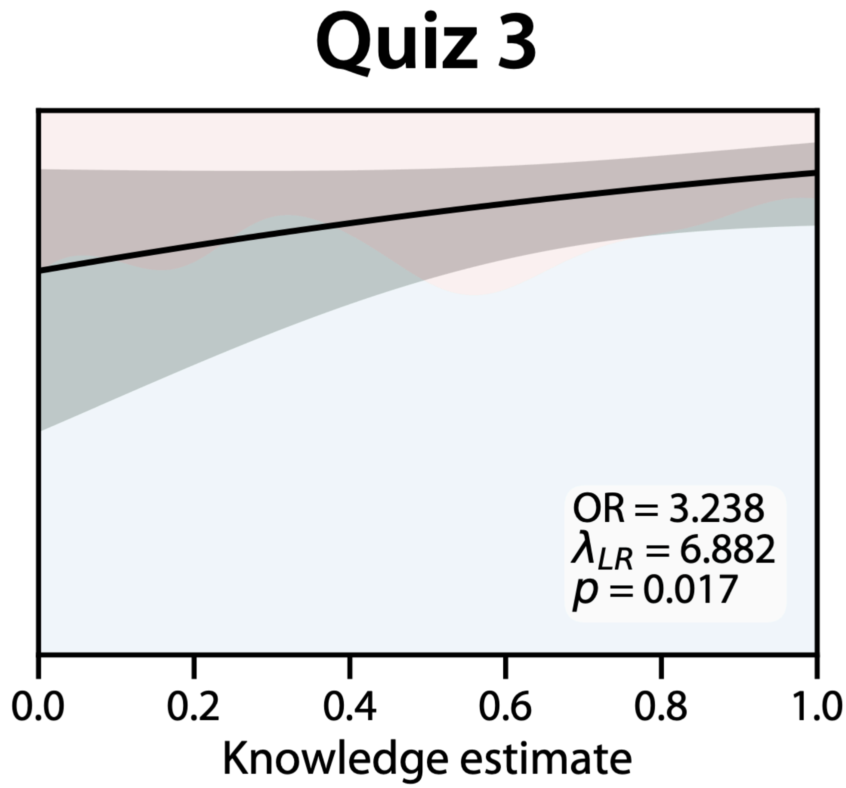
While the results in Figure 6 are statistically significant for the appropriate quizzes/topics, the patterns themselves are a bit puzzling: in many of the analyses, the distribution of estimated knowledge for questions answered incorrectly is highly skewed towards 1 (perfect knowledge). Even though this distribution is significantly less skewed than the distribution of estimated knowledge for correctly answered questions, the "effect size" of the skewness appears large to me. Specifically, between 20% and 50% of these incorrectly answered questions are predicted to have a knowledge of 1. This is quite a problem for any actual application of the authors' method, and puts a strong limit on the claim that this method provides high-resolution insights about participants' knowledge.

We appreciate the reviewer's point that the distributions of knowledge estimates for correctly versus incorrectly answered questions that we had included in our previous submission are confusing. Those distributions (as presented in our previous submission) are somewhat difficult to interpret in the way the reviewer is attempting here, since each "observation" reflected one knowledge estimate from one participant (i.e. the full distributions comprised many observations from each person—so, for example, participants who answered nearly all questions correctly would contribute to the extremes of those distributions). We have overhauled the entire figure (Fig. 6) and analysis using the GLMM framework the reviewer had suggested in their prior comments. In our revised Figure 6, we show a more intuitive (we think!) depiction of how predicted performance changes with estimated knowledge, in each sub-panel. Specifically, for each level of estimated knowledge we display the proportion of correctly versus incorrectly answered questions.

For example, a very rough and inaccurate estimation from the top right panel of Figure 6 suggests that the probability of a correctly answered question having a higher estimated knowledge than an incorrectly answered question is ~0.56 (chance is 0.5). This is very low for any practical purpose. I did this by creating a ROC curve that attempts to classify questions as "correct" or "incorrect" based on estimated knowledge and computing the area under the curve (AUC). The AUC also equals the Mann-Whitney U statistic divided by the product: num_answered_correctly × num_answered_incorrectly. My sincere apologies if my calculation is way off and I made a mistake somewhere. But if it is in the ballpark, then the "predictivity" of knowledge estimates, while significant, is quite limited. In that case, the claims in the papers should be made less bold.

