Reviewer #1 (Remarks to the Author):

Fitzpatrick and colleagues present a neat method for quantifying the content of real-world educational lecture videos and show that the resulting embedding space can be used to better understand and visualize quiz results assessing what knowledge students have learned. This manuscript is well-written and the figures are well-designed; I'm enthusiastic about the approach and the results are compelling. I have a couple clarifying questions about the methods, but my main concerns revolve around making sure the interpretations match the scope of the present data.

We thank the reviewer for their positive assessment!

1. I have two related high-level comments, both of which I'm having hard time articulating. First, I think the authors should be careful about the interpretative language they use throughout. Sometimes the language feels a bit lofty or overreaching. Let me try to provide an example: I'm not sure I fully buy the rhetorical parallel between "memorization" and "understanding." I don't think anyone in educational contexts really thinks of rote memorization as memorizing character strings the way a language model might. On the other hand, I don't really see how the topic model used in the present study (or even much more sophisticated language models) are doing anything above and beyond memorizing statistical (i.e. co-occurrence) structure (albeit at multiple scales). Sure, the model is not a simple, discrete lookup table, but I'm not sure we want to make any strong claim that the model captures "understanding" of physics either. If I understand correctly, the main advance from "memorization" toward "understanding" is made by embedding words in a continuous vector space, granting Shepard- (1987) style(?) generalization (i.e. the authors commitment that knowledge is fundamentally "smooth"). But the analyses simply demonstrate that the embeddings for a test set of words (i.e. quiz questions) can be interpolated from the embedding space learned from a training set of words (i.e. the video transcripts). Is this kind of interpolation fundamentally different from "memorization"? (A more radical claim here might be that much of what is seemingly "understanding" can be reduced to this kind of interpolation—which I generally agree with, cf. our Hasson et al., 2020 paper—but I doubt the authors really want to make that claim either.)

This is an important point, and one also raised by the other reviewers: to what extent are we capturing "understanding" as opposed to "memorization" with our approach? As an aside, this distinction is actually not the main claim we are trying to make in our paper—our comment about capturing understanding versus memorization was intended as an example of how these models might be *useful* (to educators, and so on). Nonetheless, there are a number of important details to unpack here.

First, we agree that the reviewer's example of memorizing character strings (like a language model might) is quite different from what we generally think of as "rote memorization" in educational settings. For example, memorizing a textbook to the point that a student could "regurgitate" the information on an exam still requires some basic *semantic* understanding of the material, as opposed to doing anything like learning co-occurrence statistics of the textbook's words.

As we clarify on page 3, we are not trying to claim that the topic models (or even other more sophisticated language models we could have used instead, like ChatGPT) are themselves "understanding" the material in any colloquial sense of the word. (We view those models more along the lines of the reviewer's description—essentially learning co-occurrence statistics of the text.) Rather, our approach is to use the *embedding spaces* that these models define as a way of constructing *maps* of what people might know (or even think about). Specifically, the embedding spaces these models learn have the property that nearby coordinates (e.g., in Euclidean distance) tend to reflect related concepts. In that sense, the spaces these models learn are "smooth" (differentiable everywhere) representations that we can use to efficiently describe (and infer) what *people* know or understand. The property that ends up being most useful is smoothness. Because the embedding spaces (learned by text embedding models) have the property that nearby coordinates reflect similar concepts, we reason that *knowledge* or *understanding* about the concept reflected by a given coordinate might imply knowledge or understanding about related concepts—which will tend to be at nearby coordinates in the embedding space.

We view the question of whether the embedding space captured by our (or other) text embedding model(s) is a "useful" representation of someone's knowledge as an empirical one. In other words, these maps are "useful" to the extent that they capture something predictive or informative about what people actually know. We show that we can predict individual participants' answers to held-out questions (Fig. 6), so this tells us that something about what the model "considers" to be conceptually related is also predictive of how participants' knowledge "spreads" across different concepts. We characterize this spread more directly in our revised manuscript (Fig. 7) by showing the distance (in topic space) over which the knowledge representations we derive using our approach provide predictive power over and above simply computing the proportions of correctly answered questions.

The deeper question of whether the knowledge maps we construct from participants' quiz performance capture true understanding versus something closer to memorization is trickier to get at. We think this comes down to the specific *questions* that we asked participants, as opposed to something about the embedding spaces. If the questions themselves require deep understanding (e.g., answering them correctly requires something beyond what can be achieved through memorization alone), then the maps we learn from participants' responses will also reflect deep understanding. If the questions themselves can be answered by memorization, then our maps will also reflect "memorized" concepts as opposed to true "understanding."

The questions we asked participants in our experiment vary with respect to how much we think they might capture memorization versus deeper understanding. For example, we suspect that questions like "*In the famous equation attributed to Albert Einstein, E = mc^2, what does the letter 'm' represent?*" could be answered by something close to pure memorization. But other questions, like "*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. Which of the following explains why?*" might require deeper understanding. The

"amount of understanding" a given question "requires" to answer is somewhat subjective, and might even come down to the "strategy" a given participant is using to answer that question at that particular moment. We have added a note to this effect on page 31 of our revised manuscript.

From a methodological standpoint, what we see as the major contributions of our approach are (1) a systematic means of mapping out and characterizing knowledge in maps that have infinite (arbitrary) numbers of coordinates, and (2) ways of filling those maps using relatively small numbers of multiple choice quiz questions. Our experimental finding that we can use these maps to predict responses to held-out questions has several psychological implications as well. For example, concepts that are assigned to nearby coordinates by the text embedding model also appear to be "known to a similar extent" (as reflected by participants' responses to held-out questions; Fig. 6). This suggests that participants also conceptualize similarly the content reflected by nearby embedding coordinates. The "spatial smoothness" of participants' knowledge (as estimated using quiz performance) in text embedding space is being captured by the knowledge maps we are inferring from their quiz responses (Fig. 7). In other words, our study shows that knowledge about a given concept implies knowledge about related concepts, and we also show how estimated knowledge falls off with distance in text embedding space.

To address these issues, we have added clarifying text throughout our revised manuscript. In addition, on page 8 we clarify how we interpret and use the text embeddings from the topic models, and on page 23 we clarify what we see as the major contributions of our approach.

*2a. Second, I think the authors should be very clear about the limits/scope of their approach, particularly based on the dataset used herein. For instance, in the Introduction, you use aspirational language about "the sum total of a person's knowledge" and "the nearly infinite set of possible things a person could know"—but the actual data/model used here is relatively small-scale (and there's no explicit evidence for scalability). I can't imagine that the co-occurrence statistics across only <20 minutes of video lectures are particularly dense or rich. Just to confirm, the model is not pretrained in any way and does not incorporate any conceptual information from outside the lecture videos, right? For example, the authors say (line 492) "Conceptually, each topic is intended to give larger weights to words that are semantically related *or* tend to co-occur in the same documents" (my emphasis on "or"). Does the model know that words are semantically related above and beyond the co-occurrence statistics? My understanding was that "semantically related" is derived entirely from the co-occurrences statistics. In another example, you say "Knowledge estimates need not be limited to the content of the lectures." But the embedding space is strictly limited to the co-occurrence structure extracted from these particular lectures—right?*

We have modified our text throughout our paper (including the statements the reviewer mentioned in the introduction) to clarify what we see as the scope of our work. As the reviewer notes, our current study does not show that we can map out *everything* someone knows; rather, we are focused only on a small "region" of what people know. To be clear, there *are* infinitely many coordinates in that small region of space (i..e, the space is continuous and differentiable everywhere within the convex hull formed by the union of all lecture timepoints' and quiz questions' embeddings). But of course the

space of concepts we focus on in our work does not (nearly) span everything someone could know, learn, or think about since, as the reviewer notes, we train the model using a relatively small number of documents that are (presumably) about a relatively small set of "discretizable" learning objectives.

We confirm that the embedding models are not pre-trained on data outside of the lecture transcripts and quiz questions. At a high level, the reviewer is correct that our topic modeling approach (Latent Dirichlet Allocation, or LDA) is driven by co-occurrence statistics. However, LDA does impose some additional structure on the embeddings beyond *pure* co-occurrence statistics (unlike, for example, Latent Semantic Analysis, or LSA, which is essentially like carrying out PCA on the dataset's word counts matrix). LDA attempts to learn the themes or "topics" that *underlie* the observed text, and each document (observation) is cast as a weighted blend of topics (which are in turn defined as distributions of weights over words in the vocabulary). The key difference between LDA and "pure" co-occurrence based approaches is that LDA can do things like teasing apart different uses of the same word. For example, suppose a word like "bat" appears in a document. In the absence of other information, we could guess that the word might refer to a winged mammal, a piece of sporting equipment, or something one can do with eyelashes. Models like LSA would infer that themes or "factors" related to each of these themes were present in the given document. LDA, by contrast, uses other words in the document to parcel out which specific use of "bat" most likely accounts for our observation.

This ability of LDA to identify latent themes enables the model to generalize (better than models like LSA) to *new* texts, as long as those new texts include a set of words that overlap with the trained model's vocabulary. In our implementation, although individual quiz questions might use different words to describe a given concept from a lecture, our model still "matches" questions with reasonable-seeming segments of video, when we compare the model's embeddings of the videos and questions (Fig. 4). In response to one of the reviewer's related comments (2c; see below), we have also run a new analysis showing that these LDA embeddings exhibit more temporally specific matches to the videos than obtained through simpler word count overlap.

Finally, the reviewer points out an ambiguity in our wording regarding semantic relatedness versus co-occurrence. We have clarified that sentence to read (emphasis added): "Conceptually, each topic is intended to give larger weights to words that are semantically related (*as inferred from their tendency to co-occur in the same document*)."

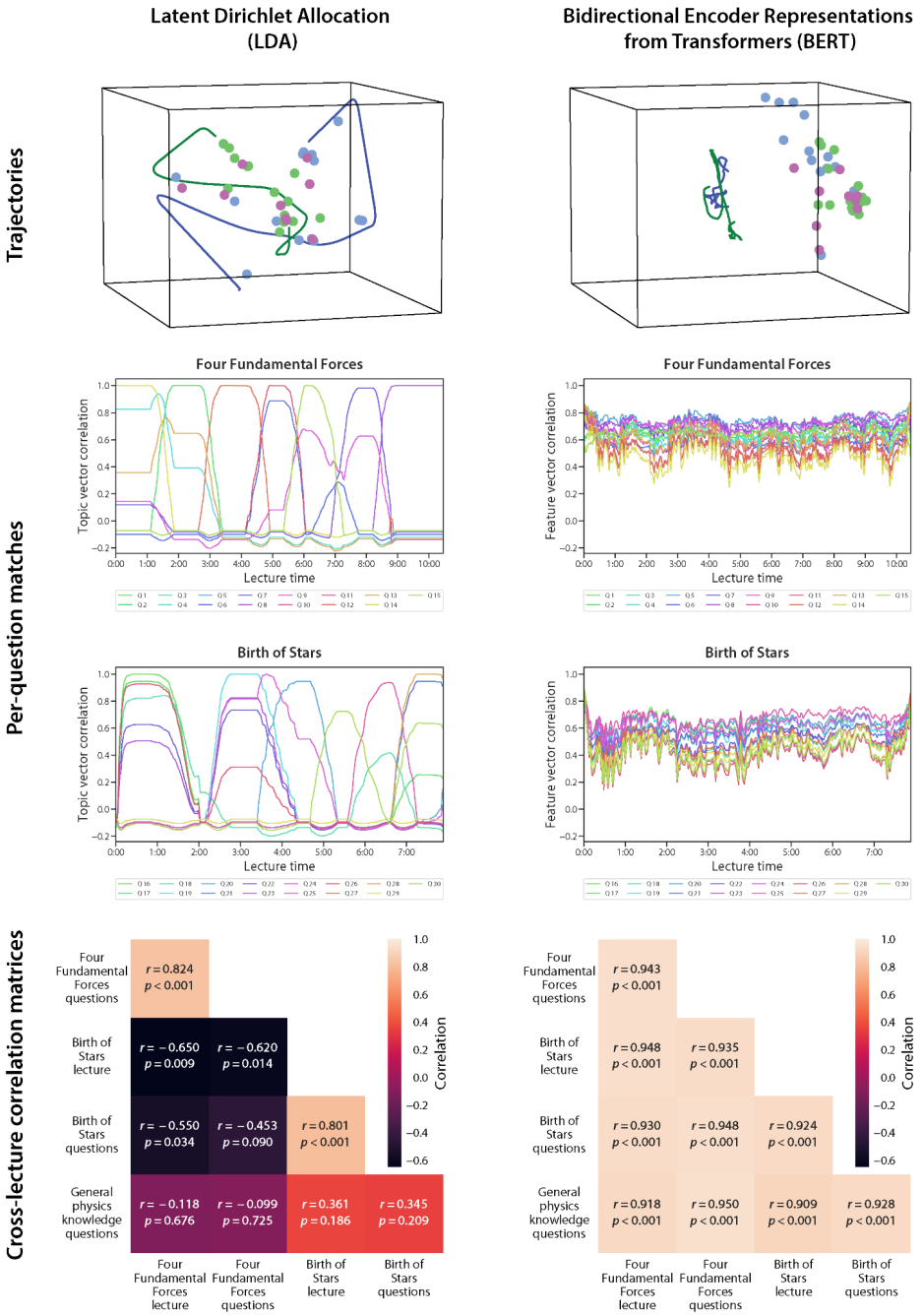
2b. Put another way, there's a tension in the writing between the potential generality of the method and the targeted nature of the embedding space (i.e. tuned to the content of the two lectures) used herein. Here's a specific example: In my understanding, the negative correlations between mismatching lecture and video topic weights (line 176) are an indicator that the embedding space reflects an "either this or that" structure specific to the two lectures. I have the intuition that near-zero correlation here would provide stronger evidence for "specificity" (line 179) than anticorrelation (as we see with the general physics questions). I wonder if this trivializes the dimensionality reduction results in Figure 7—i.e. could it be any other way? (Maybe this is fine!)

In general, we agree that there is some degree of "either this or that'' structure to the embeddings. For example, when we look at the embeddings of the lectures and questions (Figs. 3, S2), several topic dimensions appear to be "specific" to *Four Fundamental Forces* (topics 2, 3, 8, 12, 13, and 15), whereas others appear to be specific to *Birth of Stars* (topics 1, 4, 5, 6, 7). There are also some topics that appear in the average embeddings of both lectures and their associated questions (e.g., topics 11 and 14). This also explains the "negative correlations" the reviewer is referring to. From our prior work, and that of other groups (e.g., Boyd-Graber et al.'s 2014 chapter on "Care and feeding of topic models"), this follows from a fundamental tradeoff in text embedding models, between generalizability and specificity.

To illustrate, let's consider two approaches to fitting a given model. In both cases, we'll imagine using the same model architecture, the same number of parameters, the same data set sizes, and the same inference procedure. The difference is that in hypothetical approach A, we'll choose our training corpus to span a wide variety of content from a diverse set of sources. By contrast, in hypothetical approach B, we'll choose our training corpus to contain documents that cover a relatively narrow scope. At a high level, since both models had the same number of observations and both models are of the same "complexity" (i.e., matching architectures, numbers of parameters, goodness of fit to their respective training datasets, and so on), these could be considered "equivalent" models. Another way to think about these models, though, is about the tradeoffs each makes between *generalizability* (i.e., something like the chances the model would be able to generate a "reasonable" embedding of some randomly selected held-out text) and *specificity* (i.e., something akin to the degree to which the embeddings are able to reliably differentiate between subtly different texts or concepts). Essentially, a model's design, parameters (e.g., number of embedding dimensions), and the amount and quality of training data collectively determine the effective "resolution" of the model– i.e., related to the fidelity of the overall embedding space. But within those constraints, the choice of training corpus determines whether the embedding space spans a wide range of content at a relatively lower resolution (as in the "approach A" example above), or whether the embedding space "zooms in" on a narrower range of content at a comparatively higher resolution (as in the "approach B" example above). In other words, the particular set of documents we use to train an embedding model ends up determining the breadth and depth of the embedding space.

One way this can play out is in the sort of "either this or that" representations the reviewer is noticing in our model, as well as the tension they're identifying between "generality of the method versus the targeted nature of the embedding space." Since we trained our text embedding model on documents that were primarily "about" two similar but partially separable sets of concepts, the subtle conceptual variations within a video are each "assigned" several topic dimensions that enable the model to represent those variations within each individual video. That's the aspect of the model's features that looks like "switching" or like an "either this or that" representation. If we had used a greater diversity of documents in our training corpus, the within-video resolution would have suffered, and that "switching" behavior would have been more subtle.

