

¹ Text embedding models yield high-resolution insights
² into conceptual knowledge from short multiple-choice
³ quizzes

⁴ Paxton C. Fitzpatrick¹, Andrew C. Heusser^{1, 2}, and Jeremy R. Manning^{1, *}

¹Department of Psychological and Brain Sciences

Dartmouth College, Hanover, NH 03755, USA

²Akili Interactive Labs

Boston, MA 02110, USA

*Corresponding author: Jeremy.R.Manning@Dartmouth.edu

⁵

Abstract

⁶

We develop a mathematical framework, based on natural language processing models, for tracking and characterizing the acquisition of conceptual knowledge. Our approach embeds each concept in a high-dimensional representation space, where nearby coordinates reflect similar or related concepts. We test our approach using behavioral data from participants who answered small sets of multiple-choice quiz questions interleaved between watching two course videos from the Khan Academy platform. We apply our framework to the videos' transcripts and the text of the quiz questions to quantify the content of each moment of video and each quiz question. We use these embeddings, along with participants' quiz responses, to track how the learners' knowledge changed after watching each video. Our findings show how a small set of quiz questions may be used to obtain rich and meaningful high-resolution insights into what each learner knows, and how their knowledge changes over time as they learn.

¹⁷

Keywords: education, learning, knowledge, concepts, natural language processing

¹⁸ **Introduction**

¹⁹ Suppose that a teacher had access to a complete, tangible “map” of everything a student knows.
²⁰ Defining what such a map might even look like, let alone how it might be constructed or filled in, is
²¹ itself a non-trivial problem. But if a teacher *were* to gain access to such a map, how might it change
²² their ability to teach that student? Perhaps they might start by checking how well the student
²³ knows the to-be-learned information already, or how much they know about related concepts.
²⁴ For some students, they could potentially optimize their teaching efforts to maximize efficiency
²⁵ by focusing primarily on not-yet-known content. For other students (or other content areas), it
²⁶ might be more effective to optimize for direct connections between already known content and
²⁷ new material. Observing how the student’s knowledge changed over time, in response to their
²⁸ teaching, could also help to guide the teacher towards the most effective strategy for that individual
²⁹ student.

³⁰ A common approach to assessing a student’s knowledge is to present them with a set of quiz
³¹ questions, calculate the proportion they answer correctly, and provide them with feedback in the
³² form of a simple numeric or letter grade. While such a grade can provide *some* indication of whether
³³ the student has mastered the to-be-learned material, any univariate measure of performance on a
³⁴ complex task sacrifices certain relevant information, risks conflating underlying factors, and so on.
³⁵ For example, consider the relative utility of the theoretical map described above that characterizes
³⁶ a student’s knowledge in detail, versus a single annotation saying that the student answered 85%
³⁷ of their quiz questions correctly, or that they received a ‘B’. Here, we show that the same quiz data
³⁸ required to compute proportion-correct scores or letter grades can instead be used to obtain far
³⁹ more detailed insights into what a student knew at the time they took the quiz.

⁴⁰ Designing and building procedures and tools for mapping out knowledge touches on deep
⁴¹ questions about what it means to learn. For example, how do we acquire conceptual knowledge?
⁴² Memorizing course lectures or textbook chapters by rote can lead to the superficial *appearance*
⁴³ of understanding the underlying content, but achieving true conceptual understanding seems to
⁴⁴ require something deeper and richer. Does conceptual understanding entail connecting newly

45 acquired information to the scaffolding of one’s existing knowledge or experience [5, 10, 12, 13, 26,
46 57]? Or weaving a lecture’s atomic elements (e.g., its component words) into a structured network
47 that describes how those individual elements are related [36, 61]? Conceptual understanding
48 could also involve building a mental model that transcends the meanings of those individual
49 atomic elements by reflecting the deeper meaning underlying the gestalt whole [33, 37, 54, 60].

50 The difference between “understanding” and “memorizing,” as framed by researchers in ed-
51 ucation, cognitive psychology, and cognitive neuroscience [e.g., 21, 24, 29, 37, 54], has profound
52 analogs in the fields of natural language processing and natural language understanding. For
53 example, considering the raw contents of a document (e.g., its constituent symbols, letters, and
54 words) might provide some clues as to what the document is about, just as memorizing a pas-
55 sage might provide some ability to answer simple questions about it. However, text embedding
56 models [e.g., 6, 7, 9, 11, 14, 35, 44, 62] also attempt to capture the deeper meaning *underlying* those
57 atomic elements. These models consider not only the co-occurrences of those elements within and
58 across documents, but (in many cases) also patterns in how those elements appear across different
59 scales (e.g., sentences, paragraphs, chapters, etc.), the temporal and grammatical properties of the
60 elements, and other high-level characteristics of how they are used [38?]. To be clear, this is not
61 to say that text embedding models themselves are capable of “understanding” deep conceptual
62 meaning in any traditional sense. But rather, their ability to capture the underlying *structure* of
63 text documents beyond their surface-level contents provides a computational framework through
64 which those document’s deeper conceptual meaning may be quantified, explored, and understood.
65 According to these models, the deep conceptual meaning of a document may be captured by a
66 feature vector in a high-dimensional representation space, wherein nearby vectors reflect concep-
67 tually related documents. A model that succeeds at capturing an analogue of “understanding” is
68 able to assign nearby feature vectors to two conceptually related documents, *even when the specific*
69 *words contained in those documents have limited overlap*. In this way, “concepts” are defined implicitly
70 by the model’s geometry [e.g., how the embedding coordinate of a given word or document relates
71 to the coordinates of other text embeddings; 49].

72 Given these insights, what form might a representation of the sum total of a person’s knowledge

73 take? First, we might require a means of systematically describing or representing (at least some
74 subset of) the nearly infinite set of possible things a person could know. Second, we might want to
75 account for potential associations between different concepts. For example, the concepts of “fish”
76 and “water” might be associated in the sense that fish live in water. Third, knowledge may have
77 a critical dependency structure, such that knowing about a particular concept might require first
78 knowing about a set of other concepts. For example, understanding the concept of a fish swimming
79 in water first requires understanding what fish and water *are*. Fourth, as we learn, our “current
80 state of knowledge” should change accordingly. Learning new concepts should both update our
81 characterizations of “what is known” and also unlock any now-satisfied dependencies of those
82 newly learned concepts so that they are “tagged” as available for future learning.