In our previous version of this analysis (from our prior submission), the reviewer makes a good point that predictive power was low when the majority of knowledge estimates are within a narrow range. However, now that we have revised our analysis (at the reviewer's good suggestion!), we now have considerably more statistical power to predict performance on individual held-out questions. Previously, we were essentially considering only the knowledge estimate values themselves in "classifying" whether a question would be correctly or incorrectly answered. Now, however, using GLMMs, we are able to more closely examine the effect of estimated knowledge on the likelihood of correctly answering a held-out question while accounting for the facts that (A) some participants may be more likely to correctly answer a question than other participants, and (B) some questions may be more likely to be answered correctly than other questions (i.e., by fitting our models with random intercepts for participants and questions). In other words, while the overall range of estimated knowledge values is generally smaller when, for example, most participants answered most questions correctly (such as on Quiz 3), we are now better able to parse out the relationship between estimated knowledge and likelihood of answering a held-out question correctly *within* the range of estimates spanned by individual participants and/or questions.

To provide some additional context, the specific case the reviewer is referencing relates to our predictions of Quiz 3 responses that include all questions (regardless of content area). As can be seen in the panel, performance on Quiz 3 is quite good overall—i.e. the majority of questions are answered correctly (shown by the blue distributions). However, as also shown in the panel, the predicted probability of answering a held-out question correctly is *higher* when estimated knowledge is higher, as opposed to when estimated knowledge is lower. In other words, even in this case where performance is near ceiling, our knowledge estimates can still reliably distinguish between questions that are more vs. less likely to be answered correctly, based on a participant's performance on the other questions from the same quiz. We have pasted in the relevant panel below (including the relevant statistical results from the GLMM analysis of that panel's data) for reference; the *y*-axis denotes p(correct), blue denotes the proportions of correctly answered questions at each estimated level of knowledge, red denotes the proportions of incorrectly answered questions, and the black line displays the LME model fit to the data:



Other than estimating knowledge on held-out questions, the other insights provided by the authors' method are somewhat limited in my opinion. For example, the maps in Figure 8 only show that students learn the lecture that they've watched (i.e., this is where the "change in knowledge" is apparent). This is a sanity check about the algorithm for creating the maps, but it is in itself not an independent insight, as it is derived from the information in Figure 2 and the analyses reported in the text (shown in Supp. Figure 2): the embedding of each question is closer to the trajectory of the corresponding lecture, and farther from the trajectory of the other lecture. The word clouds in Figure 8c are also a reflection of the properties of the embedding space that were verified earlier in the text. The maps indeed have finer-grained structure, but such structure is currently not quantified and remains interpretable (so these are, at best, untested "hypotheses", not insights).

Our main goal is to build a model that helps to explain and track what people know and how they learn. In Figure 8, we show how our approach may be used to visualize detailed (high spatial resolution) "maps" of knowledge and learning. In fact, those maps effectively have "infinite" resolution, since the level of detail is constrained only by the sampling resolution used to construct those maps. However, it would be computationally intractable to attempt to directly test those predictions about the "infinitely many" concepts reflected by those sorts of maps. Instead, we focus on a similar (but more practical) variant of the same idea by estimating knowledge at the embedding coordinates of held-out questions.. Specifically, we construct knowledge estimates using some of the data and then see whether estimated knowledge about the content probed by held-out questions corresponds to some extent with participants' chances of answering those held-out questions correctly.

We show (e.g., in Figures 6 and 7, along with the analyses described above that the reviewer requested) that our approach provides predictive information about participants' knowledge of held-out material.

The maps shown in Figure 8 are intended to serve several purposes. First, they provide a proof of concept that our knowledge estimation procedure can be scaled up to estimate very large numbers of coordinates in text embedding space, to the point that we can construct smooth "maps" of knowledge and learning. Second, Figure 8 shows that the specific estimates reflected in those maps make some intuitive sense. For example, on average, participants who just watched a given video show an increase in knowledge around the coordinates spanned by that video. As the reviewer notes, this provides a nice "sanity check" that the structure captured by our framework is "reasonable." However, it is also important to note that the actual predictions we make about specific questions (or, more generally, knowledge about held-out content) come from *learner-specific* information (i.e., individuals' quiz performance, rather than the across-participants average performance), which is better captured by the individual level maps shown in Supplemental Figures 7, 8, and 9. Whereas the average maps show something akin to "what people learn in general from watching a given lecture video, or a sequence of lecture videos," the individual participants' maps tell us about what *that specific individual* knows, and how their knowledge changes over time with training. Those individual-level knowledge estimates and maps are what we claim could help drive insights, e.g., by building those maps into automated tutoring systems, (human) teacher feedback systems, and so on.

(Another potential issue with interpretability is that of the topics themselves: for instance, do the authors observe that the model recovers distinct topics for each of the four fundamental forces, i.e., a topic or few that correspond to gravity, others that correspond to the strong interaction, etc.? If so, saying so explicitly is important to demonstrate the strength of the method. If not, then this is a limitation that should be acknowledged, because it's one of the "minimum" requirements for any representation of the "topics" of that lecture, at least in the intuitive sense of "topics").