More broadly, there are real challenges to building embedding models that are *both* sufficiently high-resolution within a given topic *and* sufficiently broad so as to cover a wide range of topics. For example, embeddings derived from even large modern models like BERT, GPT-{2,3,4}, LLaMa, and others, that are also trained on enormous text corpora, end up yielding surprisingly poor resolution within the content space spanned by a single course video. For example, here are some comparisons between basic visualizations from our paper derived from our topic modeling approach (LDA) versus an example large language model (BERT):



We highlight three general differences in the above figure. In the top row ("Trajectories") we show a 3D projection (using PCA) of the content trajectories for each lecture (blue and green lines) and each question (blue, green, and purple dots). The LDA embeddings exhibit several desirable properties compared with the BERT embeddings:

- **The LDA embeddings of the lectures and questions are "near" each other– e.g., the convex hull enclosing the two lectures' trajectories is highly overlapping with the convex hull enclosing the questions' embeddings. The BERT embeddings of the lectures and questions are instead largely distinct.**
- **The LDA embeddings of the questions for each lecture and the corresponding lecture's trajectory are similar. E.g., the LDA embeddings for lecture 1 questions (blue dots) tend to appear closer to the lecture 1 trajectory (blue line), and the LDA embeddings for lecture 2 questions (green dots) tend to appear closer to the lecture 2 trajectory (green line). The BERT embeddings do not show this property.**

In the middle row ("Per-question matches"), we display the correlations between each question's embedding coordinate and the embedding of each moment of the corresponding lecture. Again, the LDA embeddings show several desirable properties compared with the BERT embeddings:

- **The time series plot of individual questions' correlations are different from each other when computed using LDA (e.g., the traces can be clearly visually separated), whereas the correlations computed from BERT embeddings of different questions all look very similar. This tells us that LDA is capturing some differences in content between the questions, whereas BERT is not.**
- **The time series plots of individual questions' correlations have clear "peaks" when computed using LDA, but not when computed using BERT. This tells us that LDA is capturing a "match" between the content of each question and a relatively well-defined time window of the corresponding lectures. The BERT embeddings appear to blur together the content of the questions versus specific moments of each lecture.**

In the bottom row ("Cross-lecture correlation matrices") we compare the pairwise correlations between embeddings of lectures and questions within versus across content areas (i.e., content covered by the individual lectures, lecture-specific questions, and by the "general physics knowledge" questions). The LDA embeddings show a strong contrast between same-content embeddings versus across-content embeddings. In other words, the embeddings of questions about the lecture 1 ("Four fundamental forces") material are highly correlated with the embeddings of lecture 1, but *not* with the embeddings of lecture 2, questions about lecture 2, or general physics knowledge questions. We see a similar pattern with the LDA embeddings of the lecture 2 ("Birth of stars") questions. In contrast, the BERT embeddings are *all* highly correlated with each other.

Taken together, these comparisons illustrate how LDA (trained on the specific content in question) provides both coverage of the requisite material and specificity at the level of the content covered by individual questions. BERT, on the other hand, essentially assigns both lectures and all of the questions (which are all broadly about "physics") into a tiny region of its embedding space, thereby blurring out meaningful distinctions between different *specific* concepts covered by the lectures and questions.

We note that these are not criticisms of BERT (or other large language models trained on large and diverse corpora). Rather, our point is that simple fine-tuned models trained on a relatively small but specialized corpus can outperform much more complicated models trained on much larger corpora, when we are specifically interested in capturing subtle conceptual differences at the level of a single course lecture or question. Of course if our goal had been to find a model that *generalized* to simultaneously capturing many different content areas, we would expect our approach to perform comparatively poorly relative to BERT or other much larger models. We suggest that bridging the tradeoff between high resolution within each content area versus the ability to generalize across many different content areas using a single set of model weights will be an important challenge for future work in this domain.

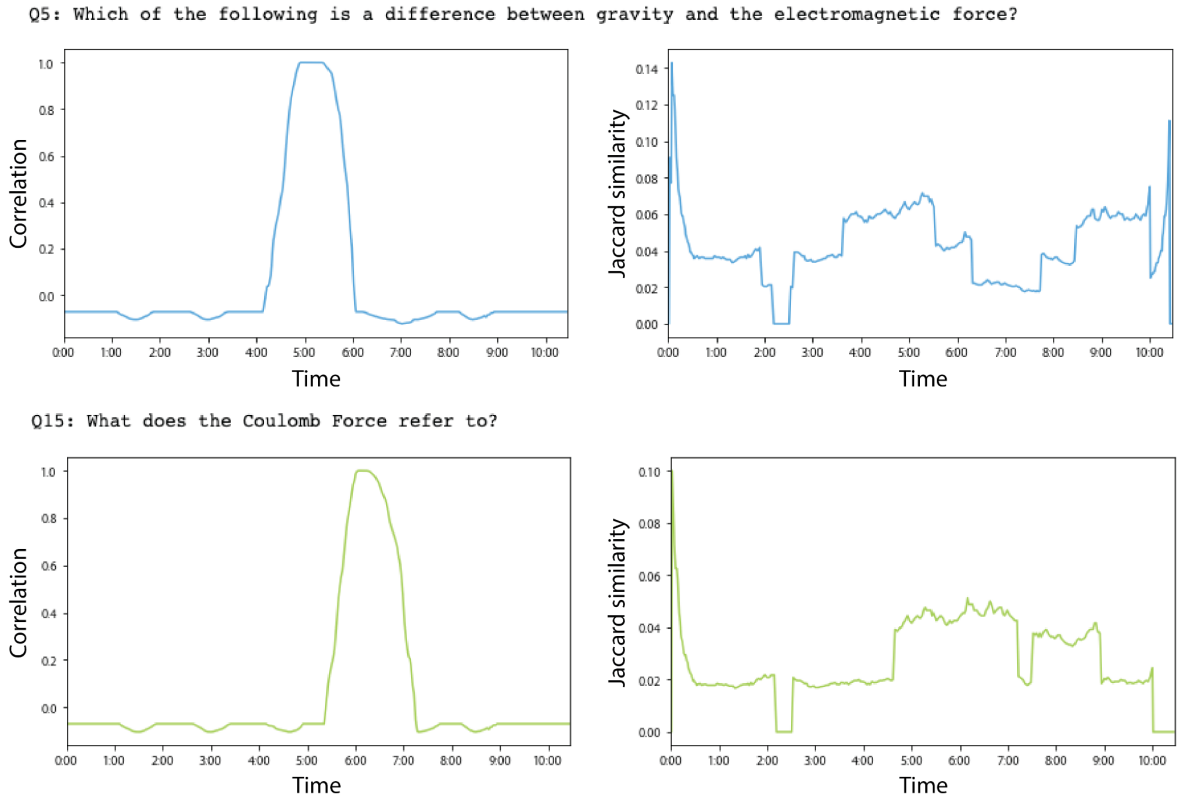
We have added a discussion of these issues to pages 23–25 of our revised manuscript, and have also added the figure above to our supplementary materials (Supp. Fig. 6).

2c. Another specific example: The authors want to make a strong claim that the model doesn't rely on surface-level similarity and that the "specific words" used in the quiz questions have "very little overlap" with the lectures? But there must be some overlap (i.e. co-occurrence), right? In my understanding, a completely disjoint set of words, none of which occurred in these video transcripts, would not map onto this embedding space in any meaningful way. To be clear, I don't think this is really a problem or a weakness of the method!—I'm just trying to articulate some discomfort when reading.

The reviewer is correct that there often is *some* overlap between the specific words used in the quiz questions and the corresponding parts of the lecture that discuss the "matching" concepts. What we meant (and have now attempted to clarify throughout our revised manuscript) is that the topic models provide additional characterization of the conceptual content of the lectures and questions *over and above* what we see using simple word matching alone.

To compare topic-based matching (correlation between topic vectors) with word-based matching (e.g., overlap between the words in a question vs. in a sliding window of a lecture's transcript), we implemented a simple measure of word overlap using Jaccard similarity. (The Jaccard similarity between text A and text B is computed as the number of unique words in the intersection of words from A and B, divided by the number of unique words in the union of words from A and B.) To carry out a fair comparison with our topic modeling approach, we ran the text through the same preprocessing steps we used prior to fitting the topic models, including removing stop words and

punctuation, converting digits to word-form, and so on. We can then compare the time series of correlation-based similarities (using text embeddings) versus the time series of word-based (Jaccard) similarities for each question and lecture. In the figure below, we show the time series plots for two example questions (rows; question text is displayed above the corresponding panels), each computed using either topic correlations (left panels) or Jaccard similarities (right panels):



The question in the top row (Question 5 about Lecture 1) is about comparing gravity and the electromagnetic force. When we use topic correlations to match up the question's embedding with Lecture 1 timepoints (left column), we can see a clearly defined "peak" where the embeddings show strong correlations. When we examine the Lecture 1 transcript during that interval, we can see that the instructor is describing how those two forces relate (emphasis added):

> **"...disappears as an actual force as an actual interaction now the next the next force up the hierarchy which is one that we are more familiar with it is something it's what actually dominates most of the chemistry that we deal with and electromagnetism that we deal with and *that's the electro magnetic force we write it in magenta electro magnetic magnetic force and just to give a sense this is this is 10 to the 36 times the strength of gravity 10 to 36 times the strength of gravity* so it kind of puts the weak force in its place it's 10 to the 12 times stronger than the weak force so these are huge numbers that we're talking about either this relative that or even this relative to gravity and so you might be saying well you know *the electromagnetic force that's unbelievably strong why doesn't that apply over over over these these kind of macro***

scales like gravity **let me write there macro scales macro scales why doesn't it apply to macro scales and actually there's nothing about the electromagnetic force why it can't it or it actually does apply over large distances the reality though is** *you don't have these huge concentrations of either electric coulomb charges or magnetism the way you do mass so the mass that you have such huge concentrations it can operate over huge huge distances even though it's way way way weaker than the electromagnetic force the electromagnetic force what happens is because it's both attractive and repulsive* **it tends to kind of sort itself out so you don't have these huge huge huge concentrations of charge now the other thing you might be wondering about is you know why is it called the electromagnetic force in our everyday life there's things like there's things like the Coulomb force that or the electrostatic force which we're familiar with positive charges or like charges want to repel if both of these were negative the same thing would be…"**

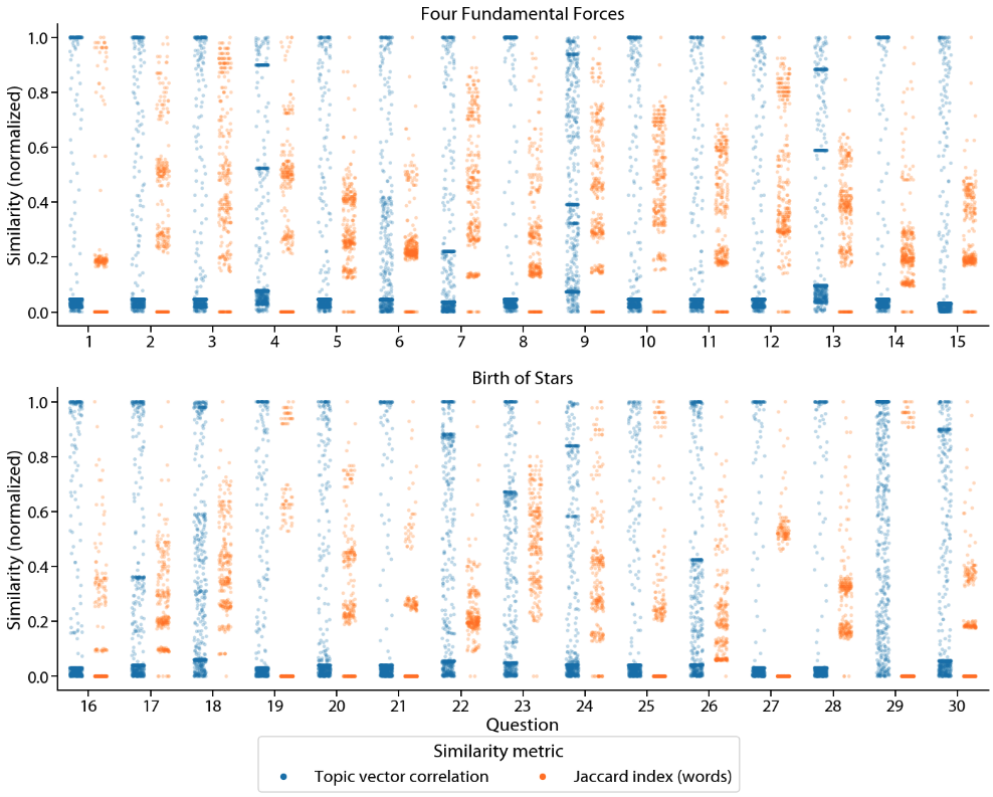
In contrast, simple word matching (right column) identifies moments when any of the *words* **in the question are used during the lecture (but not necessarily all together)-- e.g., mentions of "following," "difference," "gravity," "electromagnetic," and "force" at other moments in the lecture also evoke increases in Jaccard similarity.**

Similarly, the question in the bottom row (Question 15 about Lecture 1) is about the Coulomb Force. Again, our topic correlations approach (left column) yields a well-defined match interval that corresponds to an interval in the lecture when the instructor is describing the Coulomb Force (emphasis added):

"...electromagnetic force in our everyday life there's things like there's things like *the Coulomb force that or the electrostatic force which we're familiar with positive charges or like charges want to repel if both of these were negative the same thing would be happening and different charges like to attract we've seen this multiple times this is the Coulomb force or the electrostatic force electro static* **and then on the other side of the word yes you have the magnetic part and magnets you know you have you've played with magnets on your fridge they what's what you know if they're the same side of the magnet they're going to repel each other if they…"**

On the other hand, simple word matching (right column) shows non-specific matches throughout the lecture to moments when the transcript includes "coulomb," "force," or "refer," but not necessarily all together (e.g., other uses of "force" will evoke an increase in Jaccard similarity, whether or not the lecture at those moments is specifically referring to the *Coulomb* Force).

We see analogous patterns for each question. Although the topic correlations are not always perfect (for example, our topic model appears to occasionally confuse forces that are strong with the Strong Force), using an explicit model of topics that uses co-occurrence patterns in the text to identify latent themes provides our analyses with structure that enables us to match up questions and lecture timepoints well in practice:



In the above figure, each dot denotes a single comparison between a lecture timepoint (top: Four Fundamental Forces; bottom: Birth of Stars) and a single question. The blue dots denote the correlations between topic vectors and the orange dots denote Jaccard similarities. In each case, the topic correlations exhibit a clear "peak" correlation at just a few well-defined moments in the lectures, and most of the remaining timepoints have correlations near 0. The Jaccard similarities, however, often exhibit much broader ranges of values, reflecting that Jaccard similarities yield matches that are less temporally specific (i.e., more diffuse) than those computed using topic correlations.

We have added a brief discussion of the benefits of using a topic model (versus simple word matching) on pages 25–26, along with three new supplemental figures (combining elements of the figures pasted into our response above; Supp. Figs. 3, 4, and 5) and a new supplemental table (Supp. Tab. 3) showing lecture text from the best-matching segments of the lectures, for each question.

3. How exactly were the questions constructed for each lecture? For example, Figure 2 suggests that the questions follow the trajectories of the lectures, at least to some extent; obviously this is sensible and "by design." It's great that the authors provide the full set of questions/answers (Table S1)! However, I think it's important that the authors describe their thought process in designing the questions a bit more explicitly. For example, how exactly did they minimize "surface-level" similarity while ensuring enough conceptual overlap that the questions map onto the embedding space in a meaningful way?

One of our lab's undergraduate research assistants worked alongside a Masters student who was rotating in the lab for a month to develop a list of questions (these researchers are acknowledged in the paper for their valuable contributions to the work, but they did not meet the criteria for authorship discussed with all team members at the start of the project, as determined by JRM). The senior author (JRM) tasked the pair of researchers with coming up with "15 conceptual questions about each lecture, along with 9 additional questions about general physics knowledge." To help broaden the set of lecture-specific questions, the researchers were further instructed to work through each lecture in small segments, identify what each segment was "about" conceptually, and then write a question about that concept. The general physics questions were drawn from the researchers' coursework along with internet searches and brainstorming with the project team and other members of JRM's lab. The final set of questions (and response options) was reviewed and approved by JRM before collecting or analyzing the text or experimental data.

We have added a note to this effect on pages 30–31 of our revised manuscript. We have also toned down our language describing our question pool as reflecting "surface-level" versus "conceptual" details where appropriate, as we acknowledge that we did not have an explicit process for optimizing or designing questions to be "conceptual" beyond our own intuitions.

4. The authors correlated the topic weights for each question with the topic weights at each time point in the lecture, and in Figure 4 show that "each question appeared to be temporally specific" (line 208). Intuitively, this seems reasonable… But couldn't this result also be interpreted as creating tension with the idea that the questions aren't simply keying into "surface-level" similarities? Put another way, if I thought this approach captured some kind of holistic "understanding", I might hypothesize that good questions would integrate multiple concepts introduced across different parts of the lecture. (I might be conflating the overall validity of the approach for capturing "understanding" with the quality of the questions here.)