83 Here we develop a framework for modeling how conceptual knowledge is acquired during
84 learning. The central idea behind our framework is to use text embedding models to define the
85 coordinate systems of two maps: a *knowledge map* that describes the extent to which each concept is
86 currently known, and a *learning map* that describes changes in knowledge over time. Each location
87 on these maps represents a single concept, and the maps’ geometries are defined such that related
88 concepts are located nearby in space. We use this framework to analyze and interpret behavioral
89 data collected from an experiment that had participants answer sets of multiple-choice questions
90 about a series of recorded course lectures.

91 Our primary research goal is to advance our understanding of what it means to acquire deep,
92 real-world conceptual knowledge. Traditional laboratory approaches to studying learning and
93 memory (e.g., list-learning studies) often draw little distinction between memorization and under-
94 standing. Instead, these studies typically focus on whether information is effectively encoded or
95 retrieved, rather than whether the information is *understood*. Approaches to studying conceptual
96 learning, such as category learning experiments, can begin to investigate the distinction between
97 memorization and understanding, often by training participants to distinguish arbitrary or random
98 features in otherwise meaningless categorized stimuli [1, 18, 19, 22, 27, 52]. However the objective
99 of real-world training, or learning from life experiences more generally, is often to develop new
100 knowledge that may be applied in *useful* ways in the future. In this sense, the gap between modern

learning theories and modern pedagogical approaches that inform classroom learning strategies is enormous: most of our theories about *how* people learn are inspired by experimental paradigms and models that have only peripheral relevance to the kinds of learning that students and teachers actually seek [24, 37]. To help bridge this gap, our study uses course materials from real online courses to inform, fit, and test models of real-world conceptual learning. We also provide a demonstration of how our models can be used to construct “maps” of what students know, and how their knowledge changes with training. In addition to helping to visually capture knowledge (and changes in knowledge), we hope that such maps might lead to real-world tools for improving how we educate. Taken together, our work shows that existing course materials and evaluative tools like short multiple-choice quizzes may be leveraged to gain highly detailed insights into what students know and how they learn.

Results

At its core, our main modeling approach is based around a simple assumption that we sought to test empirically: all else being equal, knowledge about a given concept is predictive of knowledge about similar or related concepts. From a geometric perspective, this assumption implies that knowledge is fundamentally “smooth.” In other words, as one moves through a space representing an individual’s knowledge (where similar concepts occupy nearby coordinates), their “level of knowledge” should change relatively gradually. To begin to test this smoothness assumption, we sought to track participants’ knowledge and how it changed over time in response to training. Two overarching goals guide our approach. First, we want to gain detailed insights into what learners know at different points in their training. For example, rather than simply reporting on the proportions of questions participants answer correctly (i.e., their overall performance), we seek estimates of their knowledge about a variety of specific concepts. Second, we want our approach to be potentially scalable to large numbers of diverse concepts, courses, and students. This requires that the conceptual content of interest be discovered *automatically*, rather than relying on manually produced ratings or labels.



Figure 1: Experimental paradigm. Participants alternate between completing three 13-question multiple-choice quizzes and watching two Khan Academy lectures. Each quiz contains a mix of 5 questions about Lecture 1, 5 questions about Lecture 2, and 3 questions about general physics knowledge. The specific questions reflected on each quiz, and the orders of each quiz's questions, were randomized across participants.

127 We asked participants in our study to complete brief multiple-choice quizzes before, between,
 128 and after watching two lecture videos from the Khan Academy [32] platform (Fig. 1). The first
 129 lecture video, entitled *Four Fundamental Forces*, discussed the four fundamental forces in physics:
 130 gravity, strong and weak interactions, and electromagnetism. The second, entitled *Birth of Stars*,
 131 provided an overview of our current understanding of how stars form. We selected these particular
 132 lectures to satisfy three general criteria. First, we wanted both lectures to be accessible to a broad
 133 audience (i.e., with minimal prerequisite knowledge) so as to limit the impact of prior training
 134 on participants' abilities to learn from the lectures. To this end, we selected two introductory
 135 videos that were intended to be viewed at the start of students' training in their respective content
 136 areas. Second, we wanted the two lectures to have some related content, so that we could test
 137 our approach's ability to distinguish similar conceptual content. To this end, we chose two videos
 138 from the same Khan Academy course domain, "Cosmology and Astronomy." Third, we sought to
 139 minimize dependencies and specific overlap between the videos. For example, we did not want
 140 participants' abilities to understand one video to (directly) influence their abilities to understand the
 141 other. To satisfy this last criterion, we chose videos from two different lecture series (Lectures 1 and
 142 2 were from the "Scale of the Universe" and "Stars, Black Holes, and Galaxies" series, respectively).

143 We also wrote a set of multiple-choice quiz questions that we hoped would enable us to
 144 evaluate participants' knowledge about each individual lecture, along with related knowledge



Figure 2: Modeling course content. **A. Building a document pool from sliding windows of text.** We decompose each lecture’s transcript into a series of overlapping sliding windows. The full set of transcript snippets (across all windows) may be treated as a set of “documents” for training a text embedding model. **B. Constructing lecture content trajectories.** After training the model on the sliding windows from both lectures, we transform each lecture into a “trajectory” through text embedding space by joining the embedding coordinates of successive sliding windows parsed from its transcript. **C. Embedding multiple lectures and questions in a shared space.** We apply the same model (trained on the two lectures’ windows) to both lectures, along with the text of each question in our pool (Supp. Tab. 1), to project them into a shared text embedding space. This results in one trajectory per lecture and one coordinate for each question. Here, we have projected the 15-dimensional embeddings onto their first 3 principal components for visualization.

145 about physics concepts not specifically presented in either video (see Supp. Tab. 1 for the full list
 146 of questions in our stimulus pool). Participants answered questions randomly drawn from each
 147 content area (Lecture 1, Lecture 2, and general physics knowledge) on each of the three quizzes.
 148 Quiz 1 was intended to assess participants’ “baseline” knowledge before training, Quiz 2 assessed
 149 knowledge after watching the *Four Fundamental Forces* video (i.e., Lecture 1), and Quiz 3 assessed
 150 knowledge after watching the *Birth of Stars* video (i.e., Lecture 2).