To clarify, within the Latent Dirichlet Allocation framework (Blei et al., 2003), a "topic" is defined as a distribution (of weights) over words in the model's vocabulary. (The "weights" are modeled as a draw from a Dirichlet distribution, hence the name of the approach.) In other words, a topic model's "topics" do not come with ready-made human-readable labels. This is the case for nearly all modern text embedding approaches: the "dimensions" or "features" discovered by text embedding models tend to be more complicated or nuanced than easily namable concepts like "gravity" or "strong interaction" and so on. Therefore one would not expect arbitrary human-readable concepts (e.g., selected post-hoc) to "fall out" along fully separable topics in the model as the reviewer is suggesting. Blei (2012, Communications of the ACM) provides a nice intuitive overview of what the "topics" in topic models mean, versus what might be colloquially meant by the word "topic." In our paper we use the word "topic" to mean "a topic discovered by LDA"—i.e., a weighted distribution over words, whereas we use the word "concept" to represent something closer to what the reviewer is suggesting here. Formally, however, a "concept" in our framework would simply be a coordinate in text embedding space (like would be obtained for any other text).

What we *can* ask is whether the representations obtained by the model can describe the content of the lectures, questions, and students' knowledge in meaningful or useful ways. We test this in several ways:

- **In Figure 2C, we show trajectories of each lecture along with the embeddings of each question. We can see visually that:**
  - **The blue line (Lecture 1 video) and green line (Lecture 2 video) unfold over an extended region of text embedding space, as opposed to being confined to a single point. This shows that different moments in each video are being embedded in different ways (i.e., as different blends of discovered topics).**
  - **The blue dots (questions about Lecture 1) tend to map onto coordinates near the blue trajectory (Lecture 1 video), whereas the green dots (questions about Lecture 2) tend to map onto coordinates near the green trajectory (Lecture 2 video). This tells us that the deeper conceptual content of the questions about each video are being captured using similar blends of topics as the videos themselves. In other words, the questions' embeddings "match up" with the appropriate regions of text embedding space (that are covered by the corresponding lecture videos). This is an important result, because the questions are not included in the topic model's training set.**
- **In Figure 3, we can see (in a different way) that the topic embeddings "match up" questions about each video with the correct videos:**
  - **The specific topics reflected by the videos overlap with the specific topics reflected by questions *about* each video (Fig. 3A)**
  - **The specific topics reflected by questions about Lecture 1 do *not* match up with the topics reflected in Lecture 2, and vice versa (Fig. 3B)**
- **In Figure 4, we can see an even more nuanced "matching," whereby we show estimates of the precise time course over which the content of a given question is covered by the corresponding lecture. In other words, it's not simply that every question is being mapped onto the "average" coordinates of the appropriate lecture in text embedding space. Rather, each question reflects a specific (limited) part of each lecture.**
- **In Figure 6, we show that we can use the knowledge estimates (derived using the topic embeddings of each question) to partly explain which held-out questions participants answered correctly versus incorrectly. In other words, the knowledge estimates relate to participants' behaviors.**
- **In Figure 7, we show how the proportion of correctly answered questions falls off with spatial distance from the text embedding coordinate of a correctly or incorrectly answered question. That those plots are structured (i.e., neither random nor "flat") indicates that the text embeddings are capturing some conceptual content of the questions *in ways that are predictive of participants' actual performance*.**
- **In Figure 8, we show how estimated knowledge at specific (arbitrary) regions in text embedding space changes with training. Again, that these maps are structured (and that they behave in reasonable/expected ways as training progresses) indicates that we are capturing meaningful aspects of what the participants know.**

Similarly, the "moment by moment" content mastery is an interesting visualization, but there is no quantification of whether it is accurate. If I understand correctly, this visualization is derived from the information in Figure 4, so it again does not provide any independent insight that can be tested / demonstrated. More importantly, the identification of lecture snippets that correspond to each question, shown in Figure 4, is still unfortunately not quantitatively verified. I have proposed several ways in my previous comments about how the accuracy of this result can be measured, and I think they are relatively straightforward to implement in an online study (or the authors could just ask physics instructors to map each question to the corresponding part of the lecture).

The authors manually inspect the lecture snippets that were automatically identified based on each question, and report that they are overall good (providing all the data in the Supplementary Materials is very helpful!). But this is a subjective impression. For instance, Question 26 is said to correspond to two parts of the lecture, but one of them provides the right answer, whereas the other provides the wrong answer—so giving both snippets to a student who answered the question wrong might cause further confusion (as an anecdote, ChatGPT could answer this question correctly even without the lecture text, so there is nothing inherently "hard" about the question). Similarly, for Question 7, one of the identified paragraphs is quite irrelevant. For Question 4, the algorithm identifies over 2.5min of text, and this entire text does not contain the right answer ("the weak interaction really applies to very small distances"), perhaps because the word "computer" throws the algorithm off. Same for Question 11, where the identified text does not include anything about the strong force. I hope that these examples demonstrate why I believe that a quantitative estimate of the accuracy of the authors' method is important.