The reviewer raises a number of interesting questions here about whether the questions we asked participants are specifically capturing deeper understanding as opposed to solely capturing surface-level details. As we note above, since we don't have an explicit definition of what specifically differentiates "conceptual" versus "non-conceptual" questions, to some extent we end up relying on our subjective judgements in these matters. Our sense is that, across the set of questions we asked participants, there is substantial variance in how "deeply" each question probes conceptual understanding versus something more akin to rote memorization. We also discussed this issue, along with some example questions that we see as varying in "depth," in our response to the reviewer's comment 1. In summary, we acknowledge that at least some of our questions likely do *not* capture what we mean by deep understanding, and for other questions it is difficult to objectively say how deep an understanding is required. To address this point, we have toned down how we describe our experiment's questions throughout our revised manuscript, and we have also added a note to this effect on page 31.

The above said, we *can* say *some* things about what our approach is versus isn't capturing. For example, consider an embedding space that fails to capture any deeper semantic meaning (e.g., as in a "random" embedding space for which Euclidean or correlation-based distance was fully uncoupled from semantic meaning). In such a space, there would be no predictive value of the embeddings. In other words, knowing that a participant answered a given question correctly would provide no additional information (on average) about the participant's knowledge about the content covered by questions that were nearby in the embedding space as compared to questions that were far away. On the other hand, if an embedding space *does* capture similarities in semantic meaning in a way that follows similarities in participants' knowledge, then participants should have similar "levels of knowledge" about questions whose embeddings were nearby in the space.

Of course, knowledge itself cannot be directly measured; it can only be inferred "indirectly" through behaviors or other observations. But to the extent that participants' knowledge affects their chances of answering questions correctly, the embedding spaces learned by our topic modeling approach, and our framework for estimating what we call "knowledge," seem to have some explanatory power (in terms of predicting responses to held-out questions).

We also lack any direct means of defining or determining the "depth" of someone's knowledge. We found that the embedding coordinates of a held-out question, along with participants' performance on other questions at nearby coordinates, reflect how the participant will perform on the held-out question (Fig. 6). We also found that explanatory power falls off smoothly with distance in the text embedding space (Fig. 7). This suggests that we are capturing something more like "actual meaning" (as reflected in semantic relationships between different questions), over and above exact text matches (arguably the most basic framing of "surface level details").

With respect to the reviewer's point about how concepts may integrate across different timepoints, we do not see temporal specificity as necessarily in tension with how "deep" or "complex" (or even "holistic") a particular concept is. We see this as depending on several factors including what the concept itself *is*, the particulars of the lecture in which the concept is explained or described, the instruction style, and so on. Presumably the temporal specificity also depends on the text embedding model's sensitivity to distinct aspects of the same overarching concept.

Another way of considering the reviewer's comment about temporal specificity versus depth may be captured in part by our response to their comment 2c, above. In our response to that comment, we report on an analysis specifically aimed at comparing the temporal specificity of question-lecture matches computed using topic models versus simpler measures of word overlap. We found that our topic modeling approach, which inherently captures deeper structure and co-occurrence patterns than word overlap alone, results in more temporally specificity. This tells us that, at least for those two example approaches, greater temporal specificity is consistent with a model that captures *more* semantic detail (which one could interpret as a "deeper" semantic representation, or as reflecting greater nuances in "similarity").

5. The authors used windows comprising 30 lines(?) of the transcript to define documents for input to the topic model. This seems like an important parameter as it impacts the co-occurrence structure observed by the topic model, right? How was this number 30 chosen? I assume it represents some happy medium between providing a wide enough window to yield rich co-occurrence structure while retaining some temporal specificity. Transcript "lines" seems like a somewhat arbitrary unit; I assume it relates to how YouTube renders the transcript? How many words tend to occur in each of these windows? What's the average duration of these windows in seconds?

The reviewer is raising a number of important questions about how we chose the "parameters" of our general approach. In our prior work (Heusser et al., 2021, Nature Human Behaviour), where we developed a similar "topic modeling of sliding text windows" approach, we were computing text embeddings of human-generated annotations. We reasoned in that work that each sentence of the annotations could be treated roughly as a "unit of thought" for the purposes of our analyses. Essentially choosing the level of resolution of the analysis (letters, words, sentences, paragraphs, entire documents, etc.) determines the "resolution" at which differences between successive observations can be detected or characterized. Sentences seemed like a reasonable choice in that prior work.

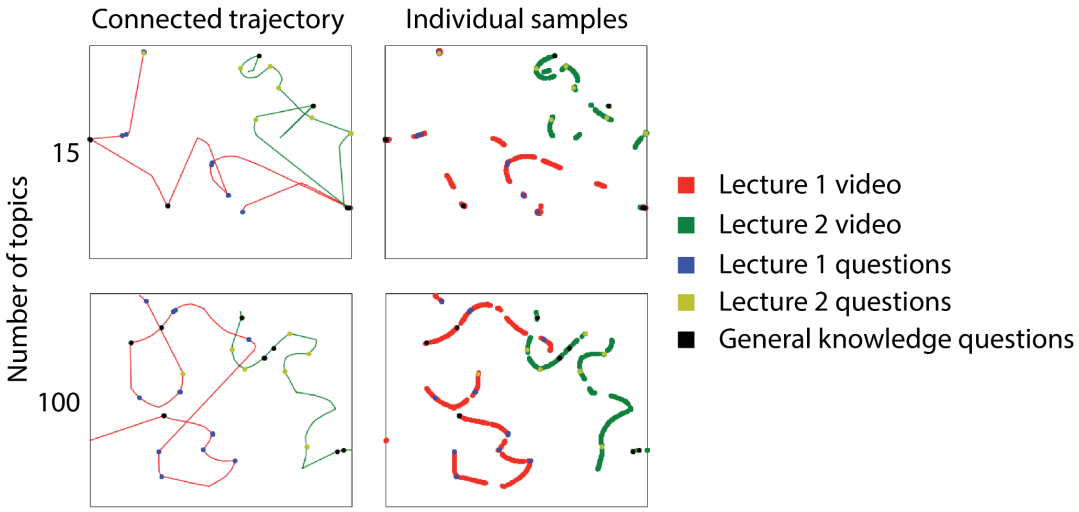
In our present study, we tried to follow our prior approach as closely as we could, but due to differences in the text and requirements of our approach, some adaptations were needed. For example, in our present work we used automatically generated (via speech to text algorithms) transcripts that did not include punctuation. Further, Khan Academy lecture videos are often narrated in a sort of "stream of consciousness" style that often doesn't map cleanly onto sentences per se. Instead, we treated each "line" of the automatically generated transcript as a stand in for a "sentence." To give a sense, each line of the automatically generated lecture transcripts corresponds to a single block of closed captioning text that is displayed on-screen during a given part of the lecture. We selected a window length of 30 "lines" in our current paper to match the same average number of words per sliding window (before preprocessing, rounded to the nearest word) as in the sentence-based transcripts we used in our prior work (Heusser et al., 2021). The sliding windows had an average of 73.8 words and lasted for an average of 62.22 seconds.

We have added some additional explanation of how we chose the sliding window lengths in our study, along with the average word count and duration per window, to page 33 of our revised manuscript.

6. I don't fully understand the motivation for choosing 15 topics for most of the analyses, or why you increase to 100 topics for Figure 7. I don't think there's anything wrong with picking a somewhat arbitrary round number in these cases; just curious if there's any particular reasoning behind the present choices.

This choice was motivated by some observations about how UMAP works and some choices that affect how accurately the embeddings of the data in the low dimensional space reflect the original

high-dimensional data. To give some context, the UMAP algorithm is what we use to visualize the 2D "maps" by converting correlations in the original (high-dimensional) topic space onto Euclidean distances in the 2D map space. This property makes it possible to intuitively visualize the similarity relations (correlations between topic vectors) we think are most reflective of actual semantic similarities. However, in "solving" for the mapping between the high and low dimensional spaces, UMAP ends up distorting the data in a few different ways. Some of these distortions, such as information loss when we move from high-dimensional to low-dimensional representations, are unavoidable. But others appear to depend in part on properties of the original high-dimensional data, including the number of dimensions, number of observations, etc., and on the number of dimensions in the target low-dimensional space. We noted two types of distortions in particular. First, UMAP has a tendency to "snap" nearby points in the original high-dimensional feature space onto the same (or very nearby) coordinates in the low-dimensional projection. Second, because the low-dimensional embedding is computed using clusters of nearby observations in the original (high-dimensional) space, observations in the low-dimensional space can be "clumped" so that they appear closer together in the low-dimensional projection than they are in the original space. We can see how these two types of distortions appear in the 2D projections of 15 versus 100 dimensional text embeddings of the same lectures and questions:



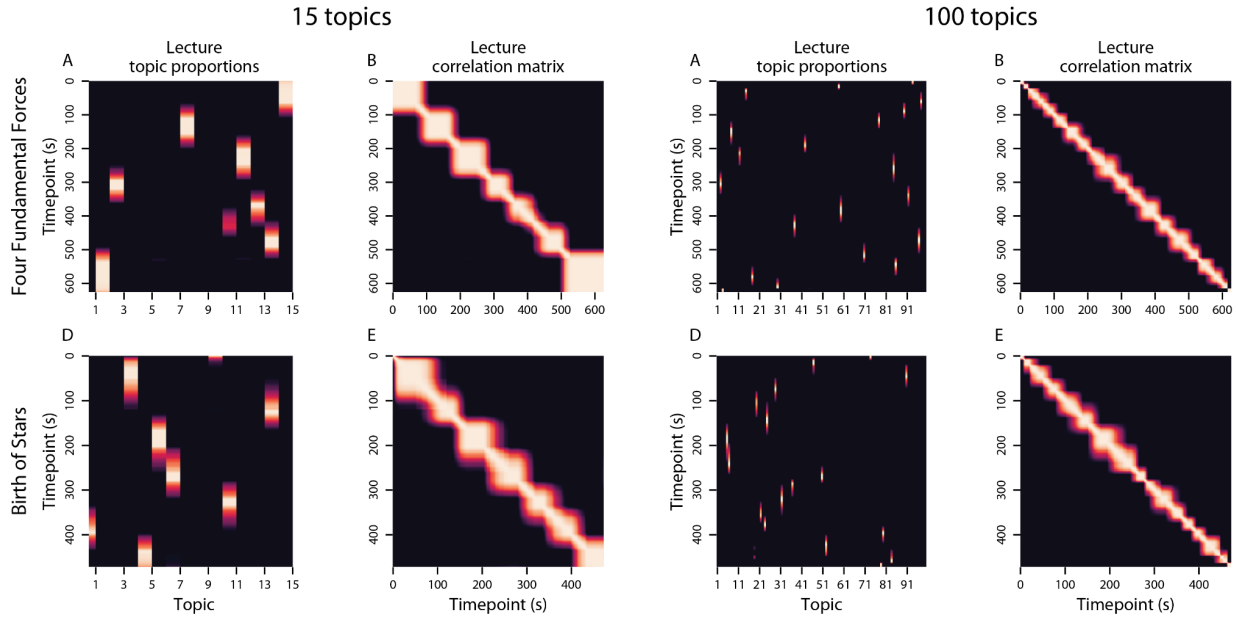
When we "connect the dots" between successive observations in the lectures (left column), both sets of trajectories look "reasonable" (e.g., they're relatively smooth, the two trajectories are partially overlapping but mostly distinct, the questions "about" each lecture correctly map onto the appropriate parts of the 2D space, and so on). But when we examine the embeddings of the individual timepoints of the lectures (right column), it's clear that there are more "gaps" in the 15-topic version. Since we know that both trajectories are smooth in the original high-dimensional feature space (because we computed them using overlapping sliding windows and resampled each trajectory to have a constant sampling rate), it's clear that the 100-topic versions more accurately reflect how the original high-dimensional data behaves. We thought fewer gaps would also be desirable when visualizing the 2D knowledge maps, which require interpolating from nearby observations (so gaps in the data

require more interpolation and less reliance on actual observations). To select an appropriate number of topics (*k*) for our model, as a starting point, we identified the minimum number of topics that yielded at least one "unused" topic (i.e., in which all words in the vocabulary were assigned uniform weights) after training. This suggested that the number of topics might be sufficient to capture the set of latent themes present in the two lectures. We found this value to be $k = 15$ topics. We found that with a limited number of additional adjustments following Boyd-Graber et al. (2014; e.g., removing corpus-specific stop-words), the model yielded (subjectively) sensible and coherent topics.

Although the above reasoning suggests that the 100-topic embeddings are better suited for creating the 2D knowledge map visualizations, we also identified several important differences in the topic activation patterns when we used 15 vs. 100 topics:

- **In the 15-topic model, topic activations persist for longer durations than in the 100-topic model. We thought that these longer durations might boost the sensitivity of our approach by integrating over longer spans of time (e.g., under the assumption that the estimated topic weights are "noisy").**
- **In the 15-topic model, topic activations often overlap (e.g., in a given timepoint, several topics often have large weights), whereas this happens only rarely in the 100-topic model. This property of the 15-topic model seemed potentially useful for capturing interesting co-variation across topics.**
- **The 15-topic model (roughly) "segments" the lectures into on the order of 5-10 chunks (roughly one per minute). When we examine the temporal correlation matrix, we can see roughly 5-10 square "blocks" of strong correlations along the diagonal. The 100-topic model segments the lectures much more finely (approaching around 10 seconds per "block"). Intuitively, we wanted to capture broad concepts that we thought might take on the order of a minute to "explain," as opposed to much lower-level concepts that might be mentioned in a few seconds. This again led us to prefer the 15-topic model in our main analyses.**

These patterns may be seen in the figure below (left: 15-topic model; right: 100-topic model; first and third columns show topic activations over time; second and fourth columns show the temporal correlation matrices; top row corresponds to the *Four fundamental forces* lecture and the bottom row corresponds to the *Birth of stars* lecture):



We briefly discuss some of this logic on pages 18–19. We note that all of our *analyses* rely on the 15-topic model, and we carry out the relevant comparisons, distance computations, and statistical tests in the original 15-dimensional space for that model. We used the 100-dimensional model solely for visualization, when we created the knowledge maps (Fig. 8).

Minor comments:

Regarding my previous comments on scalability, do the authors think this kind of approach could be extended from topic models to the contextual embeddings learned by large language models? Could be worth mentioning if this is a valid future direction, given the current popularity of these models. My guess is that if we want a method that can capture "the sum total of a person's knowledge" and "the nearly infinite set of possible things a person could know," large language models are currently the best approach.

We share the reviewer's intuition that large language models may be the most promising way to scale up our approach to a broader range of concepts. As we noted in our response to the reviewer's comment 2b above, simply swapping in a large language model trained on a much larger and more diverse corpus does not seem to "just work." Rather, as we discuss above, we've found an overarching tradeoff: when we include a larger and more diverse set of content in our training corpus, or when we used substantially larger models (e.g., BERT), we found that the resulting models are more generalizable to broad ranges of content, but less able to pick up on low-level details (e.g., at the level of individual questions or small segments of individual lectures). We've been starting to think through how one might construct models that work well at a variety of conceptual scales. One idea we've had is to potentially combine multiple models into a single "hybrid multi-scale model."

Another is to re-think how the embeddings themselves are used internally by the models, or how we use those embeddings to construct maps. While we're excited about these possibilities, we also feel that they're beyond the scope of our present study.

In Figures 2 and 7, you use PCA to visualize two- and three-dimensional projections of the embedding space. What proportion of cumulative variance do these first two and three PCs account for?

In Figure 2C, we use PCA to visualize the 3D projections of the embedding space. Those first 3 components explain 36.2% of the variance in the original embeddings of the lecture videos and questions. In Figure 7 we use UMAP (not PCA) to visualize the embedding spaces as 2D maps. Unlike with PCA, there's no straightforward way of estimating variance explained by UMAP, since it is a form of nonlinear dimensionality reduction. (Leland McInnes, the lead author of the UMAP algorithm, explains this issue further here: https://github.com/lmcinnes/umap/issues/122).

All participants saw the Four Fundamental Forces video first, followed by the Birth of a Star video, right? Should we be worried about some order effect?

All participants in our study viewed the two videos in the same order. Therefore it's of course possible that there might be order effects that we are not accounting for (i.e., we did not counterbalance the viewing order). That said, we are not aware of any reason why viewing order should specifically influence or affect our main findings.

If the two lectures had a dependency relationship (e.g., understanding one lecture required first having learned the content of the other) then this might introduce more serious order effects. But we specifically selected the two lectures in our study to be self-contained (and therefore conceptually independent of each other). We mention this as one of the criteria for choosing these two particular videos on page 6.

Line 283: These t-values have 600+ degrees of freedom; I assume this is determined by the total number of questions across all participants. Should we be worried that this kind of statistical test doesn't take into account subject-level variability?