151 To study in detail how participants’ conceptual knowledge changed over the course of the
 152 experiment, we first sought to model the conceptual content presented to them at each moment
 153 throughout each of the two lectures. We adapted an approach we developed in prior work [25]
 154 to identify the latent themes in the lectures using a topic model [7]. Briefly, topic models take
 155 as input a collection of text documents, and learn a set of “topics” (i.e., latent themes) from their
 156 contents. Once fit, a topic model can be used to transform arbitrary (potentially new) documents
 157 into sets of “topic proportions,” describing the weighted blend of learned topics reflected in their
 158 texts. We parsed automatically generated transcripts of the two lectures into overlapping sliding
 159 windows, where each window contained the text of the lecture transcript from a particular time

span. We treated the set of text snippets (across all of these windows) as documents to fit the model (Fig. 2A; see *Constructing text embeddings of multiple lectures and questions*). Transforming the text from every sliding window with the model yielded a number-of-windows by number-of-topics (15) topic-proportions matrix describing the unique mixture of broad themes from both lectures reflected in each window’s text. Each window’s “topic vector” (i.e., column of the topic-proportions matrix) is analogous to a coordinate in a 15-dimensional space whose axes are topics discovered by the model. Within this space, each lecture’s sequence of topic vectors (i.e., corresponding to its transcript’s overlapping text snippets across sliding windows) forms a *trajectory* that captures how its conceptual content unfolds over time (Fig. 2B). We resampled these trajectories to a resolution of one topic vector for each second of video (i.e., 1 Hz).

We hypothesized that a topic model trained on transcripts of the two lectures should also capture the conceptual knowledge probed by each quiz question. If indeed the topic model could capture information about the deeper conceptual content of the lectures (i.e., beyond surface-level details such as particular word choices), then we should be able to recover a correspondence between each lecture and questions *about* each lecture. Importantly, such a correspondence could not solely arise from superficial text matching between lecture transcripts and questions, since the lectures and questions often used different words (Supp. Fig. 5) and phrasings. Simply comparing the average topic weights from each lecture and question set (averaging across time and questions, respectively) reveals a striking correspondence (Supp. Fig. 2). Specifically, the average topic weights from Lecture 1 are strongly correlated with the average topic weights from Lecture 1 questions ($r(13) = 0.809$, $p < 0.001$, 95% confidence interval (CI) = [0.633, 0.962]), and the average topic weights from Lecture 2 are strongly correlated with the average topic weights from Lecture 2 questions ($r(13) = 0.728$, $p = 0.002$, 95% CI = [0.456, 0.920]). At the same time, the average topic weights from the two lectures are *negatively* correlated with the average topic weights from their non-matching question sets (Lecture 1 video vs. Lecture 2 questions: $r(13) = -0.547$, $p = 0.035$, 95% CI = [-0.812, -0.231]; Lecture 2 video vs. Lecture 1 questions: $r(13) = -0.612$, $p = 0.015$, 95% CI = [-0.874, -0.281]), indicating that the topic model also exhibits some degree of specificity. The full set of pairwise comparisons between average topic weights for the lectures and question sets



Figure 3: Lecture and question topic overlap. A. Topic weight variability. The bar plots display the variance of each topic's weight across lecture timepoints (top row) and questions (bottom row); colors denote topics. The top-weighted words from the most “expressive” (i.e., variable across observations) topic from each lecture are displayed in the upper right (orange: topic 2; yellow-green: topic 4). The top-weighted words from the full set of topics may be found in Supplementary Table 2. **B. Relationships between topic weight variability.** Pairwise correlations between the distributions of topic weight variance for each lecture and question set. Each row and column corresponds to a bar plot in Panel A.

188 is reported in Supplementary Figure 2.

189 Another, more sensitive, way of summarizing the conceptual content of the lectures and questions is to look at *variability* in how topics are weighted over time and across different questions 190 (Fig. 3). Intuitively, the variability in the expression of a given topic relates to how much “information” [20] the lecture (or question set) reflects about that topic. For example, suppose a given 191 topic is weighted on heavily throughout a lecture. That topic might be characteristic of some 192 aspect or property of the lecture *overall* (conceptual or otherwise), but unless the topic’s weights 193 changed in meaningful ways over time, the topic would be a poor indicator of any *specific* conceptual 194 content in the lecture. We therefore also compared the variances in topic weights (across time 195 or questions) between the lectures and questions. The variability in topic expression (over time 196 or questions) was similar for the Lecture 1 video and questions ($r(13) = 0.824, p < 0.001$, 197 95% CI = [0.696, 0.973]) and the Lecture 2 video and questions ($r(13) = 0.801, p < 0.001$, 95% 198 CI = [0.539, 0.958]). Simultaneously, as reported in Figure 3B, the variabilities in topic expression 199 across *different* videos and lecture-specific questions (i.e., Lecture 1 video vs. Lecture 2 questions; 200 201

202 Lecture 2 video vs. Lecture 1 questions) were negatively correlated, and neither video’s topic
203 variability was reliably correlated with the topic variability across general physics knowledge
204 questions. Taken together, the analyses reported in Figure 3 and Supplementary Figure 2 indicate
205 that a topic model fit to the videos’ transcripts can also reveal correspondences (at a coarse scale)
206 between the lectures and questions.

207 While an individual lecture may be organized around a single broad theme at a coarse scale,
208 at a finer scale, each moment of a lecture typically covers a narrower range of content. Given the
209 correspondence we found between the variabilities in topic expression across moments of each
210 lecture and questions from its corresponding set (Fig. 3), we wondered whether the text embedding
211 model might additionally capture these conceptual relationships at a finer scale. For example, if a
212 particular question asks about the content from one small part of a lecture, we wondered whether
213 the text embeddings could be used to automatically identify the “matching” moment(s) in the
214 lecture. To explore this, we computed the correlation between each question’s topic weights
215 and the topic weights for each second of its corresponding lecture, and found that each question
216 appeared to be temporally specific (Fig. 4). In particular, most questions’ topic vectors were
217 maximally correlated with a well-defined (and relatively narrow) range of timepoints from their
218 corresponding lectures, and the correlations fell off sharply outside of that range (Supp. Figs. 3, 4).
219 We also qualitatively examined the best-matching intervals for each question by comparing the
220 question’s text to the transcribed text from the most-correlated parts of the lectures (Supp. Tab. 3).
221 Despite that the questions were excluded from the text embedding model’s training set, in general
222 we found (through manual inspection) a close correspondence between the conceptual content
223 that each question probed and the content covered by the best-matching moments of the lectures.
224 Two representative examples are shown at the bottom of Figure 4.