We appreciate that the reviewer is asking for a specific way of characterizing the "quality" of the automatically identified matches between the identified timepoints in each lecture that each lecture-specific question is "asking about." We provided a supplemental table reporting *all* of the matches, for every question in our dataset, to enable interested/motivated readers to judge for themselves the quality of each match. To our reading of that table (Supp. Tab. 3), most (though, as the reviewer points out, not *all*!) of these automatically identified matches appear to align quite nicely with the corresponding lecture snippets. For example, in nearly every case, the identified snippets discuss or describe whatever concept the question is asking about. The reviewer is also correct, of course, that there are a few examples where the alignment appears to be "off" or incorrect.

If our end goal was to show that we could match up questions with lecture snippets, then a new experiment along the lines of what the reviewer is suggesting might be worth running. However, the goal of the *current* paper goes beyond matching. If the matching process was truly random or uninformative, then (a) questions should not align to the correct lectures, as shown in Figure 3B, and (b) knowledge estimates generated using the questions' embeddings should carry no predictive information about participants' quiz performance. Further, (c) the knowledge maps in Figure 8 would not show increases in knowledge around the just-watched lectures (e.g., an increase in knowledge around Lecture 1's region of topic space in Quiz 2, and an increase in knowledge around Lecture 2's region of topic space in Quiz 3). That we can predict participants' performance in several ways, and that using the text embedding to characterize performance adds interpretable structure to the problem of characterizing and tracking participants' knowledge, suggests that the embeddings are indeed capturing meaningful content about the course materials. Our main claim is that *leveraging* that

captured content can help us understand what learners know at a finer-grained level of detail than simply ignoring those detailed characterizations about question and lecture content that are captured by text embeddings.

My second main concern regards the usefulness of the authors' method with respect to other, existing tools. The authors now provide an analysis of BERT, which is nice an informative, but I would like to still push back on two fronts. The first one is less important: given that BERT represents so many topics, as well as language structure, it is expected that its lecture trajectories would be "squashed", and that the questions would be represented somewhat separately (e.g., they have different syntax). But this does not necessarily mean that the critical information is not there. For instance, in static word embeddings like Word2Vec, antonyms ("good" vs. "bad") are located in close proximity, and this "squashing" of opposites could be said to be bad for a representation of meaning; nonetheless, the direction in the embedding space that connects antonyms defines a subspace that is meaningful, such that even words that are quite far from the "good" and "bad" vectors can have their valence estimated based on their position relative to that subspace. This is just one example of how complex knowledge exists, and can be extracted, even when at first glance it appears to lack desired properties. (By the way, similarities in BERT are also known to be influenced by "rogue" dimensions, such that the activations of each artificial unit should be z-scored based on a corpus; for instance, see: https://arxiv.org/abs/2109.04404. Personally, I don't think this is going to make much of a difference; I'm mentioning it just in case the authors use similar analyses for other projects).

The main issue we're describing with the BERT embeddings (e.g., in Supp. Fig. 6) is not only that the embeddings are "squashed," or that the lectures and questions use different syntax (although we agree that these both pose challenges). Rather, the larger "problem" is that the questions are not represented (very) differently from *each other*. An easy way to see this is to visually compare the timecourses of the correlations of each lecture 1 and 2 question for the text embeddings derived using LDA versus BERT, in Supplementary Figure 6. Whereas the LDA-derived embeddings for each question have a (largely) unique time course, the BERT embeddings for each question neither distinguish between different moments of the lectures, nor do they distinguish between lectures. This is quantified more formally in the bottom panels of Supplementary Figure 6. The LDA-derived embeddings for each lecture are strongly correlated with the embeddings of the *matching* lecture, and they are strongly *negatively* correlated with the embeddings of the other lecture, or of questions about different material. This tells us that the LDA embeddings are *specific* to each content area. In contrast, the BERT embeddings for every content area (across both lectures and all three sets of questions) are *all* strongly positively correlated. This tells us that BERT embeddings are *not* specific to each content area at the level of detail required for this study.

The reviewer makes some interesting points about how BERT's difficulties in this domain might arise due to the number of feature dimensions, and we appreciate the "rogue dimensions" point along with the preprint reference. Our intuition has been that those difficulties are more about *training* than about the *dimensionality* per se. Because BERT is *trained* on a very general corpus, its embeddings need to "explain" a very broad range of content. Even though BERT has more feature dimensions than our LDA model, BERT needs to "spread" those dimensions across a much wider range of content than

our LDA model (which only needs to "explain" the content of two lectures and a few dozen questions).