We generate the knowledge estimates for each held-out question using only the same participant's responses to other questions on the same quiz (Fig. 6). So there is no "leakage" across participants in our predictions (e.g., the fact that other participants tended to get a particular question wrong will not inform our predictions of how the *current* participant answered that question). Our statistical tests are carried out on the full distribution of knowledge estimates; specifically, we compare the estimated knowledge of held-out questions that were actually answered correctly vs. held-out questions that were actually answered incorrectly. Given that our overall approach is already relatively complicated, we opted for a simple and easy-to-interpret statistical test, rather than attempting to model participant-level effects or other factors (e.g., of question identity, content area, etc.). While an

imperfect solution, we made an incremental improvement to our statistical framework by using non-parametric Mann-Whitney *U*-tests instead of *t*-tests.

Line 492: Do you use a specific set of stop words? (e.g. the NLTK package has a standard list of stop words)

We use the NLTK package's "English" stop words list, augmented with the following additional list of words (identified following Boyd-Grayber et al., 2014): ['actual', 'actually', 'also', 'bit', 'could', 'e', 'even', 'first', 'follow', 'following', 'four', 'let', 'like', 'mc', 'really', 'saw', 'see', 'seen', 'thing', 'things', 'two']. We clarify this point on page 33.

Line 557: "insure" > "ensure"

Fixed!

Reviewer #2 (Remarks to the Author):

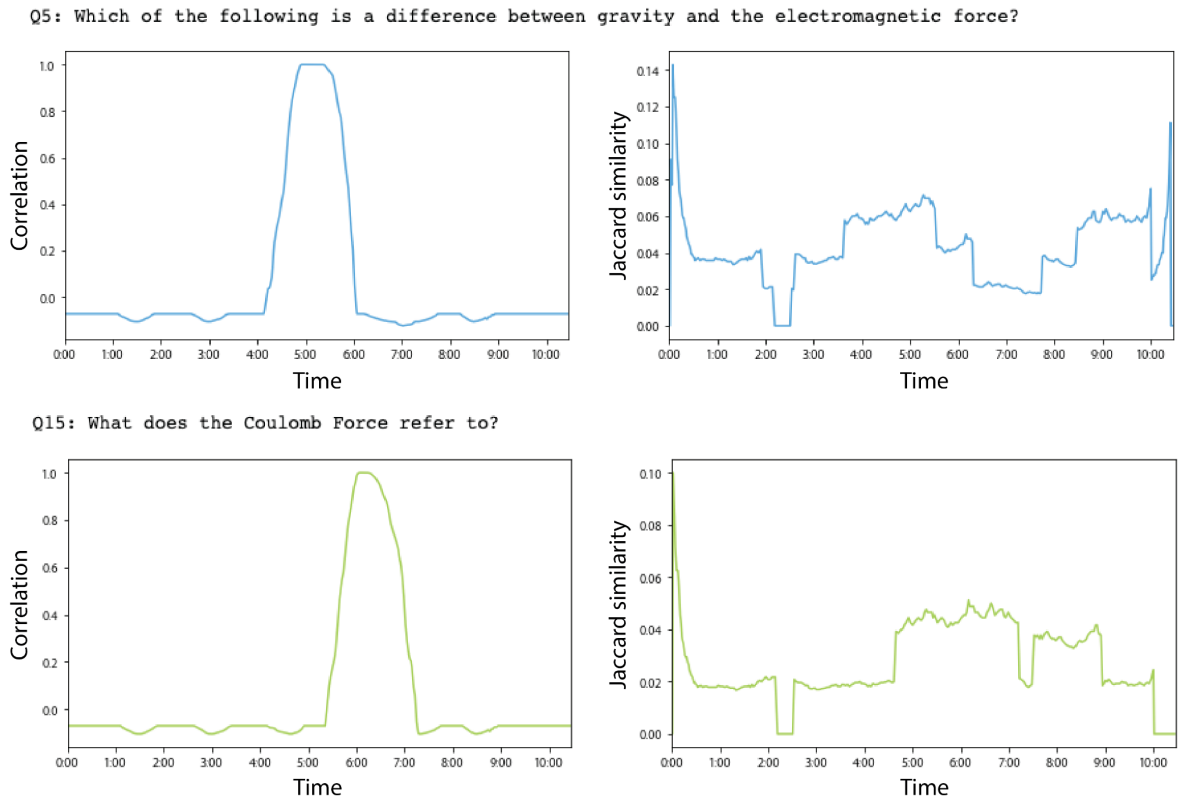
The authors develop a framework for characterizing the acquisition of knowledge across time, specifically in the context of educational lectures. Participants watched two physics-related lectures on the Khan Academy platform, and completed quizzes before the first lecture, in between lectures, and after the second lecture. Questions on all quizzes tapped into knowledge from the two distinct lectures (as well as general physics knowledge). The information contained in the lectures and associated quiz questions was embedded in a topic space derived from word co-occurrences contained therein. This approach enabled the authors to (1) define each of the two lectures as a trajectory through this embedding space, (2) quantitatively determine the differences in conceptual content between the two lectures (by assessing topic weight and variability), (3) match quiz questions to specific moments in the lecture with relatively high resolution, (4) determine how much knowledge participants had about different points in the lecture, (5) predict whether a participant got a question correct or incorrect based on their estimated knowledge at that relevant coordinate in the embedding space, and (5) visualize how knowledge spreads through the embedding space as a result of watching the lectures. I am convinced that this framework yields more informative feedback on students' quiz results, relative to calculating percent correct. I find the authors' approach to be an exciting one, and that it provides methods others can use to quantify and visualize the specific content of one's knowledge and how it might change over time. The methodology is sound, and the introduction and discussion are written well. This work primarily proposes a methodological framework rather than presenting new empirical results. Despite my overall enthusiasm, some other comments/suggestions are summarized below.

We appreciate the reviewer's summary and positive overall assessment.

1. It is mentioned multiple times that correspondences between quiz questions and lectures (e.g., localizing each quiz question to a specific point along a lecture's trajectory) are not superficial, because "the lectures and questions used different words." As far as I can see, there is no quantification of word overlap between lectures and quizzes, and we are supposed to take the authors' word on this. I find it difficult to believe that there is no word overlap between lectures and quiz questions, and would like to see some quantified measure of word overlap. The degree to which

correspondences in the embedding space are "superficial" or "deep" is unclear and as it stands, I am not fully convinced that the current analyses tap into deep conceptual knowledge.

Reviewer 1 raised a similar point about comparing word overlap in topic correlations in their comment 2c, above. In brief, to explore this we ran a new analysis to explicitly compare word overlap with topic vector correlations. We used Jaccard similarity (number of unique words in the intersection divided by number of unique words in the union) to quantify word overlap. As shown in this figure (copied below for convenience), the topic vectors for different questions often match up with a temporally well-defined segment of the corresponding lecture, whereas word matches alone are more diffuse:



Essentially, this is because simple word matching (right column) identifies moments when any of the *individual* words in the question are used during the lecture, whereas topic vector correlation identifies moments when the *underlying themes* (most probably) expressed by the question's specific combination of words were also expressed in the lecture. For example, in the first question (top row), mentions of "following," "difference," "gravity," "electromagnetic," and "force" at other moments in the lecture also evoke increases in Jaccard similarity.

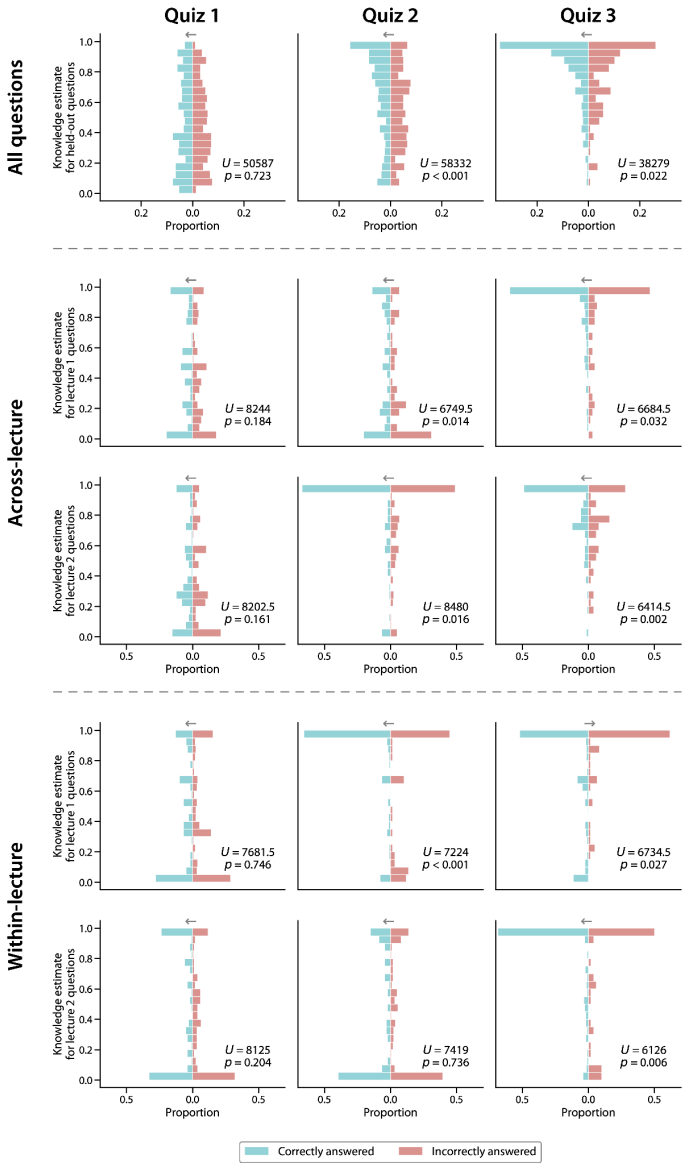
Reviewer 1 also raised a similar point about whether our approach is capturing "superficial" or "deep" knowledge (comment 1, above). For convenience, we've copied the relevant portion of our response to that comment below:

The deeper question of whether the knowledge maps we construct from participants' quiz performance capture true understanding versus something closer to memorization is trickier to get at. We think this comes down to the specific questions that we asked participants, as opposed to something about the embedding spaces. If the questions themselves require deep understanding (e.g., answering them correctly requires something beyond what can be achieved through memorization alone), then the maps we learn from participants' responses will also reflect deep understanding. If the questions themselves can be answered by memorization, then our maps will also reflect "memorized" concepts as opposed to true "understanding."

The questions we asked participants in our experiment vary with respect to how much we think they might capture memorization versus deeper understanding. For example, we suspect that questions like "In the famous equation attributed to Albert Einstein, E = mc^2, what does the letter 'm' represent?" could be answered by something close to pure memorization. But other questions, like "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. Which of the following explains why?" might require deeper understanding. The "amount of understanding" a given question "requires" to answer is somewhat subjective, and might even come down to the "strategy" a given participant is using to answer that question at that particular moment. We have added a note to this effect on page 31 of our revised manuscript.

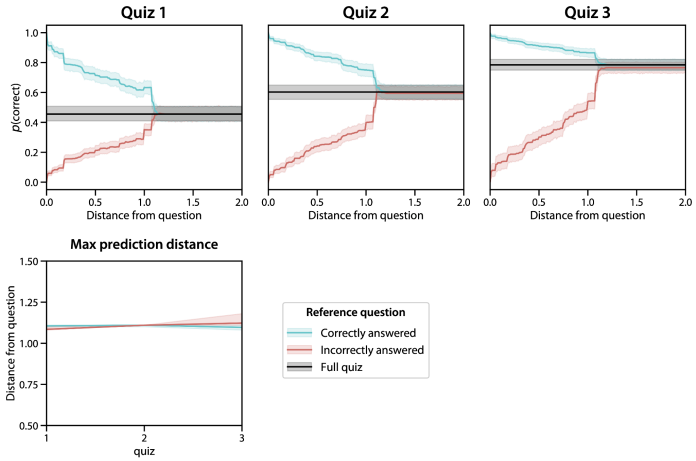
2. I would like to see the "Estimating Knowledge" analyses to go a step farther. The authors estimate participants' knowledge at the position of each quiz question's coordinate in a leave-one-out fashion, and show that these estimates reliably track whether the question was answered correctly or incorrectly. However, this analysis currently includes within-lecture prediction e.g., Lecture 1 quiz question performance to estimate a left-out Lecture 1 question. This analysis would be much more convincing and empirically interesting if a cross-lecture approach was used. That is, if for Quiz 2, performance on Lecture 1 questions could be used to predict accuracy on (unwatched) Lecture 2 quiz questions. This would be more convincing evidence that the embedding space is capturing deep conceptual meaning.

We've carried out the analysis the reviewer suggested (see Fig. 6, also copied below). In the previous version of our analysis (corresponding to the top row of the figure, labeled "All questions"), we used *all* (held-in) questions (about lecture 1, lecture 2, and general knowledge) from each quiz to predict knowledge on a single held-out question.



The reviewer's suggested analysis appears as a panel in the "Across-lecture" part of the figure (middle rows). Specifically, they suggest using Quiz 2 responses to Lecture 1 questions to predict accuracy on Lecture 2 questions from the same quiz (i.e., before Lecture 2 had been watched). This corresponds to the second row and middle column of the "Across-lecture" part of the figure. Indeed, we see that the predicted knowledge for held-out Lecture 2 questions that participants answered correctly is significantly higher than for held-out Lecture 2 questions that participants answered incorrectly. We also carried out other variants of this analysis, whereby we predicted knowledge for Lecture 1 questions using responses to Lecture 2 questions, and we also carried out the analysis separately for responses to each of the three quizzes. As the reviewer suggests, we see this as showing how the predictions obtained using our model can generalize across different content areas.

For completeness, we also carried out a third general class of test ("Within-lecture"), whereby we used the responses to Lecture 1 questions to predict knowledge for a held-out Lecture 1 question from the same quiz. We repeated this procedure using Lecture 2 questions to predict knowledge for a held-out Lecture 2 question, and for each of the three quizzes. This latter analysis was suggested by Reviewer 3, and helps to demonstrate the *specificity* of the predictions– i.e., the ability to generate predictions about conceptual knowledge at the resolution of different concepts covered within the *same* lecture: In addition to the analysis suggested by the reviewer, their comment led us to consider the broader question of "how far in the text embedding space knowledge estimates extend." For example, suppose we know that a participant answers a question (at embedding coordinate X) correctly. As we move away from X in the embedding space, how does knowledge (as estimated by quiz performance) "fall off" with distance? Or, suppose the participant instead answered that same question *in*correctly. Again, as we move away from X in the embedding space, how does what the participant *doesn't* know about the content change with distance? We reasoned that, assuming our space is capturing something about how participants actually organize their knowledge, conceptual knowledge right around X should be similar to the participant's knowledge of the content at X. And at another extreme, at some distance (after moving sufficiently far away from X), our guesses about what participants know (based on their response to the question at location X) should be no better than guessing based on their overall proportion of correctly answered questions—i.e., if Y is very far away from X, all we can do with the participant's response to X is guess that "their performance on quiz questions about Y is about equal to their average performance on quiz questions about *any* material." With these ideas in mind, we asked: conditioned on answering a question correctly, what proportion of *all* questions (within some radius *r* of that question's embedding coordinate) were answered correctly? We could then plot the proportion as a function of *r*. Similarly, we could ask, conditioned on answering a question *incorrectly*, how the proportion of correct responses changed with *r*. We found that quiz performance falls off smoothly with distance, and the "rate" of the falloff doesn't appear to change across the different quizzes, as measured by the distance at which performance becomes statistically indistinguishable from a simple proportion-correct score (Fig. 7):