225 The ability to quantify how much each question is “asking about” the content from each moment
226 of the lectures could enable high-resolution insights into participants’ knowledge. Traditional
227 approaches to estimating how much a student “knows” about the content of a given lecture entail
228 administering some form of assessment (e.g., a quiz) and computing the proportion of correctly
229 answered questions. But if two students receive identical scores on such an exam, might our



Figure 4: Which parts of each lecture are captured by each question? Each panel displays time series plots showing how each question’s topic vector correlates with each video timepoint’s topic vector (Panel A.: correlations for the *Four Fundamental Forces* lecture and associated questions; Panel B.: correlations for the *Birth of Stars* lecture and associated questions). The colors denote question identities. The diamonds in each panel denote the moment of peak correlation between the indicated question and the lecture trajectory. The associated questions’ text and snippets of the lectures’ transcripts from the surrounding 30 seconds, are displayed at the bottom of the figure.

modeling framework help us to gain more nuanced insights into the *specific* content that each student has mastered (or failed to master)? For example, a student who misses three questions that were all about the same concept (e.g., concept *A*) will have gotten the same *proportion* of questions correct as another student who missed three questions about three *different* concepts (e.g., *A*, *B*, and *C*). But if we wanted to help these two students fill in the “gaps” in their understandings, we might do well to focus specifically on concept *A* for the first student, but to also add in materials pertaining to concepts *B* and *C* for the second student. In other words, raw “proportion-correct” measures may capture *how much* a student knows, but not *what* they know. We wondered whether our modeling framework might enable us to (formally and automatically) infer participants’ knowledge at the scale of individual concepts (e.g., as captured by a single moment of a lecture).

We developed a simple formula (Eqn. 1) for using a participant’s responses to a small set of multiple-choice questions to estimate how much the participant “knows” about the concept reflected by any arbitrary coordinate x in text embedding space (e.g., the content reflected by

any moment in a lecture they had watched; see *Estimating dynamic knowledge traces*). Essentially, the estimated knowledge at coordinate x is given by the weighted proportion of quiz questions the participant answered correctly, where the weights reflect how much each question is “about” the content at x . When we apply this approach to estimate the participant’s knowledge about the content presented in each moment of each lecture, we can obtain a detailed time course describing how much “knowledge” that participant has about the content presented at any part of the lecture. As shown in Figure 5A and C, we can apply this approach separately for the questions from each quiz participants took throughout the experiment. From just a few questions per quiz (see *Estimating dynamic knowledge traces*), we obtain a high-resolution snapshot (at the time each quiz was taken) of what the participants knew about any moment’s content, from either of the two lectures they watched (comprising a total of 1,100 samples across the two lectures).

While the time courses in Figure 5A and C provide detailed *estimates* about participants’ knowledge, these estimates are of course only *useful* to the extent that they accurately reflect what participants actually know. As one sanity check, we anticipated that the knowledge estimates should reflect a content-specific “boost” in participants’ knowledge after watching each lecture. In other words, if participants learn about each lecture’s content upon watching it, the knowledge estimates should capture that. After watching the *Four Fundamental Forces* lecture, participants should exhibit more knowledge for the content of that lecture than they had before, and that knowledge should persist for the remainder of the experiment. Specifically, knowledge about that lecture’s content should be relatively low when estimated using Quiz 1 responses, but should increase when estimated using Quiz 2 or 3 responses (Fig. 5B). Indeed, we found that participants’ estimated knowledge about the content of *Four Fundamental Forces* was substantially higher on Quiz 2 versus Quiz 1 ($t(49) = 8.764, p < 0.001$) and on Quiz 3 versus Quiz 1 ($t(49) = 10.519, p < 0.001$). We found no reliable differences in estimated knowledge about that lecture’s content on Quiz 2 versus 3 ($t(49) = 0.160, p = 0.874$). Similarly, we hypothesized (and subsequently confirmed) that participants should show greater estimated knowledge about the content of the *Birth of Stars* lecture after (versus before) watching it (Fig. 5D). Specifically, since participants watched that lecture after taking Quiz 2 (but before Quiz 3), we hypothesized that their knowledge estimates



Figure 5: Estimating knowledge about the content presented at each moment of each lecture. **A. Knowledge about the time-varying content of *Four Fundamental Forces*.** Each trace displays the weighted proportion of correctly answered questions about the content reflected in each moment of the lecture (see *Estimating dynamic knowledge traces*), using responses from a single quiz (color). The traces are averaged across participants. **B. Average estimated knowledge about *Four Fundamental Forces*.** Each bar displays the across-timepoint average knowledge, estimated using the responses to one quiz's questions. **C. Knowledge about the time-varying content of *Birth of Stars*.** The panel is in the same format as Panel A, but here the knowledge estimates are for the moment-by-moment content of the *Birth of Stars* lecture. **D. Average estimated knowledge about *Birth of Stars*.** The panel is in the same format as Panel B, but here the knowledge estimates are for the content of the *Birth of Stars* lecture. All panels: error ribbons and error bars denote 95% confidence intervals, estimated across participants.

271 should be relatively low on Quizzes 1 and 2, but should show a “boost” on Quiz 3. Consistent
272 with this prediction, we found no reliable differences in estimated knowledge about the *Birth of*
273 *Stars* lecture content on Quizzes 1 versus 2 ($t(49) = 1.013, p = 0.316$), but the estimated knowl-
274 edge was substantially higher on Quiz 3 versus 2 ($t(49) = 10.561, p < 0.001$) and Quiz 3 versus 1
275 ($t(49) = 8.969, p < 0.001$).