Although it is beyond the scope of our current paper to fully explore this, we suspect that a fine-tuned version of BERT, or fine-tuned GPT model, could do a much better job at describing the nuanced lecture and question content– perhaps even better than our LDA approach. The deeper point we were trying to make by comparing LDA vs. BERT isn't that LDA is necessarily the "best" model, or even that it's necessarily better than BERT or other transformer-based models. We are simply arguing (e.g., pages 27–28) that there may be a fundamental tradeoff between "generalizability" and "specificity." When models are trained or fine-tuned on a highly specialized corpus, they can achieve high resolution in that content area. However, those specialized models would be unlikely to generalize well to other domains. When the same model architectures are trained on a much broader corpus, they can generalize well but have relatively poor performance when you "zoom in" sufficiently far in any given domain. The same is true even for *very* large models, like ChatGPT-4. You can see this if you attempt to do something highly specialized, like playing a game of chess against the default version of ChatGPT-4 (i.e., a model that contains knowledge about the rules of chess through its training corpus, but that has no chess-specific fine-tuning).

In other words, our use of LDA isn't intended as a statement about LDA alone. LDA could be easily swapped in for another model that is able to capture the lecture and question content to a sufficient extent, without changing any of the fundamental assumptions or approaches we are proposing. When we compare LDA's performance (when LDA is fit to sliding windows of the the lecture transcripts) versus BERT's performance (when BERT is fit to a much more general corpus) we are not trying to claim (nor do we believe) that LDA is somehow a "better model" than BERT, or any other model. Rather, our point is that even a much larger model, fit to a much larger corpus, fails to capture fine-grained aspects of the lecture and question content to the extent that a much simpler model like LDA can, when fit to this very specific dataset. It could very well be the case that a fine-tuned version of BERT (or nearly any other widely used modern language model) could explain the lecture and question content even better than LDA.

In our paper, we show a "proof of concept" that a model that captures the conceptual content of lectures and questions, up to some "sufficient" level of detail, will be able to uncover aspects of knowledge and learning that go beyond what we can achieve using traditional "proportion correct" measures alone. And in the best case, we can start to construct complete maps (of the sort we propose in Figure 8) of what people know and how their knowledge changes over time, e.g., with training.

Comparing LDA versus other models is tangential to that goal of establishing a proof of concept. In other words, if a simple "fine-tuned" model can explain the relevant content sufficiently well to begin to explain some aspects of participants' performance on the quizzes (over and above "proportion correct" measures that do not directly attempt to formally capture the conceptual content of the lectures or questions), then we might assume that some more sophisticated model that captures the conceptual content even better might *also* explain participants' behaviors more fully. But our goal is

not to find the "best possible" model; we are solely trying to establish that proof of concept that capturing the conceptual content of the lectures and questions can provide some explanatory power of participants' behaviors.

Second, and more importantly, BERT is a model that is quite good at linguistic representation, but not very good at reasoning (it is also quite small by today's standards). It does show some evidence of learning "topics", similarly to those the authors talk about (for instance, https://arxiv.org/pdf/2203.14680.pdf), but it is still quite limited. In contrast, models trained using reinforcement learning from human feedback exhibit much better reasoning abilities (for instance: https://www.nature.com/articles/s41562-023-01659-w). It is possible that these are the abilities that are required for, e.g., estimating knowledge based on questions. The ability to identify which span of text is relevant for answering a particular question can already be done with high accuracy by modern models (e.g., https://paperswithcode.com/sota/question-answering-on-squad20). Moreover, state-of-the-art, publicly available systems might have some important advantages over the authors' model.

We appreciate these points and these references. While interesting, the "reasoning" capabilities of the text embedding models we employ is (in our view) somewhat tangential to our main focus. Neither BERT *nor* LDA are good at "reasoning" in any direct sense of the word. As explained above, our goal is not to pit different model *architectures* against each other. Our point in including comparisons between LDA and BERT is to show that not *every* model (even if large and complex, and even if fit to a much larger dataset) will be able to immediately explain the conceptual content of the lectures and questions sufficiently well. In other words, simply swapping in a model trained on a very general corpus (even if that model is large) is unlikely to yield good knowledge estimates. We think this is related to the "generalizability versus specificity" issues we discuss on pages 24–25. In particular, for a given model architecture and number of parameters, training the model on a broader corpus will yield greater generalizability across content areas, but less specificity in any given content area.

We also wish to clarify that the objective of our text matching demonstrations using LDA (Fig. 4, Supp. Tab. 3) is *not* only to match up which snippets of the lecture transcript are providing the information relevant to a given question. The text matching results we present in Supplemental Table 3 are intended to show that, overall, the moment-by-moment match we compute using correlations between topic vectors (Fig. 4) appear to also, in most cases, correspond to reasonable matches between the lecture transcripts and question text. But that is only a sanity check. What we actually "care about" in our approach is the *quantifications* of those matches—i.e., we need to know how much to *weight* the given question when we want to estimate knowledge about a given part of the lecture. Finding the best matching lecture snippet does not (in and of itself) provide a "number" that we can use to weight the contributions of different questions in estimating knowledge.