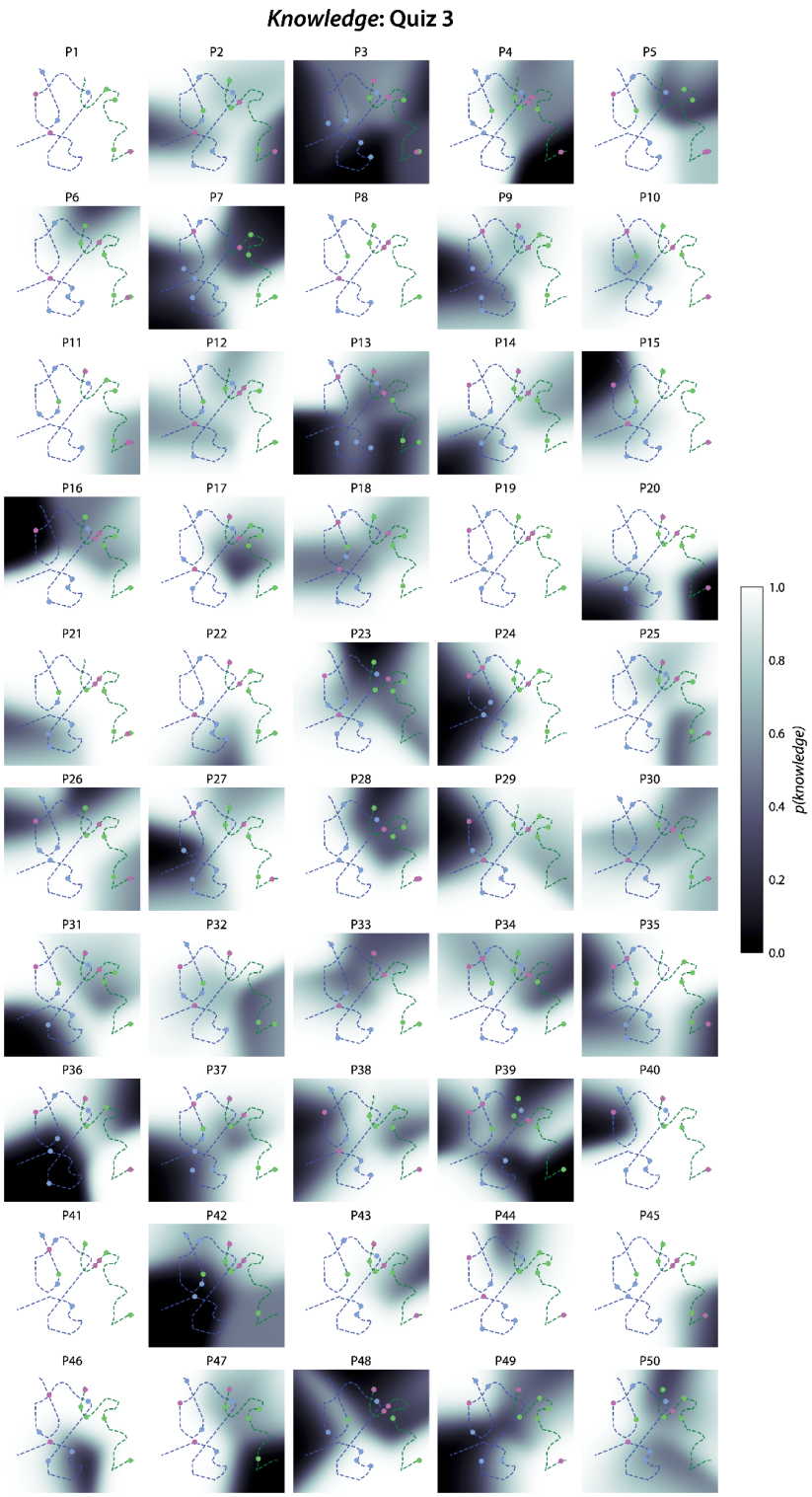
Taken together, this new analysis suggests that participants' knowledge (as reflected by their quiz performance on specific questions) changes relatively smoothly and gradually across our text embedding space.

3. In the current "Estimating Knowledge" analyses, I'm having a hard time understanding what is happening in Quiz 3. It makes sense that there's no difference in the correct/incorrect distributions in Quiz 1, and it's great that it emerges in Quiz 2, but I would have also expected to see the effect for Quiz 3 (and indeed a weak effect was found). The authors explain this as a "ceiling effect", but it doesn't look like the knowledge estimates are at ceiling in Fig6-Quiz3. Further, if the knowledge estimates based on the embedding space are not sufficiently predicting quiz question accuracy, then what does? In other words, if the knowledge estimates for all regions of the space were high, then why are the participants' responses wrong? Is the model not capturing other pertinent knowledge, or do the authors think participants were choosing answers randomly?

There are a few things going on here. First, as the reviewer notes, we do see a reliable effect for Quiz 3 (the test the reviewer is referring to is in the upper right panel of the figure we pasted in with our response to the previous comment).

In our prior submission, we speculated that the reason Quiz 2 responses were "easier" to predict than Quiz 3 responses had to do with the "contrast" between the highest and lowest knowledge within the embedding space. When participants take Quiz 2, they tend (on average) to have relatively high knowledge for Lecture 1 content and relatively low knowledge for Lecture 2 content. On the other hand, when participants take Quiz 3, they tend (again, on average) to have relatively high knowledge *everywhere* in the embedding space. Our "ceiling effect" comment was meant to reflect that pattern, but we can see how the wording was confusing. We ended up entirely re-writing that part of our results section (pages 14–17).

Regarding the reviewer's question about why some participants' responses are still wrong on Quiz 3, this is simply a reflection of the difference between knowledge on *average* versus knowledge for an *individual participant*. As the reviewer points out, and as shown in Figure 8, the average "Quiz 3" knowledge map shows relatively high levels of knowledge throughout the embedding space. Further, there is relatively little contrast between the most and least known content in the average map for Quiz 3 (e.g., as compared with the Quiz 2 map). However, when we examine the Quiz 3 knowledge maps for individual participants (Supp. Fig. 9), we can see that there is substantial variability:



As shown in the individual maps, not all individual participants are at "ceiling" even after watching both lectures. That we can predict knowledge for individual held-out questions shows that (at least to

some extent), participants are *not* simply responding at random. Rather, their responses appear to track with our predictions about what they do versus don't know.

4. I don't find the "Moment-by-moment" plots in Figure 5 to be useful, and in a sense they are misleading. With time on the x-axis, the implication is that the plot will be showing learning over time, as the lecture unfolds. This leads to the initial question of why the plots don't show an increase over time. However, this analysis does not reveal increasing knowledge over time, but rather the degree of knowledge relating to each moment in the lecture, assessed after the lecture was completed. What extra information are we getting from these moment-to-moment plots, that we're not getting from the averages presented in Fig 5B and 5D? Especially since there's no indication of how topics are changing from moment to moment, these plots seem superfluous. An alternate possibility is to show what topics participants know in each of the quizzes.

As the reviewer notes, the Figure 5 time course plots (Figs. 5A and 5C) show "*the weighted proportion of correctly answered questions about the content reflected in each moment of the lectures*" (Fig. 5 caption). The "averages" presented in Figures 5B and 5D show a sort of corrected version of an overall "proportion-correct" score for each of the three quizzes (the correction accounts for the fact that the questions do not tile the lecture content perfectly evenly). The time courses in Figures 5A and 5C show how much participants "know" about different *parts* of the lectures at the time they took each quiz. This can reveal interesting differences across the different content presented in the lectures. For example, before watching Lecture 1, participants tend to be relatively unknowledgeable about the content presented around the 3–4 minute range of Lecture 1 (as compared with, say, the content presented around the 0–2 minute range, or the 6–7 minute range). *After* watching Lecture 1, however (when participants take Quizzes 2 and 3), we see two main effects. First, participants' knowledge about every part of the Lecture 1 material is (on average) greater, relative to before they watched Lecture 1. Second, not all of their knowledge increases by the same amount. In particular, on average, participants' knowledge about the content from minutes 3–4 of Lecture 1 ends up being even *greater* than their knowledge about the content from 0–2 or 6–7 minutes. This suggests that some aspects of knowledge might be more "malleable" than others. This could reflect something about how the content is presented, an interaction with other aspects of participants' knowledge, or some other factor.

We see this ability to estimate participants' knowledge for individual moments of the lectures as an important advance of our method, enabled by our ability to map each quiz question onto the specific parts of the lectures that conveyed information relevant to correctly answering it (Fig. 4). For example, one could imagine these time-resolved knowledge estimates as the basis for an automated tool that helps students identify specific sections of course content to review, based on their performance on simple quizzes similar to those used in our study. Real-world instructors might also find value in identifying systematic differences in how successfully a class learns different sections of a lesson's content. In terms of the information that the time courses shown in Figure 5A and 5C provide over and above the averages shown in panels B and D, one could imagine two different series of moment-by-moment knowledge estimates: one in which estimated knowledge is generally low for

moments from the first half of the lecture and high for moments from the second half, and another in which estimated knowledge is generally high for the first half and low for the second half. Averaging both of these time courses across moments as we do in 5B and 5D would result in mean knowledge estimates that are similar or identical, despite the time courses themselves being meaningfully different. Importantfully, if these two time courses of knowledge estimates were derived from different individuals (or groups of individuals), both the contents of those individuals' knowledge and what an instructor could do to most effectively help them "fill in gaps" in their knowledge would also be meaningfully different.

Although we continue to see the Figure 5A and 5C time course plots as interesting and informative, we do take the reviewer's point that they can be misinterpreted. To help clarify, we have updated the figure's caption to make it more explicit that we are estimating knowledge about the *content* presented at each moment of the lectures, as opposed to estimating "how much is known *overall*" at each moment of training.

5. The Topic Weight Variability analysis (and corresponding Fig 3A) feels a bit lackluster. The authors have set up this rich, high-dimensional space in which to perform their other analyses, so it would be more satisfying if this analysis was also done using a geometric approach. For example, could variance be captured by the distance spanned by a lecture's trajectory in each of the 15 dimensions? Could variability be visualized in 3D space? I also think the decision to visualize the topic variance, rather than topic weights, could be better justified—it's not clear in the manuscript's current form why we should care more about variance than weights, which is more intuitively relevant.

We appreciate the reviewer's feedback and we share their enthusiasm for an aesthetically pleasing presentation. We have explored several ways of displaying the content of Figure 3A. We can provide some additional background to clarify our process, and our decision to ultimately maintain the prior presentation format.

Geometrically, topic overlap simply looks like "points that are nearby." We show in Fig. 2C how the lectures' and (individual) questions' embeddings relate when rendered in 3D. The challenge is in capturing the full embedding space, since it's difficult to visualize more than 3 dimensions. Our "bar graph" approach (while, we admit, is less "flashy") lets us zoom in on each individual topic dimension, which shows more completely the specific dimensions along which lectures and questions are similar or different. This is in contrast to projecting the lecture trajectories or questions into a lower-dimensional (e.g., 3D) space, where information is necessarily lost.

We also appreciate the reviewer's comment about topic variability versus topic weights. To clarify, both approaches yield similar looking visualizations (we provide the "topic weights" version of Fig. 3 in Supp. Fig. 2). That said, we think that the "variability" version is closer to what we *mean* when we think of two texts being "about" the same thing. One reason why weights alone can be misleading is somewhat technical and specific to constraints placed on LDA's topic vectors. Because topic vectors

must sum to 1, the number of topics "present" in any given document will influence the weights of the other topics. Therefore the raw (absolute) weights can conflate the presence of a given "theme" with the *absence* of other themes.

Even beyond the constraints that are specific to LDA, if a given topic (or a feature dimension of some other embedding model) is consistently set to the same (large) value across all timepoints in a lecture, it can't express any specific concepts in the lecture that span less than its entire duration. But if a given topic shows up during a specific interval in one lecture, that tells us that topic has something to do with whatever content was covered during that interval. We want to be able to pick out those sorts of "transient" patterns that reflect concepts at a lower level than "what the entire lecture is about on average." So we think that the notion of identifying topics for which some moments (or questions) are weighted strongly and others are weighted *weakly* is more indicative of those topics being important. In other words, that contrast between strong and weak weights that depend on the specific content being covered is what we see as diagnostic of the topic being "important." In practice, the two measures (variability vs. means) end up being more similar than we had anticipated (e.g., compare Fig 3 vs. Supp. Fig 2).

6. This a relatively minor qualm, but referring to a coordinate in the embedding space as "a single concept" or "individual concept" doesn't seem quite right (e.g., "We wondered whether our modeling framework might enable us to…infer participants' knowledge at the scale of individual concepts (e.g., as captured by a single moment of a lecture). The notion of a single concept implies that concepts are units of some kind, and capturing conceptual knowledge in a high-dimensional space runs counter to that view. Further, what does a "single concept" mean if the authors count "fish", "water" and "fish swim in water" to each be a concept (p. 3)? In this case, a full sentence would count as a concept, or even the information contained in a whole document. Maybe the authors take this view, but it feels particularly odd given the initial use of "fish" and "water" as initial examples.

This question of what a concept "is" is an important one. There are many potential ways to define what a concept is, and to construct hierarchies of subordinate and superordinate concepts. We view this as an ontological question far beyond the scope of the current manuscript, so in our manuscript we chose to treat a "concept" as equivalent to a text embedding coordinate. Since any amount of text can be embedded, our notion of a "concept" as framed in our paper is consistent with all of the examples the reviewer proposes—a single word, a phrase, a sentence, a paragraph, an entire document, or even a collection of documents. One way to "think about" what a concept is, in this framing, is to read word clouds of the sort presented in Figure 7C. If each individual word evokes some mental representation, then a "concept" is the weighted sum of the representations of all the words in the cloud (where the weights are given by the words' sizes). Under this framing, most concepts are not easily namable (at least not as single words or even short sentences). Still, this framing *is* still compatible with *some* concepts being nameable. For example, an "edge case" word cloud might contain a single word (e.g., "fish"), in which case the "meaning of that single word" is a "concept" by our definition. So in that sense, our "fish" and "water" examples aren't incompatible with our overall view of what concepts are. Nonetheless, we take the reviewer's broader point that some clarity is

needed. We have added a note about what we mean by "concept" (within the context of our paper) on page 3:

> **"...'concepts' are defined implicitly by the model's geometry (e.g., how the embedding coordinate of a given word or document relates to the coordinates of other text embeddings; Piantadosi and Hill, 2022)"**

Reviewer #3 (Remarks to the Author):

The authors apply a computational method called "topic modeling" to text transcribed from instructional videos (on two topics in physics). The result is a quantitative representation of the content of each moment of the video in the form of vectors embedded in a multidimensional space. They then represent the content of questions about these lectures in the same space. They report that this data-driven modeling approach (1) can match questions to their appropriate lectures, and even to the specific section of the lecture which covers the relevant content; (2) can capture how the knowledge of human learners evolves from pre-lecture to post-lecture; (3) estimate whether a learner will answer a question correctly or not based on their answers to other, related questions; (4) estimate the learner's knowledge of parts of the embedding space that were not explicitly covered by the lectures (or questions).

Whereas I find the overall approach interesting, I do not believe this work, in its current form, merits publication in this journal. The main reason is that I think there is a discrepancy between the claims that the authors make and the data they present. Specifically:

1. Most generally, throughout the paper different research goals are mentioned, and I found it somewhat difficult to discern what exactly the authors were aiming to do. The paper would benefit from clarifying the research goals. Of course, a single paper can achieve multiple goals, but in such a case it would be useful to state all of them together.

Our primary research goal is to advance our understanding of what it means to acquire deep, real-world conceptual knowledge (page 4). We break this down into two sub-goals (page 5): first, we want to gain detailed insights into what learners know at different points in their training. Second, we want our approach to be potentially scalable to large numbers of diverse concepts, courses, and students.

To state these goals differently, we want to map and track the acquisition of complex conceptual knowledge, like what students might learn in a course. Accomplishing this goal required developing an approach to *representing* knowledge (when the "number" of unique concepts one might want to consider is nearly infinite). We also wanted our approach to easily transfer to other sorts of courses, e.g., as opposed to being "hard coded" or as opposed to requiring human-generated labels.

1.1. The authors might be trying to come up with a new tool that would be applicable to educators. If application is one of the goals, I am not sure why the authors are not using available technology that can be more quickly adopted, such as large language models (LLMs) like GPT-2. Activity vectors from hidden layers in such models are available, so text excerpts from lectures and questions could be all embedded in the space of hidden activity, and it's plausible that those embeddings would be just as good as those obtained with the topic model. Moreover, such a model would not have to be re-trained for every new domain (whereas the author's model would likely have to be re-trained on

topics that are far from physics, such as literature, history, political science, computer science, etc.). One might argue that LLM representations are uninterpretable, in contrast to the "topics" (dimensions) of a topic model, but the current paper seems to have no interest in interpreting those "topics" anyway (they are not used in any analysis), so that argument is irrelevant. In fact, one might even imagine doing away with hidden activations altogether, and just giving ChatGPT a lecture text and a question, asking it to identify which part of the lecture is referred to by a question, and using the number of utterances between this part of the lecture and any other part to conduct similar analyses to those proposed by the authors (e.g., constructing knowledge maps). If the goal is to propose scalable tools for wide adoption, it appears that testing such models to see whether they could actually work is an important first step before proposing methods like those advanced by the authors.

We appreciate the reviewer's suggestions here. There are a few points we want to unpack.