276 If we are able to accurately estimate a participant’s knowledge about the content tested by
277 a given question, our estimates of their knowledge should carry some predictive information
278 about whether they are likely to answer that question correctly or incorrectly. We developed a
279 statistical approach to test this claim. For each quiz question a participant answered, in turn,
280 we used Equation 1 to estimate their knowledge at the given question’s embedding space co-
281 ordinate based on other questions that participant answered on the same quiz. We repeated
282 this for all participants, and for each of the three quizzes. Then, separately for each quiz, we
283 fit a generalized linear mixed model (GLMM) with a logistic link function to explain the like-
284 lihood of correctly answering a question as a function of estimated knowledge for its embed-
285 ding coordinate, while accounting for random variation among participants and questions (see
286 *GLMM METHODS SECTION PLACEHOLDER*). To assess the predictive value of the knowledge
287 estimates, we compared each GLMMs to an analogous (i.e., nested) “null” model that did not
288 consider estimated knowledge using parametric bootstrap likelihood-ratio tests.

289 We carried out three different versions of the analyses described above, wherein we considered
290 different sources of information in our estimates of participants’ knowledge for each quiz question.
291 First, we estimated knowledge at each question’s embedding coordinate using *all* other questions
292 answered by the same participant on the same quiz (“All questions”; Fig. 6, top row). This test was
293 intended to assess the overall predictive power of our approach. Second, we estimated knowledge
294 for each question about a given lecture using only the other questions (from the same participant
295 and quiz) about that *same* lecture (“Within-lecture”; Fig. 6, middle rows). This test was intended to
296 assess the *specificity* of our approach by asking whether our predictions could distinguish between
297 questions about different content covered by the same lecture. Third, we estimated knowledge
298 for each question about one lecture using only questions (from the same participant and quiz)

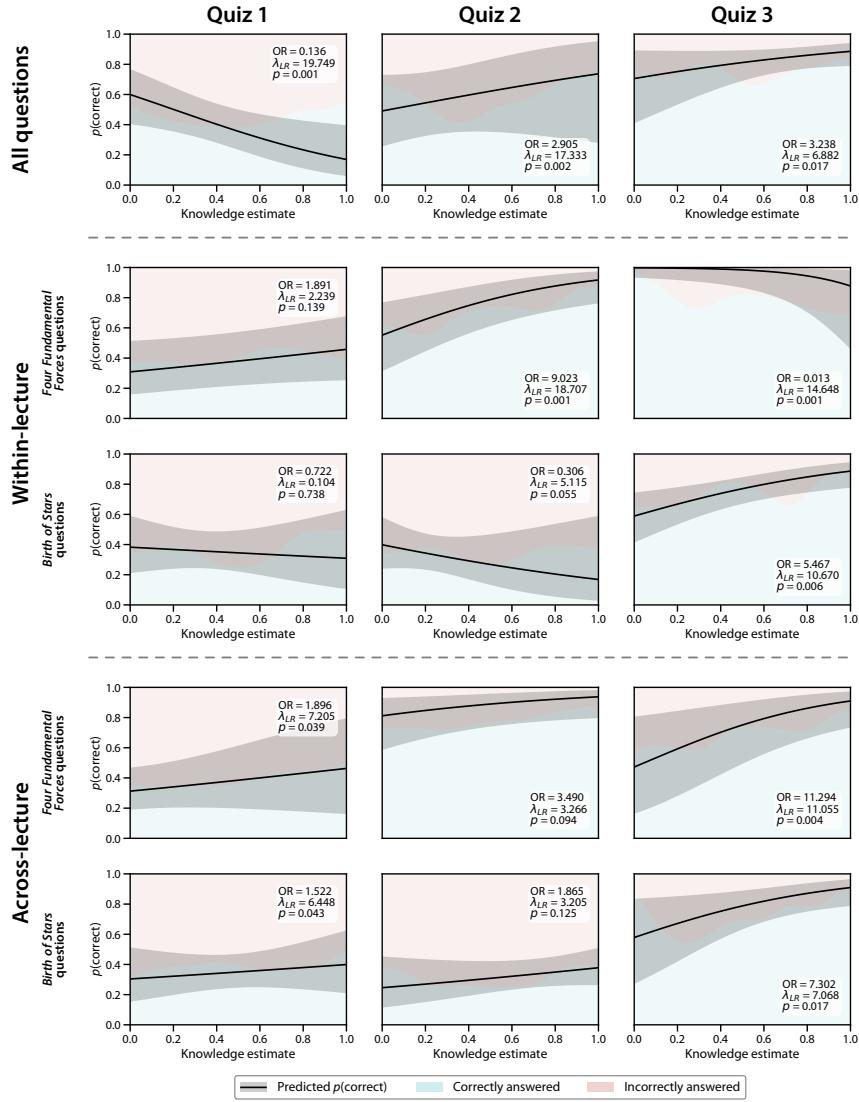


Figure 6: Predicting success on held-out questions using estimated knowledge. We used generalized linear mixed models (GLMMs) to model the likelihood of correctly answering a quiz question as a function of estimated knowledge for its embedding coordinate (see *GLMM METHODS SECTION PLACEHOLDER*). Separately for each quiz (column), we examined this relationship based on three different sets of knowledge estimates: knowledge for each question based on all other questions the same participant answered on the same quiz (“All questions”; top row), knowledge for each question about one lecture based on all other questions (from the same participant and quiz) about the *same* lecture (“Within-lecture”; middle rows), and knowledge for each question about one lecture based on all questions (from the same participant and quiz) about the *other* lecture (“Across-lecture”; bottom rows). The background in each panel displays the kernel density estimates of the relative observed proportions of correctly (blue) versus incorrectly (red) answered questions, for each level of estimated knowledge along the x -axis. The black curves display the (population-level) GLMM-predicted probabilities of correctly answering a question as a function of estimated knowledge. Error ribbons denote the 95% confidence intervals.

299 about the *other* lecture (“Across-lecture”; Fig. 6, bottom rows). This test was intended to assess the
300 *generalizability* of our approach by asking whether our predictions held across the content areas of
301 the two lectures.