The authors state in their rebuttal "suppose one were to ask ChatGPT to match up each question with some part of a lecture. What would one "do" with that response?". My answer is: whatever one would do with the information in Figure 4. If the data in that figure are used as a test of the authors' model, then I am assuming those data have some use. For instance, you could refer a student who answered a question incorrectly to the relevant part of the lecture for review. The authors also state "ChatGPT also has no built-in mechanism for keeping track of what the student *knows, or how that knowledge might relate to the content of a course the student is learning from"; and "ChatGPT has no internal machinery for representing or tracking the learner's knowledge, nor does it maintain a "theory of mind" of the learner, nor does it (in and of itself, to the best of our own understanding) have any deep understanding of the material itself". The authors' model also does not have any deep understanding of the material (and, if anything, understands less because it does not capture the structural dependencies between words in a text). The model also has no substantial theory of mind: what it does is implement a notion of "knowledge = vectors in a continuous space", but ChatGPT and similar models rely on the same assumption, and hence implement this same notion. Being a high-dimensional embedding, ChatGPT has the capacity to hold and process all the information that the authors' model stores; the fact that it is very hard (at least for me) to conceive of how it might do so, does not mean that it cannot do so or does not do so. There is active work on the ability of large language models to have a "situation model" or a "discourse model" and update it as they process more input. Also, professors are using these models to read students' papers and fill out detailed rubrics with feedback, so these models can convey high-resolution insights about student performance. Models like ChatGPT are also generative, so they can go well beyond the authors' model by, e.g., writing new questions. In this sense, these models could be more useful in practice because they can translate insights into actions.*

We suggest that the reviewer may be misinterpreting the intended message of Figure 4. To clarify, the goal of Figure 4 is to show that LDA-derived topic vectors yield temporally "specific" matches between the questions and lectures. In other words, most questions' correlation time series plots have a single "peak" (or some have two peaks), and most questions' peaks occur at different parts of the lectures. That tells us that the particular "blend" of questions that will be used to estimate knowledge about each moment in the lectures (e.g., as in Figure 5) is largely unique to each moment of each lecture. If we contrast those correlation time series plots with the plots obtained using BERT (Supp. Fig. 6, right panels), it becomes clear that the BERT embeddings behave very differently: individual questions show relatively equal "weight" (i.e., correlations) over lecture timepoints, and all of the questions' time series plots look very similar. Therefore the BERT-derived "blend" of questions used to estimate knowledge about any given lecture moment would be highly similar across moments (approaching something much closer to a "percent correct" measure).

Our point about the need to "do something" with ChatGPT's response refers to what is needed if we want to quantify knowledge, as opposed to solely describing what someone knows in a qualitative way. Even if we know which snippets of a lecture transcript "match" with a given question, we still want to turn those matches into a knowledge estimate. That lets us move beyond conceptualizing knowledge as a binary "correct or incorrect" feature, towards instead capturing more continuous estimates of knowledge that integrate performance across *several* related questions.

The reviewer also points out that LDA has no internal representation of what a participant knows. We agree! What we are doing is using the text embeddings to define the coordinates (dimensions) of a *separate* space for characterizing and tracking knowledge. In other words, LDA gives us a way of mapping out where concepts "go" in the space, so that we can appropriately combine participants' responses in order to build accurate estimates of what they know. This requires only that LDA maps

similar concepts onto nearby coordinates– it does *not* require LDA to maintain its own internal representation of what the participant knows.

As an aside, the activations from ChatGPT's transformer module (if one had access to them) or from a similar model (e.g., Llama 2 70B) could (in principle) be used in a similar way to how we are using the LDA embeddings. The downside of using large "off the shelf" language models that are trained on huge (and diverse) corpora is that (due to the generalizability versus specificity tradeoff we discuss above and on pages 24–25), those models won't necessarily have good "resolution" within an arbitrarily chosen small area within their embedding space. However, a fine-tuned version of ChatGPT (or another similar model), if one had access to the weights, could potentially be used to define even better or more accurate "knowledge maps."

Finally, the reviewer makes a good point about how ChatGPT "translates insight into action." Using ChatGPT to generate questions related to some content area, or to auto-grade essay questions is fully compatible with and complementary to the framework we are proposing. Those automatically generated and scored questions could serve as the input data to the sort of knowledge estimates and "maps" that we are deriving with our framework. One could even imagine an "active sampling" approach whereby mapping out what someone knows entails (a) continually identifying the part of their "knowledge space" (i.e., a region of text embedding space) that we are maximally uncertain about, (b) using ChatGPT to generate and score a question about that content area, and (c) updating the estimated knowledge map with the new information.