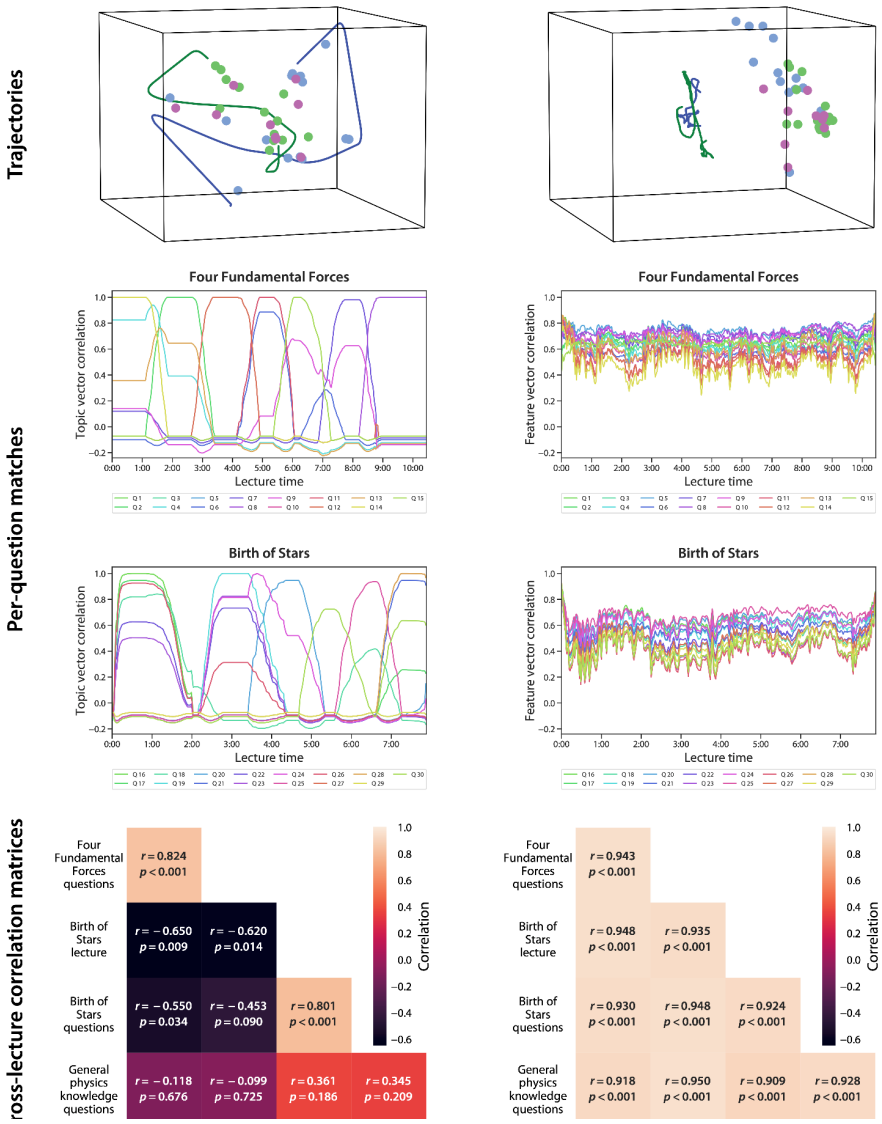
First, regarding our choice of model, one benefit of using simple topic models (LDA) over more sophisticated models like GPT-2 is that training is actually relatively inexpensive (with respect to time, compute power, and memory). For example, on the consumer-grade laptop we used to write and perform our analyses, training the model and transforming our complete document corpus takes 803 ms (+/- 5.23 ms; M +/- SD over 1,000 iterations). Therefore re-training our topic models on new domains is "scalable" in a way that would not be feasible for more sophisticated models or much larger corpora. Topic models can even be fit or updated to accommodate streaming data, often in near-real time (e.g., Hoffman, Bach, and Blei, 2010, NeurIPS).

That said, we've certainly considered using LLMs like GPT-2, ChatGPT, etc. In fact, in our other ongoing work (e.g., https://github.com/ContextLab/chatify) we've been developing LLM-based "tutors" to help students learn course materials. We agree that incorporating LLMs into education technologies to help teachers and students presents an exciting opportunity!

In our current study, our goal is to *map and track* what our participants know, and how they learn with training. This requires maintaining a formal representation of what material is being learned, what is known, and so on. For example, suppose one were to ask ChatGPT to match up each question with some part of a lecture. What would one "do" with that response? Ultimately, in our view, we would need to turn those "match labels" into numbers of some sort. And that would seem to be best accomplished by looking directly at the text embeddings (e.g., internal to ChatGPT), since ChatGPT's outputs can be unreliable, heavily dependent on the specific prompt used, and so on. Ignoring ChatGPT's reliability, though, 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. Building in those mechanisms would appear to require something like what we built in our current paper.

Another aspect of the reviewer's comment touches on which model might be "best" for constructing or learning text embeddings of lecture content. This idea relates to a comment also raised by Reviewer 1 in their comment 2b, about a "tension" between the potential generality of our approach versus the targeted nature of our particular embedding space (trained to the content of these two lectures). Although LLMs like BERT, GPT-{2, 3, 4}, ChatGPT, LLaMA, and so on are incredibly flexible,

generalizable, and powerful, they are surprisingly *bad* at distinguishing between subtle conceptual differences at the scale of parts of a single lecture video. As a quick example (and reproducing some of our response to Reviewer 1's comment 2b for convenience), we used BERT (a transformer-based LLM trained on a large corpus, with 768 feature dimensions) to carry out some of the core analyses from our main paper (i.e., swapping out BERT for the topic modeling approach we reported on):



Quoting from our previous response:

> We highlight three general differences in the above figure. In the top row ("Trajectories") we show a 3D projection (using PCA) of the content trajectories for each lecture (blue and green lines) and each question (blue, green, and purple dots). The LDA embeddings exhibit several desirable properties compared with the BERT embeddings:

The LDA embeddings of the lectures and questions are "near" each other– e.g., the convex hull enclosing the two lectures' trajectories is highly overlapping with the convex hull enclosing the questions' embeddings. The BERT embeddings of the lectures and questions are instead largely distinct.
The LDA embeddings of the questions for each lecture and the corresponding lecture's trajectory are similar. E.g., the LDA embeddings for lecture 1 questions (blue dots) tend to appear closer to the lecture 1 trajectory (blue line), and the LDA embeddings for lecture 2 questions (green dots) tend to appear closer to the lecture 2 trajectory (green line). The BERT embeddings do not show this property.

In the middle row ("Per-question matches"), we display the correlations between each question's embedding coordinate and the embedding of each moment of the corresponding lecture. Again, the LDA embeddings show several desirable properties compared with the BERT embeddings:
The time series plot of individual questions' correlations are different from each other when computed using LDA (e.g., the traces can be clearly visually separated), whereas the correlations computed from BERT embeddings of different questions all look very similar. This tells us that LDA is capturing some differences in content between the questions, whereas BERT is not.
The time series plots of individual questions' correlations have clear "peaks" when computed using LDA, but not when computed using BERT. This tells us that LDA is capturing a "match" between the content of each question and a relatively well-defined time window of the corresponding lectures. The BERT embeddings appear to blur together the content of the questions versus specific moments of each lecture.

In the bottom row ("Cross-lecture correlation matrices") we compare the pairwise correlations between embeddings of lectures and questions within versus across content areas (i.e., content covered by the individual lectures, lecture-specific questions, and by the "general physics knowledge" questions). The LDA embeddings show a strong contrast between same-content embeddings versus across-content embeddings. In other words, the embeddings of questions about the lecture 1 ("Four fundamental forces") material are highly correlated with the embeddings of lecture 1, but not with the embeddings of lecture 2, questions about lecture 2, or general physics knowledge questions. We see a similar pattern with the LDA embeddings of the lecture 2 ("Birth of stars") questions. In contrast, the BERT embeddings are all highly correlated with each other.

Taken together, these comparisons illustrate how LDA (trained on the specific content in question) provides both coverage of the requisite material and specificity at the level of the content covered by individual questions. BERT, on the other hand, essentially assigns both lectures and all of the questions (which are all broadly about "physics") into a tiny region of

its embedding space, thereby blurring out meaningful distinctions between different specific concepts covered by the lectures and questions.

We note that these are not criticisms of BERT (or other large language models trained on large and diverse corpora). Rather, our point is that simple fine-tuned models trained on a relatively small but specialized corpus can outperform much more complicated models trained on much larger corpora, when we are specifically interested in capturing subtle conceptual differences at the level of a single course lecture or question. Of course if our goal had been to find a model that generalized to simultaneously capturing many different content areas, we would expect our approach to perform comparatively poorly relative to BERT or other much larger models. We suggest that bridging the tradeoff between high resolution within each content area versus the ability to generalize across many different content areas using a single set of model weights will be an important challenge for future work in this domain.

We have added a discussion of these issues to pages 23–25 of our revised manuscript, and have also added the figure above to our supplementary materials (Supp. Fig. 6).

1.2. To follow up on the issue of technology: ways to evaluate the specific knowledge of students already exist. For instance, some instructors give specific feedback on specific questions in online quizzes; I personally weigh different questions based on whether they require knowing things that were explicitly stated in class vs. near-transfer to new examples vs. far-transfer that also requires combination of knowledge across different lectures; and, more generally, there is a vast literature spanning 2-3 decades on the importance of specific feedback to students. The question is why these methods are not more wildly adopted. If the reason is related to scalability / time / effort, it is currently unclear whether the authors' method is better suited for wide adoption compared to using something like ChatGPT (I am by no means a fan of ChatGPT, but the fact remains that it is overall surprisingly good at tasks that are not dissimilar to what the authors are trying to do).

Personalized human instruction is hard to beat; we're not aware of any modern approaches, including ChatGPT, that approach the level of an experienced and expert human instructor. The reviewer's examples of providing personalized feedback, manually tagging (in advance) which questions are associated with which content areas, considering a variety of questions that test understanding at different levels, and so on are all excellent ways to effectively teach. One could imagine that an experienced one-on-one tutor could employ all of these techniques, perhaps combined with creative ways of teaching the material, to serve as a "gold standard" of sorts, with respect to what one might hope to achieve with an automated tutoring approach if all of the appropriated pieces were in place.

What might be required to build such a system? One consideration is that effective tutors might hold in mind a representation of what their student *already* knows, what they are ready to learn, what the learning objectives are, what previous concepts they learned quickly versus struggled with, and so on. For a given content area or skill (e.g., "computing derivatives of trigonometric functions") there will be a set of relevant concepts the student will need to grasp in order to have mastered the material. Our current framework for mapping out participants' knowledge (and how their knowledge changes over

time with training) serves as a "prototype" of what an automated tutor might incorporate as its internal representation of "what the student knows" at a given moment. A desirable aspect of our approach is that our mapping procedure can generalize to any content area (i.e., there is nothing inherent to the method that makes it better suited to the particular content we happened to select; we could have just as easily chosen courses in art history instead of physics, and one could reasonably expect similar results and performance). Of course we have not explicitly tested every possible content area, but we see our work as a "first step" in the direction of a generalized mapping approach.

Another consideration is how the learner's knowledge might be estimated. If one has the benefit of a human tutor, the tutor can (as the reviewer suggests) manually "match up" each question with core concepts that their student might need to know. But *without* a human tutor, could this matching process be carried out automatically? Again, we show a proof of concept demonstration that we can *automatically* match up specific questions with specific moments in existing lectures (Fig. 4, Supp. Tab. 3). Further, we provide a formal approach to estimating knowledge using the learner's responses to a relatively small number of questions. And by showing that we can accurately predict knowledge for held-out questions (Fig. 6), we show that our approach can generalize to concepts outside of the immediate training set used to construct the knowledge estimates. This is a critical requirement for building scalable tutoring systems that can work across a variety of content areas.

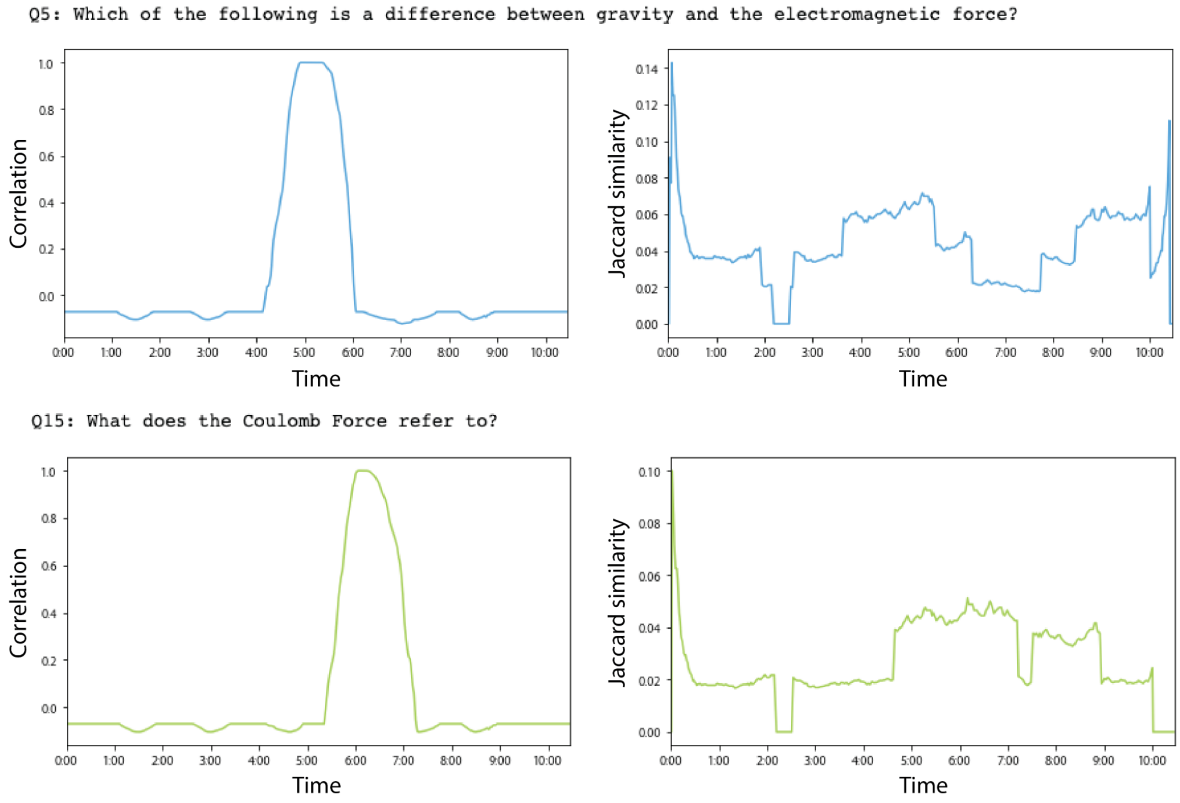
As we noted above, while we (like the reviewer) see some limitations of tools like ChatGPT, we also agree with the reviewer that ChatGPT can serve as a useful component of automated tutoring systems. But ChatGPT has no internal machinery for representing or tracking the learner's knowledge, nor does 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. So neither ChatGPT nor other related tools or systems we are aware of can solve the problem of mapping out or tracking knowledge in the scalable (i.e., automated) and generalizable way we present in our paper.

To address this issue, we have added a comment to the discussion section comparing our approach to other "alternatives" like ChatGPT, and clarifying what we see as some benefits and differences of our approach (page 25).

1.3. The authors might also be interested in providing a computational model that tells us something about human psychology (e.g., learning processes). In this case, the representational assumptions that the model makes require some discussion. On the one hand, it would be important to compare the model to "baseline" or "control" models in order to claim that the current model is useful (compared to, e.g., averaging word2vec embeddings of words in each lecture sentence or question and using representations; this might be what the authors mean by "superficial matching"?); on the other hand, it would be important to discuss claims that embeddings models are a non-starter for complex knowledge representation because they do not have any explicit notion of causal schemas like "knowledge graphs" / "concept maps" / "theory theory", etc. In a sense, the authors' model might occupy an "uncanny valley", being too complicated compared to simpler models that can achieve the same empirical results without claiming any psychological reality, but not structured enough to be a quantitative theory of human learning.

The reviewer brings up several points here. We'll address each in turn.

First, the reviewer proposes comparing our approach (using topic models) to some alternatives. In our response to this and Reviewer 1's comment 2c, we compare our topic modeling approach to a simpler "word matching" approach (i.e., what we meant by "superficial matching") within the context of "matching up" individual questions with specific moments of each lecture. Our topic models tend to identify well-defined time intervals in the lecture that match up well conceptually with the questions (e.g., Fig. 4 shows some example time courses and question/lecture matches, and we have added Supp. Figs. 3 and 4 and Supp. Tab. 3 to show these time courses and text matches for all of the questions). On the other hand, simpler word matching does *not* behave nearly as well. For convenience we have pasted in the figure we included in response to Reviewer 1's comment; the figure shows some example time courses for questions, as computed using topic models (left) and word matching (right):



As we also note above (in response to Reviewer 2's similar comment):

> **[T]his is because simple word matching (right column) identifies moments when any of the individual words in the question are used during the lecture, whereas topic vector correlation identifies moments when the underlying themes (most probably) expressed by the question's specific combination of words were also expressed in the lecture. For example, in the first question (top row), mentions of "following," "difference," "gravity," "electromagnetic," and "force" at other moments in the lecture also evoke increases in Jaccard similarity.**

At the other end of the "methods spectrum," one could imagine replacing our topic modeling approach with "fancier" (and orders of magnitude larger, both with respect to number of parameters and training corpus size) transformer-based models like BERT, GPT-{2, 3, 4}, LLaMa, etc. This doesn't work particularly well either, as we report in response to the reviewer's comment 1.1, above.