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

320 When we fit a GLMM to estimates of participants’ knowledge for each Quiz 1 question based
321 on all other Quiz 1 questions, we observed an outcome consistent with our null hypothesis: higher
322 estimated knowledge at the embedding coordinate of a held-out question was associated with a
323 lower likelihood of answering the question correctly (odds ratio (OR) = 0.136, likelihood-ratio test
324 statistic (λ_{LR}) = 19.749, 95% CI = [14.352, 26.545], $p = 0.001$). This outcome suggests that our
325 knowledge estimates do *not* provide useful information about participants’ Quiz 1 performance
326 when we aggregated across all question content areas. We speculated that this might either

327 indicate that the knowledge estimates are uninformative in general, or about Quiz 1 performance
328 in particular. When we repeated this analysis for Quizzes 2 and 3, the direction of this relationship
329 reversed: higher estimated knowledge for a given question predicted a greater likelihood of
330 answering it correctly (Quiz 2: $OR = 2.905$, $\lambda_{LR} = 17.333$, 95% CI = [14.966, 29.309], $p = 0.002$;
331 Quiz 3: $OR = 3.238$, $\lambda_{LR} = 6.882$, 95% CI = [6.228, 8.184], $p = 0.017$). Taken together, these
332 results suggest that our knowledge estimates *do* reliably predict participants' likelihood of success
333 on individual quiz questions, but only when there is some "contrast" in performance between
334 the content of different questions. On Quiz 1, before participants have learned the content of
335 either lecture, their overall performance is poor (average proportion correct: 0.46), and the specific
336 questions they answer correctly do not seem to be structured in a meaningful way with respect to
337 the embedding space (e.g., Fig. 8A, left panel). On Quizzes 2 and 3, after participants have begun
338 their training, their overall performance improves (average proportion correct for Quiz 2: 0.60;
339 average proportion correct for Quiz 3: 0.79), and the specific content areas they improve on are
340 captured by the lecture and question embeddings (and, therefore, our knowledge estimates).

341 We observed a similar pattern of results when we fit GLMMs to estimates of participants'
342 knowledge for each question about one lecture derived from other questions about the *same*
343 lecture. Specifically, for questions that participants answered on Quiz 1, prior to watching either
344 lecture, knowledge for the embedding coordinates of *Four Fundamental Forces*-related questions
345 estimated using other *Four Fundamental Forces*-related questions did not reliably predict whether
346 those questions were answered correctly ($OR = 1.891$, $\lambda_{LR} = 2.293$, 95% CI = [2.091, 2.622], $p =$
347 0.139). The same was true of knowledge estimates for *Birth of Stars*-related questions based on other
348 *Birth of Stars*-related questions ($OR = 0.722$, $\lambda_{LR} = 5.115$, 95% CI = [0.094, 0.146], $p = 0.738$). When
349 we recomputed these within-lecture knowledge estimates using questions from Quiz 2—which
350 participants took immediately after viewing *Four Fundamental Forces* but prior to viewing *Birth*
351 of *Stars*—we found that they now reliably predicted success on *Four Fundamental Forces*-related
352 questions ($OR = 9.023$, $\lambda_{LR} = 18.707$, 95% CI = [10.877, 22.222], $p = 0.001$) but not on *Birth of Stars*-
353 related questions ($OR = 0.306$, $\lambda_{LR} = 5.115$, 95% CI = [4.624, 5.655], $p = 0.055$). Using participants'
354 responses from Quiz 3 (taken immediately after viewing *Birth of Stars*), we found that within-lecture

knowledge estimates for *Birth of Stars*-related questions could now reliably predict success on those questions ($OR = 5.467$, $\lambda_{LR} = 10.670$, 95% CI = [7.998, 12.532], $p = 0.006$). In contrast, within-lecture knowledge estimates for *Four Fundamental Forces* questions answered on Quiz 3 were no longer directly related to the likelihood of successfully answering them and instead exhibited the inverse relationship we would expect to arise from unstructured knowledge (with respect to the embedding space; $OR = 0.013$, $\lambda_{LR} = 14.648$, 95% CI = [10.695, 23.096], $p = 0.001$). In principle, this “prediction failure” might arise from several possible scenarios. One possibility is that the participants are answering nearly all of the questions about *Four Fundamental Forces* correctly on Quiz 3, and (as with the Quiz 1 questions) there might be insufficient contrast in performance across different questions to reliably estimate participants’ knowledge. However, when we examined participants’ performance on Quiz 3 questions about *Four Fundamental Forces*, we found that they answered these questions correctly only 76% of the time, on average (i.e., unlikely to reflect a “ceiling” effect). Alternatively, we speculated that participants might be forgetting some of the material from the *Four Fundamental Forces* lecture, and that this forgetting might be happening in a relatively “random” way with respect to spatial distance within the embedding space. This might lead our approach to over-estimate knowledge for held-out questions about “forgotten” knowledge that participants answered incorrectly. Indeed, participants’ performance on *Four Fundamental Forces*-related questions was slightly (thought not dramatically) lower on Quiz 3 than on Quiz 2 (mean proportion correct on Quiz 2: 0.77). Taken together, these results suggest that our approach can distinguish between questions about different content covered by a single lecture when participants have sufficiently structured knowledge about its contents, though this specificity may decrease further in time from when the lecture in question was viewed.

Finally, when we fit GLMMs to estimates of participants’ knowledge for questions about one lecture using questions they answered (on the same quiz) about the *other* lecture, we also observed a (largely) similar pattern. On Quiz 1, we found that participants’ abilities to correctly answer questions about *Four Fundamental Forces* could be predicted from their responses to questions about *Birth of Stars* ($OR = 1.896$, $\lambda_{LR} = 7.205$, 95% CI = [6.224, 7.524], $p = 0.039$) and similarly, that their ability to correctly answer *Birth of Stars*-related questions could be predicted from their responses