As an anecdotal example, I fed ChatGPT the text of the Four Fundamental Forces lecture, followed by some prompts and questions. Below are the first responses that I received in this interaction, unless otherwise specified, so there was no "tweaking" (sure, slightly different prompts might have given different results, but that is beside the point in my opinion; once we know the right prompt to use, there is no reason to use others). .

[the input below was given immediately after the text of the lecture]

Answer the question below:
Why don't you and your computer experience any attraction or repulsion due to the Weak Interaction?
A. The weak interaction only acts over extremely small distances
B. The weak interaction between you and your computer is counteracted by the other forces
C. You and your computer have no net charge
D. Neither you nor your computer has enough mass to induce a significant Weak Interaction

ChatGPT
A. The weak interaction only acts over extremely small distances

You

Please type the part of the lecture that helped you answer the question above

ChatGPT

The part of the lecture that helps answer the question is:

"Now, the next strongest force-- and just to give a sense of how weak gravity is even relative to the weak interaction, the weak interaction is 10 to the 25th times the strength of gravity. And you might be saying, if this is so strong, how come this does it operate on planets or us relative to the Earth? Why doesn't this apply to intergalactic distances the way gravity does? And the reason is the weak interaction really applies to very small distances, very, very small distances. So it can be much stronger than gravity, but only over very, very-- and it really only applies on the subatomic scale. You go anything beyond that, it kind of disappears as an actual force, as an actual interaction."

[note that the above is a better than the authors' algorithm, which does not identify the relevant lecture snippet for this question; in addition, the snippet is ~1min long, compared to the author's algorithm which identified ~2.5min of text]

You

A student read the lecture above, and answered the following questions correctly:
1. Why is the gravitational attraction between you and your computer too small for you to notice?
2. Roughly how many times stronger is the Weak Interaction than gravity?
3. Why don't you and your computer experience any attraction or repulsion due to the Weak Interaction?
But that student got the following questions wrong:
1. Why does the universe have a very uneven distribution of mass but a relatively equal distribution of charge?
2. In your body, there are a tremendous amount of negatively-charged electrons. Your computer also contains a huge number of negatively-charged electrons. We know that like charges repel, but you and your computer are not repelled apart. Why?
3. What does the Coulomb Force refer to?
Based on this pattern, which concepts did the student successfully learn, and which concepts did the student not learn well?

[I first tried to get ChatGPT to immediately generate new questions based on the above prompt, instead of asking first which concepts the student learned well; that did not work well, which is why I added this "intermediate" step above]

ChatGPT

The student seems to have a good understanding of the concepts related to gravitational attraction, the relative strengths of the weak interaction compared to gravity, and why the weak interaction is not noticeable between everyday objects. However, there are challenges with concepts related to the distribution of mass and charge in the universe, the behavior of like charges in your body and the computer, and the specific reference to the Coulomb Force.

Concepts the student successfully learned:

1. **Gravitational Attraction:** The student understands that the gravitational attraction between objects like you and your computer is too small to be noticeable in everyday situations.
2. **Weak Interaction Strength:** The student correctly grasps the idea that the weak interaction is much stronger than gravity, but its effects are limited to very small distances.

IB **You**

Given what this specific student seems to know vs. what they do not know, write one new question that you expect them to answer correctly and one new question that you expect them to answer wrong.

ChatGPT

Expected Correct Answer:

Question: Why is the weak interaction more noticeable on a subatomic scale rather than on larger scales like intergalactic distances?

Expected Incorrect Answer:

Question: What is the primary factor responsible for the attractive forces between positively charged protons in an atomic nucleus?

Another example:

You

Here is another question: Which force is stronger than the Electromagnetic Force? Could you please write which part of the lecture would help someone answer that question?

ChatGPT

The force that is stronger than the Electromagnetic Force is the Strong Force. The part of the lecture that provides this information is:

"Now, the strongest of the force is probably the best named of them all. And that's the strong force. That is the strong force... And once again, just to keep it in mind relative to gravity, it is 10 to the 38th times the strength of gravity. Or it's about 100 times stronger than the electromagnetic force."

This part of the lecture compares the strength of the Strong Force to that of the Electromagnetic Force, highlighting that the Strong Force is much stronger than the Electromagnetic Force.

[This is Question 11, for which the authors' algorithm did not provide the correct snippet]

These examples are, of course, anecdotal. I did not conduct a thorough evaluation of ChatGPT. But I am showing to suggest that the authors are perhaps too hasty in dismissing large language models as lacking high-resolution insights because their semantic spaces are so large. These models might be able to do a lot of what the authors' model does, and even go beyond it by, e.g., constructing new questions. True, they cannot provide visualizations of continuous maps of a student's knowledge (Figure 5, 8), but such maps are only interpretable and useful insofar as they lead to some action (e.g., identifying which concepts to focus on for re-learning).