In this sense, we would argue that our approach actually occupies the *opposite* of an "uncanny valley"—rather, we think our approach occupies a sort of "sweet spot" that enables us to capture the relevant content at the appropriate semantic scale. Our approach enables us to accurately and consistently identify each question's content in a way that also matches up with what is presented in the lectures. In turn, this enables us to construct accurate predictions about participants' knowledge of the conceptual content tested by held-out questions (Fig. 6).

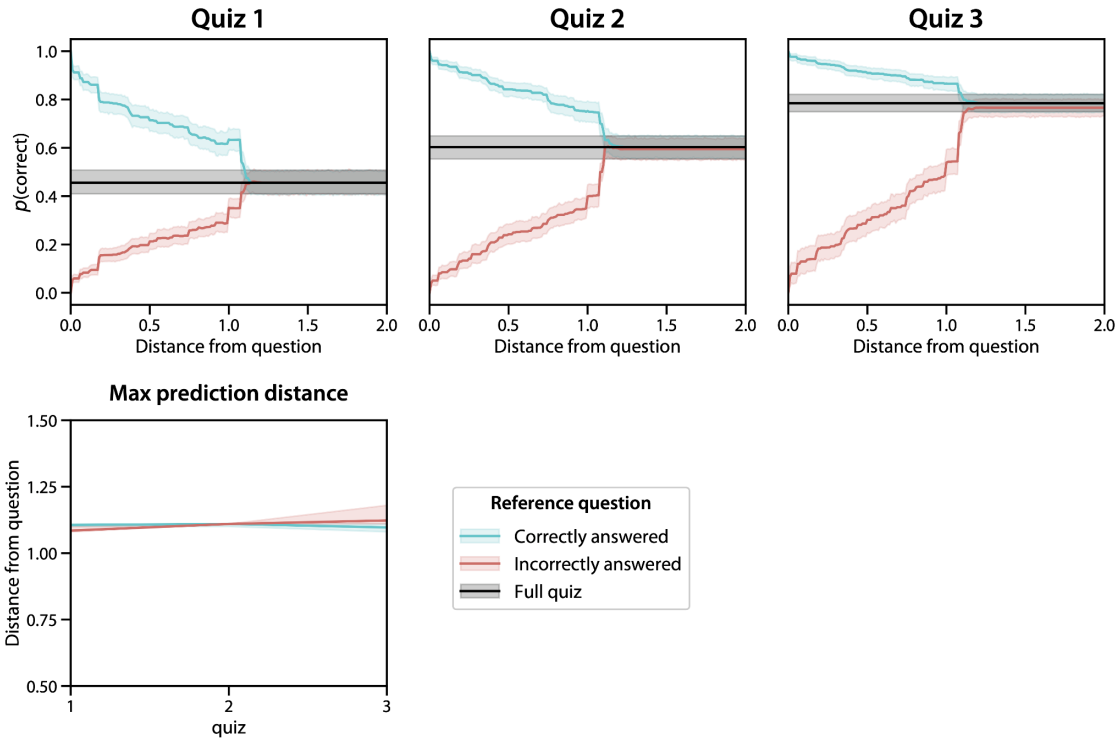
We agree with the reviewer that topic models contain no explicit internal representations of more complex aspects of "knowledge" like knowledge graphs, dependencies or associations between concepts, and so on. We have added a note (page 26) to clarify that we are not claiming that our approach incorporates these sorts of representations, but that such features might be added as extensions to our approach to more accurately and precisely capture, characterize, and track knowledge.

1.4. The authors might be interested in providing a quantitative measure of how knowledge "behaves". In this case, I think the authors' claims are over stated, especially with regard to "high-resolution insights", as I discuss in (2) below.

We have added several new tests and analyses to bolster our "high-resolution insights" claim. We describe several of them below in response to the reviewer's specific comments. We also ran an analysis specifically related to how knowledge "behaves" as one moves through the text embedding space. We described it above in Reviewer 2's comment 2, but we have copied the text and figure below for convenience:

> **In addition to the analysis suggested by the reviewer, their comment led us to consider the broader question of "how far in the text embedding space knowledge estimates extend." For example, suppose we know that a participant answers a question (at embedding coordinate X) correctly. As we move away from X in the embedding space, how does knowledge (as estimated by quiz performance) "fall off" with distance? Or, suppose the participant instead answered that same question incorrectly. Again, as we move away from X in the embedding space, how does what the participant doesn't know about the content change with distance? We reasoned that, assuming our space is capturing something about how participants actually organize their knowledge, conceptual knowledge right around X should be similar to the participant's knowledge of the content at X. And at another extreme, at some distance (after moving sufficiently far away from X), our guesses about what participants know (based on their response to the question at location X) should be no better than guessing based on their overall proportion of correctly answered questions—i.e., if Y is very far away from X, all we**

can do with the participant's response to X is guess that "their performance on quiz questions about Y is about equal to their average performance on quiz questions about any material." With these ideas in mind, we asked: conditioned on answering a question correctly, what proportion of all questions (within some radius r of that question's embedding coordinate) were answered correctly? We could then plot the proportion as a function of r. Similarly, we could ask, conditioned on answering a question incorrectly, how the proportion of correct responses changed with r. We found that quiz performance falls off smoothly with distance, and the "rate" of the falloff doesn't appear to change across the different quizzes, as measured by the distance at which performance becomes statistically indistinguishable from a simple proportion-correct score (Fig. 7):



2. The authors claim that they provide "nuanced insights into what learners know and how their knowledge changes with training". I do not believe this is the case.

We have attempted to clarify this claim in our revised manuscript. The "nuanced insights" phrasing is meant to refer to predictions and maps at the conceptual level of individual moments of the lectures (Fig. 5) and individual quiz questions (Fig. 6). Our "changes with training" phrasing simply refers to our updated estimates using each set of quizzes in turn.

We also created detailed knowledge and learning maps before/after watching each lecture (visualized in Fig. 8). This gives us another high-resolution view of participants' knowledge, and those maps change over time as the participants learn. We also test the accuracy of the maps by asking whether

the knowledge predicted at the embedding coordinates of held-out questions tracks with participants' performance on those questions (Fig. 6), and we characterize how "knowledge" (as estimated using quiz performance) changes with distance in the embedding space (Fig. 7).

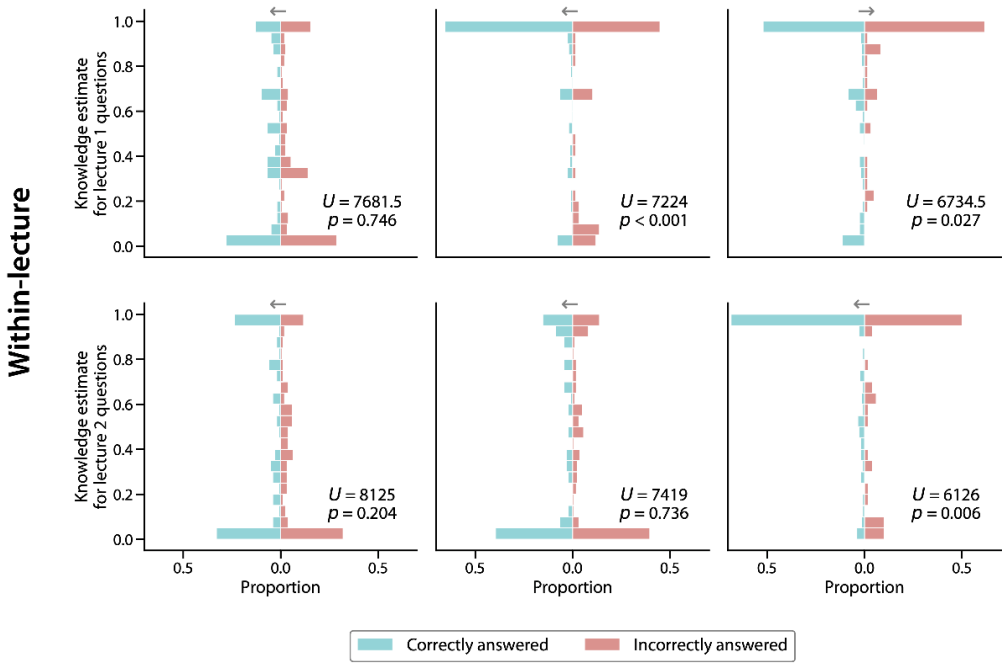
2.1. Nearly all analyses yield the same type of knowledge that would obtained from calculating "proportion correct" separately for each subset of questions (Birth of Starts, Four Fundamental Forces, general physics). Namely, students do not know any topics before a lecture; after lecture #1, they learn the content of that lecture but not unrelated content from lecture #2 which they haven't yet watched; and after viewing lecture #2, they also learn its content. Despite the authors' claims for higher-resolution quantification of how knowledge evolves throughout a lecture, their analyses in Figures 3, 5, 6, and 7 only provide evidence for this general pattern. For instance, we can compute % correct on 4 questions about Birth of Stars to predict the correctness of the 5th question; would this analyses be significantly inferior to the one proposed by the authors? Overall, my impression is that the authors do not quantitatively evaluate any pattern that is more nuanced than "knowledge of lecture #1 and knowledge of lecture #2 can be separated".

Our approach and results yield a number of insights into participants' knowledge and learning that simpler "proportion correct" measures would not (directly) support. First, in panels A and C of Figure 5, we show estimates of participants' knowledge about each moment of each of the lectures, which we obtain by leveraging our ability to map quiz questions they answered correctly and incorrectly onto specific relevant time periods of the lectures' contents. While participants do indeed tend to know more about each lecture's content overall after viewing it than they did before (as one might expect), these temporally specific ("high-resolution") knowledge estimates provide additional insights into which *parts* of each lecture participants know comparatively more or less about, as well as which parts they learn more or less successfully through viewing it. The analyses shown in Figure 6 show that these insights generalize to the level of individual quiz questions. The analyses in Figures 7 and 8 show how quiz question performance (our proxy for "knowledge") changes across locations in the text embedding space (i.e., across different "concepts"). Based on the reviewer's comment regarding our approach quantifying "*how knowledge evolves throughout a lecture*", as well as their comment 2.3 below and a related comment by reviewer 2 (comment 4), we realize that these plots can be misinterpreted as displaying how much participants know *overall at* each moment of the lectures, rather than characterizing knowledge *about the content* at each timepoint (of each lecture). We have amended the caption for Figure 5 to clarify what we are trying to show in panels A and C.

As shown in Figure 7 (also included in our response to the reviewer's comment 1.4), we show in our revised manuscript that our predictions are substantially more precise ("high-resolution") than simple proportion-correct measures, provided that we are attempting to estimate knowledge within a correlation distance of (roughly) 1 from the embedding coordinates of a question they answered. Beyond a correlation distance of ~1, our predictions approach what could be obtained through simple proportion-correct measures (Fig. 7B).

We also specifically tested the reviewer's concern about whether the knowledge patterns we identify are more nuanced than the level of different lectures (this has been incorporated into Figure 6 in our

revised manuscript). In an excerpt from this figure below, we show knowledge predictions obtained using only the other questions about the same lecture's material (top row: lecture 1; bottom row: lecture 2; left column: Quiz 1; middle column: Quiz 2; right column: Quiz 3):



We show that we can (often, though not always) accurately predict participants' knowledge of a held-out question, using only the other four questions about the same lecture, from the same quiz. We expected these predictions to be quite noisy, considering each knowledge estimate is derived from just four other responses to multiple questions, but we nonetheless show that knowledge at the coordinates of held-out questions participants answered correctly is often predicted to be higher than at the coordinates of held-out questions participants answered incorrectly.

2.2. The authors claim: "we show that our approach can automatically match the conceptual knowledge probed by individual quiz questions to the corresponding moments in lecture videos when those concepts were presented". This is inaccurate. First, the authors show that their method can match quiz questions to the overall lecture they correspond to (Figure 3B), but this is only a sanity check (as the authors acknowledge), not a technological advancement – instructors who write questions can tag them for which content they are testing, and we hardly need a computational tool to achieve this goal; in fact, this appears to be what instructors mentally do when trying to come up with questions to test different areas of content, and what the authors of this manuscript themselves did when creating questions for lecture #1 vs. #2. Second, the authors report qualitative impressions that each question can be further matched to a particular timeframe within a lecture (Figure 4), but I do not believe they have provided any test of this impression (my apologies if I missed it!). For instance, if instructors are given the questions and are asked to match them to specific timeframes within the lecture – do they provide the same estimates? (Or, if they are

given 3 timeframes identified by the model, can they tell which one corresponds to a given question?). If a student missed a question and is referred back to a particular timeframe of the lecture to re-watch, do they do better than a student referred to a different part of the lecture? Here, too, I wonder: how useful is the "high resolution" match between question and content, beyond what instructors do anyway when compose questions? (e.g., I know exactly which slides are pertinent to each question on my exams, and it's possible ChatGPT would be able to match questions to the relevant part of a lecture).

We'd like to push back on the reviewer's claim that we are matching up questions and lectures only at the level of the entire lecture, as opposed to what we describe as "corresponding moments" in individual lectures. In Figure 4, for example, we show that each question's embedding displays a time course of correlations with the moment-by-moment embeddings of the lecture that tends to "peak" at a relatively well-defined time interval within the lecture. Further, when we manually examine the question text versus the lecture transcripts at those best-matching moments, they nearly always "align" with respect to which concepts the question is asking about versus what the lecturer is discussing at those moments. This alignment can be assessed by manually comparing the lectures' and questions' text. In addition to the specific examples we provide in Figure 4, we have added two new supplemental figures showing these "peaks" for each question (which we detect automatically; Supp. Figs. 3 and 4), as well as a new supplemental table showing the question and lecture text for *all* lecture 1 and 2 questions; Supp. Tab. 3).

We also wish to clarify that while we agree that a human instructor could manually match up quiz questions with moments in the lectures, the goal in our manuscript is to do this matching *without* human-generated tags. As we mention in response to the reviewer's comment 1.2 above, we see this move from manually generated to automatically generated tags as a necessary part of scaling up automated teaching tools beyond what could be accomplished by human instructors alone.

Finally, we appreciate the reviewer's suggestion that ChatGPT might be able to match up questions with the relevant parts of each lecture. The challenge (as we also described in response to the reviewer's comment 1.2) is that it's not clear what one would then "do" with those ChatGPT-derived labels. Constructing explicit knowledge estimates using our framework requires knowing the specific embedding coordinates of each question. That also lets us say, not only that a given question is mostly "about" the content covered by some part of a lecture, but it also lets us quantitatively compare the similarity of content covered by any (arbitrary) pair of questions. In turn, we use those similarity values to construct weighted averages that form our knowledge estimates.

2.3. The authors claim: "we demonstrate how we can estimate moment-by-moment "knowledge traces" that reflect the degree of knowledge participants have about each video's time- varying content, and capture temporally specific increases in knowledge after viewing each lecture": this claim has the same problems detailed in 2.2; namely, all quantitative analyses in the paper focus on differences in knowledge between pre- to post-lecture, not on demonstrating that the time-varying trace itself captures anything meaningful.

We show (e.g., in Fig. 6) that we can predict knowledge about the content captured by individual questions (on individual quizzes), which in turn localize to specific well-defined windows in the

lectures (Fig. 4). Our pre- versus post-lecture tests (e.g., Fig. 5B, 5D) are *not* intended to be temporally specific, since those tests average across all timepoints in the relevant lecture(s). As Reviewer 2 pointed out (comment 4), and as mentioned in our response to comment 2.1, our previous description of what is shown in Figures 5A and 5C required some clarification, which we have made by editing the Figure 5 caption. In those "traces" we are characterizing knowledge about the *content* at each timepoint (of each lecture), not "what or how much is known at each moment of training."

2.4. The authors claim: "these knowledge estimates can generalize to held-out questions": there is not test demonstrating that this analysis (Figure 6) does anything more than capturing low-resolution differences between entire topics. Intriguingly, the results of Quiz 3 suggests that this might be the case: contra to the authors' claim, Quiz 2 is not the most sensitive test, because passing it merely requires distinguishing between questions about lecture #1 from questions about lecture #2, which is a very low-resolution test (students haven't learned content that they haven't been exposed to); Quiz 3 is the one that requires the highest sensitivity, assuming that students know some parts of each lecture better than other parts of that same lecture, and that there are individual differences in which parts are more vs. less understood. Such a case would allow to test the authors' claim that knowing which questions in a given lecture a student has answered correctly (rather than an overall estimate of a student's score on that lecture) can predict which other questions from that lecture they will answer correctly. If the assumption of variations in knowledge throughout a lecture and across students does not hold and, instead, there is a "ceiling effect" as the authors claim (all students know all sections of both lectures very well), then the design of the study is inappropriate for testing high-resolution knowledge.