383 to *Four Fundamental Forces*-related questions ($OR = 1.522, \lambda_{LR} = 6.448, 95\% CI = [5.656, 6.843], p = 0.043$). This reflects coarse-scale structure in participants' knowledge prior to any training in our experiment. When we repeated this analysis using questions from Quiz 2, we found participants' responses to *Four Fundamental Forces*-related questions did *not* reliably predict their success on *Birth of Stars*-related questions ($OR = 1.865, \lambda_{LR} = 3.205, 95\% CI = [3.027, 3.600], p = 0.125$), nor did their responses to *Birth of Stars*-related questions reliably predict their success on *Four Fundamental Forces*-related questions ($OR = 3.490, \lambda_{LR} = 3.266, 95\% CI = [3.033, 3.866], p = 0.094$). These "prediction failures" appear to come from the fact that any signal derived from participants' knowledge about the content of the *Birth of Stars* lecture (prior to watching it) is swamped by the much more dramatic increase in their knowledge about the content of the *Four Fundamental Forces* (which they watched just prior to taking Quiz 2). This is reflected in their Quiz 2 performance for questions about each lecture (mean proportion correct for *Four Fundamental Forces*-related questions on Quiz 2: 0.77; mean proportion correct for *Birth of Stars*-related questions on Quiz 2: 0.36). When we again carried out these across-lecture knowledge predictions using questions from Quiz 3 (when participants had now viewed *both* lectures), we could again reliably predict success on questions about both *Four Fundamental Forces* ($OR = 11.294, \lambda_{LR} = 11.055, 95\% CI = [9.126, 18.476], p = 0.004$) and *Birth of Stars* ($OR = 7.302, \lambda_{LR} = 7.068, 95\% CI = [6.490, 8.584], p = 0.017$) using responses to questions about the other lecture's content. Across all three versions of these analyses, our results suggest that (by and large) our knowledge estimations can reliably predict participants' abilities to answer individual quiz questions, distinguish between questions about similar content, and generalize across content areas, provided that participants' quiz responses reflect a minimum level of "real" knowledge about both content on which these predictions are based and that for which they are made. Our results also indicate some potential limitations of our approach: when the contrast in participants' knowledge within the embedding space is low overall, our approach (often) cannot reliably predict their abilities to answer individual quiz questions.

408 That the knowledge predictions derived from the text embedding space reliably distinguish
409 between held-out correctly versus incorrectly answered questions (Fig. 6) suggests that spatial
410 relationships within this space can help explain what participants know. But how far does this

411 explanatory power extend? For example, suppose we know that a participant correctly answered a
412 question at embedding coordinate x . As we move farther away from x in the embedding space, how
413 does the likelihood that the participant knows about the content at a given location “fall off” with
414 distance? Conversely, suppose the participant instead answered that same question *incorrectly*.
415 Again, as we move farther away from x in the embedding space, how does the likelihood that the
416 participant does *not* know about a coordinate’s content change with distance? We reasoned that,
417 assuming our embedding space is capturing something about how individuals actually organize
418 their knowledge, a participant’s ability to answer questions embedded very close to x should
419 tend to be similar to their ability to answer the question embedded *at* x . Whereas at another
420 extreme, once we reach some sufficiently large distance from x , our ability to infer whether or
421 not a participant will correctly answer a question based on their ability to answer the question
422 at x should be no better than guessing based on their *overall* proportion of correctly answered
423 questions. In other words, beyond the maximum distance at which the participant’s ability to
424 answer the question at x is informative of their ability to answer a second question at location y ,
425 then guessing the outcome at y based on x should be no more successful than guessing based on a
426 measure that does not consider embedding space distance.

427 With these ideas in mind, we asked: conditioned on answering a question correctly, what
428 proportion of all questions (within some radius, r , of that question’s embedding coordinate)
429 were answered correctly? We plotted this proportion as a function of r . Similarly, we could
430 ask, conditioned on answering a question incorrectly, how the proportion of correct responses
431 changed with r . As shown in Figure 7, we found that quiz performance falls off smoothly with
432 distance, and the “rate” of the falloff does not appear to change across the different quizzes, as
433 measured by the distance at which performance becomes statistically indistinguishable from a
434 simple proportion correct score (see *Estimating the “smoothness” of knowledge*). This suggests that,
435 at least within the region of text embedding space covered by the questions our participants
436 answered (and as characterized using our topic model), the rate at which knowledge changes
437 with distance is relatively constant, even as participants’ overall level of knowledge varies across
438 quizzes or regions of the embedding space.

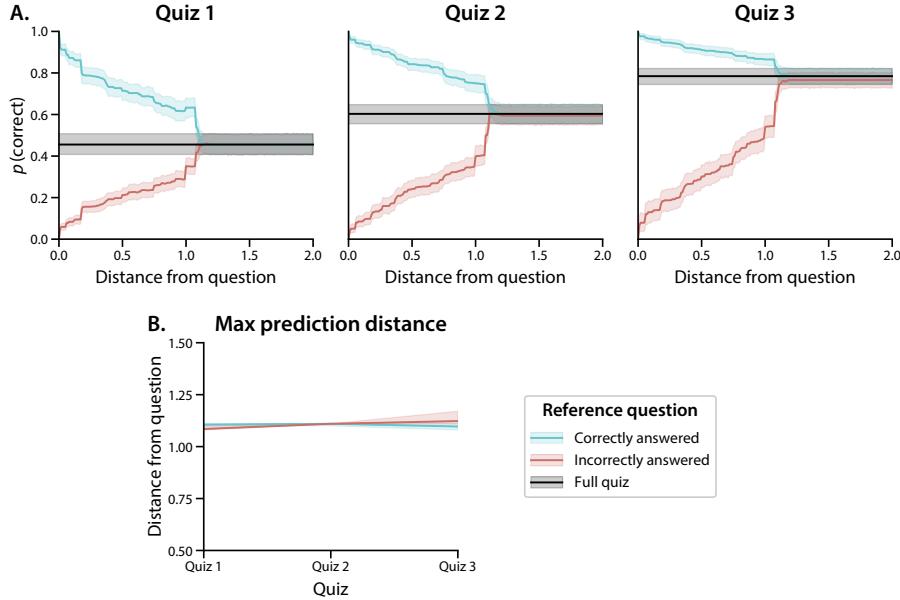


Figure 7: Knowledge falls off gradually in text embedding space. A. Performance versus distance. For each participant, for each correctly answered question (blue) or incorrectly answered question (red), we computed the proportion of correctly answered questions within a given distance of that question’s embedding coordinate. We used these proportions as a proxy for participants’ knowledge about the content within that region of the embedding space. We repeated this analysis for all questions and participants, and separately for each quiz (column). The black lines denote the average proportion correct across *all* questions included in the analysis at the given distance. **B. Maximum distance for which performance is reliably different from the average.** We used a bootstrap procedure (see *Estimating the “smoothness” of knowledge*) to estimate the point at which the blue and red lines in Panel A reliably diverged from the black line. We repeated this analysis separately for correctly and incorrectly answered questions from each quiz. **All panels.** Error ribbons denote bootstrap-estimated 95% confidence intervals.