All this brings me back to the concern I expressed in my first review: the authors' algorithm is cool, but does it sufficiently advance us beyond the capabilities that already exist in large language models, which are available to students and instructors? I think that answering this question is important for deciding whether the paper merits publication in this journal.

These examples are great! We think this is an interesting application of ChatGPT. In addition to providing a practical tool for educators, one could imagine this approach being leveraged by students as a convenient way of searching a video to identify which parts are "about" the content of a given question. But as we describe above, this is a fundamentally different problem than we are tackling in our paper. Whereas ChatGPT is being used here to directly identify a snippet of text in a lecture transcript, our approach is asking: at each moment of the lecture, how *much* does the content overlap

with this question? We then *estimate* a snippet following those moment-by-moment characterizations. But the snippets we report are a way of demonstrating that the matching is "reasonable" (as judged subjectively by the reader). We don't actually *use* those snippets in our computations. Similarly, although it's very impressive that ChatGPT can provide a digestible summary of what the student does vs. doesn't know, there is no way to quantify those estimates. The maps we propose in Figure 8 are our vision of what those quantifications of knowledge could look like—i.e., more naturally described or visualized as continuous "maps" of exactly how much a learner knows about any conceivable concept. Those maps could then be used to generate questions, search over tutorial materials, etc. (as the reviewer is suggesting). But they could also be used to quantitatively track progress over time, compare or "match up" students who might benefit from studying together, assigning tutors based on areas of expertise, and so on.

Finally, the reviewer's point that one should not dismiss large models because their spaces are so large is well-taken, and we agree with the sentiment. In fact, in some of our other work (e.g., https://github.com/ContextLab/chatify), we are building tools for incorporating large language models into interactive tutorials. We think models like ChatGPT are incredibly promising and useful tools for education. However, even though ChatGPT is useful for many education-related tasks, here we are trying to build more quantitative (formal) characterizations of what learners know at different points in their training. This is needed to test specific predictions or theories, or to systematically track a given student's knowledge over time, or to compare different students' knowledge or learning rates, and so on.

Some minor thoughts:

- *The word-overlap analysis (Supp. Figure 5) is very helpful!*

We are glad that the reviewer found Supplemental Figure 5 helpful!

- *Perhaps change the naming of the quizzes to Quiz 0, Quiz 1, and Quiz 2? That way, the number corresponds to the number of lectures the students viewed (I found myself having to do a "minus one" operation every time I was reading something about the quizzes to figure out their position in the experiment). I realize this means re-generating all the figures, so I'm not sure it's worth it…*

We have chosen to retain our current numbering system for the quizzes. We appreciate the reviewer's point that it makes intuitive sense that "Quiz 0" happens before any lectures, "Quiz 1" happens after lecture 1, and "Quiz 2" happens after lecture 2. We tend to find "0-indexing" most intuitive as well. On the other hand, we've found that, in presenting and discussing this work, our audiences have tended (most often) to find "1-indexing" more intuitive.

- *The authors test whether "knowledge about a given concept implies knowledge about related concepts" and "knowledge about a given concept is predictive of knowledge about similar or related concepts". It would be helpful to the reader is the authors explicitly articulated which theories from psychology and/or education* *do not predict this pattern of knowledge (otherwise, this is a test of the validity of the model, not a psychological implication of the findings).*

We suspect that most theories that incorporate some notion of attribute similarity or feature similarity would be compatible with our findings. It would be difficult to imagine a reasonable theory of learning whereby learning about a given concept *decreases* knowledge about other related concepts. The advance here is not the notion that learning can spread, but rather that we have developed a way to directly measure and track that spreading (through text embedding space).

- *In Figure 3, and Supp Figure 2, the negative correlation between Lecture A and questions from Lecture B might in part be an "artifact" of the inclusion of only two lectures in the model. The more lecture are included, leading to more diverse topics, the more this pattern might decrease? (It is still expected that questions about a lecture will correlate with that lecture more than with other lectures, but the difference might not be as strong, especially for lectures on related topics).*

In the extreme, if we trained a model on a much larger and more general corpus (e.g., as in the BERT example shown in Supp. Fig. 6), we might expect the negative correlations to disappear. If trained on *too* broad a corpus, even the increased match between a question and the lecture it is about becomes indistinguishable from the match between the question and other related content (again, as in BERT). As we discuss on pages 24–25, there seems to be an interesting tradeoff between fine-tuning (specificity) and general tuning (generality). In our paper we do not attempt to characterize exactly where the boundary between specificity versus generalizability lies, or how training on corpora of successively broader "scopes" might affect the correlations between embedding vectors. However, it would be interesting to explore this in future work!