Reviewer 2 had a somewhat different take on this idea (comment 2). They argue that it is specifically *across*-lecture predictions that show that we are capturing deep conceptual meaning, as opposed to something coarser. They suggested using lecture 1 questions to predict knowledge for held-out lecture 2 questions (and vice versa) from the same quiz. We show in our revised Figure 6 that estimated knowledge for held-out other-lecture questions that participants answered correctly is higher than for held-out other-lecture questions that they answered incorrectly. This holds for Quizzes 2 *and* 3 (as do our predictions using *all* questions from each quiz to predict a single held-out question).

We also ran the "within-lecture" variant of this analysis that the reviewer is proposing here, in response to the reviewer's previous comment (2.1); we pasted in the relevant results in response to that comment (above; also see Fig. 6, "within-lecture" plots). In summary, after participants watch lecture 1 (i.e., on Quiz 2), we are able to reliably predict knowledge for held-out lecture 1 questions using only other lecture 1 questions from the same quiz. Similarly, after participants watch lecture 2 (i.e., on Quiz 3), we are able to reliably predict knowledge for held-out lecture 2 questions using only other lecture 2 questions from the same quiz. (We also ran other variants of these tests; the full set of tests is reported in Fig. 6.)

2.5. The authors claim: "visual maps that provide snapshot estimates of how much participants know about any concept within the scope of our text embedding model, and how much their knowledge of those concepts changes

with training": again, their analyses seem to capture differences between two distinct lectures, not across moments / topics within a lecture.

Our new "knowledge falloff" analysis (Fig. 7; also pasted into our response for the reviewer's comment 1.4) shows that the accuracy of our knowledge estimates falls off gradually and smoothly with increasing distance in the text embedding space. This goes beyond simply distinguishing between questions from one lecture versus the other.

As we mention in our response to the reviewer's comment 2.4, our new "within-lecture" knowledge prediction analysis (Fig. 6, "within-lecture" panels) also shows that we can capture knowledge at the sub-lecture resolution (i.e., knowledge about a single held-out question, given other questions from the same lecture and from the same quiz).

2.6. The authors claim: "our work suggests a rich new line of questions about the geometric "form" of knowledge": again, the low-resolution nature of the authors' results seem quite far from studying the geometric form of knowledge, compared to existing lines of work (both recent ones, and ones that date back 20 years). For instance, some studies have focused on the intrinsic dimensionality of content manifolds (e.g., "Intrinsic Dimension Estimation for Robust Detection of AI-Generated Texts"), or on the linear geometry of word embeddings and how it reflects structured world knowledge (e.g., "How do blind people know that blue is cold? Distributional semantics encode color-adjective associations"). Other models have explored how the geometry of knowledge evolves over time (e.g., Jay McClelland's early work on connectionist models of semantic knowledge; or papers like "Structured Semantic Knowledge Can Emerge Automatically from Predicting Word Sequences in Child-Directed Speech").

By "geometric," we are referring to geometric (e.g., shape or distance-based) relations between different concepts. Our embedding spaces define geometries that describe how concepts relate: difference in "meaning" is reflected by Euclidean distance in those spaces. In general, that we can explain participants' behaviors using these representations suggests that the text embedding spaces we use to "map out" knowledge are capturing at least some aspects of how knowledge varies across different content for our participants. Our new "knowledge falloff" analysis (Fig. 7; also pasted into our response for the reviewer's comment 1.4) also shows how, given that you know "something" (e.g., at the coordinate of some question) that knowledge (or lack thereof) changes as a function of distance in the embedding space.

We appreciate the reviewer's pointers to other related work, and we have added in citations of these papers to our revised manuscript.

2.7. One way to reframe what the authors are doing is to treat the two lectures are smaller sections within a long lecture in an academic course (e.g., a 90 minute class). In this case, distinguishing between the two lectures is akin to breaking an entire lecture into smaller ("higher-resolution") parts. However, I do not believe this is how the authors conceive of their work, because if that was the case, they would not have visualizations like those in Figure 4 or 5A, which are about second-to-second changes within a 10-minute lecture.

First, if we're interpreting the reviewer's comment as intended, our understanding is that the reviewer is following up on their prior suggestions that we are distinguishing concepts only at the level of entire "lectures" as opposed to at the sub-lecture resolution (e.g., smaller concepts mentioned within a given lecture video). In our responses above (and with the new analyses we've added), we've attempted to clarify our approach and findings. For example, in Figure 6 we show that we can estimate "knowledge" (as reflected by quiz performance) at the level of individual questions, which in turn tend to map onto relatively well-defined intervals within the lectures (Fig. 4, Supp. Tab. 3).

Regarding the reviewer's comment about how we conceive of our work, we can also unpack how we are thinking about "concepts" and "lectures." As we note on page 3, we treat concepts as equivalent to embedding coordinates; we see them as defined implicitly by the embedding model's geometry. Within the individual lectures, concepts build on each other: if one were to chop up a single lecture into "conceptual chunks" and watch the different segments "out of order," the material would be very difficult to follow. In contrast, across the lectures we showed participants in our study, there are no dependencies (as noted on page 6 of our manuscript, we intentionally chose introductory videos from two different lecture series). In other words, watching lecture 1 before lecture 2 is no more difficult to follow than watching lecture 2 before lecture 1.

Regarding Figures 4 and 5A, our general goal is to show that the conceptual content of a single lecture is not a uniform thing, but rather it varies from moment to moment as different material is discussed or presented. This would be true whether we consider what we call "lectures 1 and 2" to be parts of a longer "unified" lecture, or whether we (as in our current framing) consider lectures 1 and 2 to be distinct "units" for the purposes of our analysis.

3. Some minor comments are below. Some of these are suggestions, whereas others are thoughts that I jotted down as I was reading the manuscript and would perhaps be interesting or useful to the authors, but are equally likely to be irrelevant musings.

We appreciate the author's comments and suggestions, and for taking the time to write these thoughts out!

3.1. Introduction, line 60: "These models consider not only the co-occurrences of those elements within and across documents, but also patterns in how those elements appear across different scales (e.g., sentences, paragraphs, chapters, etc.), the temporal and grammatical properties of the elements, and other high-level characteristics of how they are used". I think this claim could benefit from some clarifications. For instance, doesn't LDA treat documents as "bags of words", looking only at co-occurrences? (The researcher can cut down a document, as is done here, but this is not a feature of the text embedding model – the model treats whatever "unit" of text it receives as a bag). In static word embeddings, co-occurrences are limited to local ones (e.g., a context window of 10 words), so there is not notion of paragraph / chapter, so there is more or less a single temporal scale being modeled. And in contemporary LLMs, some of the more "complex" features, such as grammatical features, are implicitly inferred from co-occurrences (there is no separate mechanism for learning these properties; a simplified form of them even arises in GloVe or word2vec).

Great point. We have added some text to the discussion to clarify that our framework is based around a "bag of words" model, and we also provide some additional rationale for why we chose to use LDA, e.g., as opposed to more "modern" LLMs (pages 23–25, Supp. Fig. 6).

3.2. Line 63: "A model that succeeds at capturing an analogue of "understanding" is able to assign nearby feature vectors to two conceptually related documents, even when the specific words contained in those documents have very little overlap." It seems to me that this task is very easy for LLMs (they can learn which words mean similar things based on distributional patterns). The harder thing is to separate documents with very similar words that talk about conceptually very different things (e.g., distinguishing between statements about A causing B vs. A correlating with B due to a shared cause vs. B causing A...).

We agree, and we have added a note to the discussion to this effect (page 26).

3.3. Line 72: "For example, understanding the concept of a fish swimming in water first requires understanding what fish and water are." From the point of view of conceptual role semantics (or "theory theory"), understanding the relationship between fish and water is precisely what is important for learning what those terms are (rather than some "dictionary definition" of their inherent properties). See, for instance, https://arxiv.org/abs/2208.02957.

We appreciate the pointer to this paper; we've added a citation along with a note to this effect (page 3).

3.4. Figure 5A: the traces look almost like mirror images of one another. What is the reason? Is this caused by time-points which are vs. are not covered by questions?

Sort of! First, we have a similar intuition to the reviewer that the shapes of those traces reflect how the questions "cover" the different moments of the lecture (e.g., Fig. 4). Second, the traces in Figure 5A aren't *quite* mirror images. This is actually a sort of visual artifact resulting from how participants' performance on certain questions (that cover particular intervals of *Four Fundamental Forces*) changes before vs. after watching that lecture.

The main intervals in the lecture that give rise to the "mirror image" appearance in Figure 5A are (roughly) between 3–4 minutes and 9–10 minutes. In our new Supplementary Figure 3, we show time-varying correlation plots for each individual question. In that figure, we can see that Questions 2, 3, 13, and 8 are the most correlated with those intervals in the lecture (and therefore carry the most influence over the time-varying knowledge estimates at those intervals). It happens that participants tended to get those questions wrong more often (relative to the other questions) before watching the *Four Fundamental Forces* lecture, but they tended to get those questions *right* more often (again, relative to the other questions) *after* watching that lecture. So the shapes of those traces actually reflect participants' behaviors (quiz performance), which seems to differ in meaningful ways for questions that cover different content areas.

3.5. Line 327: could the authors please say more about how this localization is non-trivial? What is shows is that the topic model can identify which broad topic is addressed by each question (i.e., not which part of a lecture, but merely which lecture). Therefore, is this finding not fully predicted by Figure 1C? In what way does it provide new

information beyond Figure 1C? Moreover, if the authors simply constructed a Voronoi diagram where each point in space is assigned to its nearest question, would the result not be highly similar? (I am not saying that it is, but this is a much simpler method and, in a sense, a discretization of the authors' formula).

The "non-trivial" localization we're referring to is that even though our knowledge predictions at each coordinate on the knowledge maps are based solely on the locations of the *questions*, the increases in predicted knowledge we see from one quiz's map to the next (shown as changes in the maps' shading at each coordinate) are specific to the regions near the just-watched *lecture*'s trajectory. In other words, changes in the predictions we make about participants' knowledge based on this "map's" coordinate system track how we might intuitively expect them to change following exposure to particular new content, despite not considering that content's spatial location in computing them. This is related to the visualization in Figure 2C (we assume that the reviewer is referring to Fig. 2C, since Fig. 1 has no panel C), but the two figures are different in several important ways.

Figure 2C shows that the sets of questions we manually designed (and labeled) to be "about" each lecture tend to appear near their corresponding lecture when visualizing their first three principal components. This requires generally that the topic model capture similarities between content we expect to be conceptually related. The knowledge maps, by contrast, show more specifically that questions for which participants' abilities to answer correctly *changed* after viewing a particular lecture tend to be embedded nearby that lecture. This requires that both the topic model *and* the non-linear manifold learning algorithm we used to construct the knowledge maps' coordinate system (UMAP) preserve relationships that are meaningful to what individuals know and how they learn. It also requires that changes in knowledge are both sufficiently consistent and sufficiently specific across participants to produce the visual contrast we see in Figure 8.

We do agree that many of these patterns themselves are not necessarily specific to these maps, and can be observed elsewhere in our findings. For example, Figures 2 and 3 suggest that questions can be matched to their corresponding lectures, Figure 4 suggests that different questions will map onto different locations along those lectures' trajectories, and Figure 5 suggests that increases in knowledge from one quiz to the next will be lecture-specific and consistent across participants. However, our ability to automatically capture these patterns simultaneously in low-dimensional, intuitive, human-readable "maps" (e.g., which have been transformed to reflect Euclidean distance, and from which we can recover word clouds like those shown in Fig. 8C) is a non-trivial advance in and of itself.

The reviewer's idea of "filling in" these maps using Voronoi diagrams is an interesting one, and would certainly lead to neat visualizations! However, based on other results we report in our manuscript, we suspect this would be a less truthful representation of participants' (probable) knowledge at each individual coordinate on these maps. As described in our response to Reviewer 2's comment 2, we show in Figure 7 that knowledge tends to "fall off" gradually with respect to distance in our text embedding space, and in Figure 6 we show that participants' knowledge about a particular embedding-space coordinate can be predicted by a weighted combination of multiple coordinates for which their knowledge is "known." By contrast, a Voronoi diagram such as the reviewer describes

would cast success on each individual quiz question as the sole, static determinant of our knowledge predictions for all knowledge-map coordinates to which it is the single closest, while having zero influence on our predictions beyond that point. As such, we see our current approach (of computing accuracy-weighted sums of radial basis functions centered on each quiz question) as more directly motivated by our other findings (as well as more theoretically consistent with our method of estimating knowledge in the high-dimensional topic space) and have elected to retain it rather than opting for a more discrete alternative.

3.6. Choosing number of topics: did the authors examine the perplexity or coherence of the topics? These measures (especially coherence) appear to be accepted metrics for choosing the number of topics (see, e.g., https://aclanthology.org/D11-1024.Pdf, https://aclanthology.org/N10-1012.pdf).

Reviewer 1 (comment 6) asked about how we selected the number of topics as well. Here is the relevant part of our response to that comment:

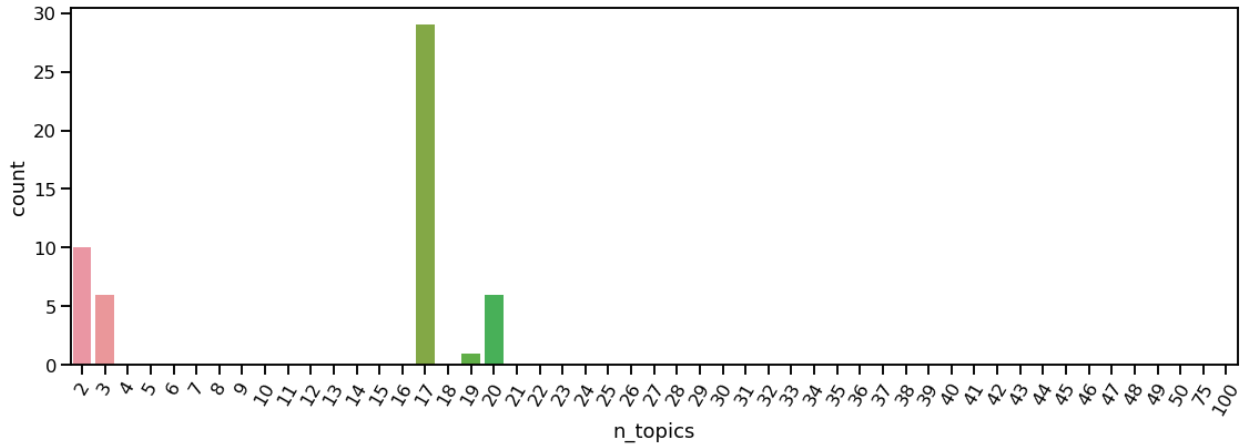
> **To select an appropriate number of topics (k) for our model, as a starting point, we identified the minimum number of topics that yielded at least one "unused" topic (i.e., in which all words in the vocabulary were assigned uniform weights) after training. This suggested that the number of topics might be sufficient to capture the set of latent themes present in the two lectures. We found this value to be k = 15 topics. We found that with a limited number of additional adjustments following Boyd-Graber et al. (2014; e.g., removing corpus-specific stop-words), the model yielded (subjectively) sensible and coherent topics.**

To elaborate, we tend to take the view proposed by Boyd-Graber et al. (2014) in their "Care and feeding of topic models" paper. Essentially, they suggest that using automated (optimization-based) approaches to selecting the number of topics can lead to poor performance. (This view is also reported in myriad other papers, many of which are cited in, or cite, the Boyd-Graber et al. paper.) Instead, they suggest that in practice, it tends to work well to manually evaluate whether a given number of topics is "appropriate" using three criteria:

1. **Are individual topics meaningful, interpretable, coherent, and useful?**
2. **Are assignments of topics to documents meaningful, appropriate, and useful?**
3. **Do topics facilitate better or more efficient document search, navigation, understanding, [or] browsing?**

Although these criteria do not perfectly align with our current goals and approach, we felt that the topics seemed well-behaved and useful (e.g., see Tab. S2 and Figs. 3, S1, and S2). Out of curiosity, and in response to the reviewer's comment, we also carried out a number of analyses using coherence to see how many topics were "optimal" by those measures. There are several ways to "compute" coherence and select a "best" number of topics, so we implemented seven different coherence measures proposed by several studies (Mimno et al., 2011; Cao et al., 2009; 4 measures from Röder et al., 2015; and Arun et al., 2010). For each measure, coherence is typically computed using some number n of the most heavily weighted on by each topic . We used n = 5, 10, 15, 20, …, 50, yielding a total of 70

approaches to selecting the optimal number of topics (k). Then we counted up the numbers of times each value of k was selected as the optimum:



We found that overall, according to these metrics, using $K = 17$ topics was most frequently selected as the optimal choice. The topic distributions using 17 topics also looked reasonable, although when we compared them manually to the 15 topic versions we felt they looked slightly less clean. (We did verify that our main results replicate using $K = 17$ topics). Ultimately we decided to stick with our prior "hand selected" approach.