439 Knowledge estimates need not be limited to the content of the lectures. As illustrated in
 440 Figure 8, our general approach to estimating knowledge from a small number of quiz questions
 441 may be extended to *any* content, given its text embedding coordinate. To visualize how knowledge
 442 “spreads” through text embedding space to content beyond the lectures participants watched, we
 443 first fit a new topic model to the lectures’ sliding windows with $k = 100$ topics. Conceptually,
 444 increasing the number of topics used by the model functions to increase the “resolution” of the
 445 embedding space, providing a greater ability to estimate knowledge for content that is highly
 446 similar to (but not precisely the same as) that contained in the two lectures. We note that we
 447 used these 2D maps solely for visualization; all relevant comparisons, distance computations, and

448 statistical tests we report above were carried out in the original 15-dimensional space, using the
449 15-topic model. Aside from increasing the number of topics from 15 to 100, all other procedures
450 and model parameters were carried over from the preceding analyses. As in our other analyses,
451 we resampled each lecture’s topic trajectory to 1 Hz and projected each question into a shared text
452 embedding space.

453 We projected the resulting 100-dimensional topic vectors (for each second of video and each quiz
454 question) onto a shared 2-dimensional plane (see *Creating knowledge and learning map visualizations*).
455 Next, we sampled points from a 100×100 grid of coordinates that evenly tiled a rectangle enclos-
456 ing the 2D projections of the videos and questions. We used Equation 4 to estimate participants’
457 knowledge at each of these 10,000 sampled locations, and averaged these estimates across par-
458 ticipants to obtain an estimated average *knowledge map* (Fig. 8A). Intuitively, the knowledge map
459 constructed from a given quiz’s responses provides a visualization of how “much” participants
460 knew about any content expressible by the fitted text embedding model at the point in time when
461 they completed that quiz.

462 Several features of the resulting knowledge maps are worth noting. The average knowledge
463 map estimated from Quiz 1 responses (Fig. 8A, leftmost map) shows that participants tended to
464 have relatively little knowledge about any parts of the text embedding space (i.e., the shading is
465 relatively dark everywhere). The knowledge map estimated from Quiz 2 responses shows a marked
466 increase in knowledge on the left side of the map (around roughly the same range of coordinates
467 traversed by the *Four Fundamental Forces* lecture, indicated by the dotted blue line). In other words,
468 participants’ estimated increase in knowledge is localized to conceptual content that is nearby (i.e.,
469 related to) the content from the lecture they watched prior to taking Quiz 2. This localization is
470 non-trivial: these knowledge estimates are informed only by the embedded coordinates of the
471 *quiz questions*, not by the embeddings of either lecture (see Eqn. 4). Finally, the knowledge map
472 estimated from Quiz 3 responses shows a second increase in knowledge, localized to the region
473 surrounding the embedding of the *Birth of Stars* lecture participants watched immediately prior to
474 taking Quiz 3.

475 Another way of visualizing these content-specific increases in knowledge after participants



Figure 8: Mapping out the geometry of knowledge and learning. **A.** Average “knowledge maps” estimated using each quiz. Each map displays a 2D projection of the estimated knowledge about the content reflected by *all* regions of topic space (see *Creating knowledge and learning map visualizations*). The topic trajectories of the two lectures are indicated by dotted lines (blue: Lecture 1; green: Lecture 2), and the coordinates of each question are indicated by dots (light blue: Lecture 1-related; light green: Lecture 2-related; purple: general physics knowledge). Each map reflects an average across all participants. For individual participants’ maps, see Supplementary Figures 7, 8, and 9. **B.** Average “learning maps” estimated between each successive pair of quizzes. The learning maps follow the same general format as the knowledge maps in Panel A, but here the shading at each coordinate indicates the *difference* between the corresponding coordinates in the indicated pair of knowledge maps—i.e., how much the estimated knowledge “changed” between the two quizzes. Each map reflects an average across all participants. For individual participants’ maps, see Supplementary Figures 10 and 11. **C.** Word clouds for sampled points in topic space. Each word cloud displays the weighted blend of words underlying the topic proportions represented at the corresponding colored star’s location on the maps. In each word cloud, the words’ relative sizes correspond to their relative weights at the starred location, and their colors indicate their relative weights in *Four Fundamental Forces* (blue) versus *Birth of Stars* (green) lectures, on average, across all timepoints’ topic vectors.

476 viewed each lecture is displayed in Figure 8B. Taking the point-by-point difference between the
477 knowledge maps estimated from responses to a successive pair of quizzes yields a *learning map*
478 that describes the *change* in knowledge estimates from one quiz to the next. These learning maps
479 highlight that the estimated knowledge increases we observed across maps were specific to the
480 regions around the embeddings of each lecture, in turn.

481 Because the 2D projection we used to construct the knowledge and learning maps is invertible,
482 we may gain additional insights into these maps' meanings by reconstructing the original high-
483 dimensional topic vector for any location on the map we are interested in. For example, this could
484 serve as a useful tool for an instructor looking to better understand which content areas a student
485 (or a group of students) knows well (or poorly). As a demonstration, we show the top-weighted
486 words from the blends of topics reconstructed from three example locations on the maps (Fig. 8C):
487 one point near the *Four Fundamental Forces* embedding (yellow), a second point near the *Birth of*
488 *Stars* embedding (orange), and a third point between the two lectures' embeddings (pink). As
489 shown in the word clouds in the panel, the top-weighted words at the example coordinate near the
490 *Four Fundamental Forces* embedding tended to be weighted more heavily by the topics expressed
491 in that lecture. Similarly, the top-weighted words at the example coordinate near the *Birth of Stars*
492 embedding tended to be weighted more heavily by the topics expressed in *that* lecture. And the
493 top-weighted words at the example coordinate between the two lectures' embeddings show a
494 roughly even mix of words most strongly associated with each lecture.

495 Discussion

496 We developed a computational framework that uses short multiple-choice quizzes to gain nuanced
497 insights into what learners know and how their knowledge changes with training. First, we show
498 that our approach can automatically match the conceptual knowledge probed by individual quiz
499 questions to the corresponding moments in lecture videos when those concepts were presented
500 (Fig. 4). Next, we demonstrate how we can estimate moment-by-moment "knowledge traces"
501 that reflect the degree of knowledge participants have about each video's time-varying content,

502 and capture temporally specific increases in knowledge after viewing each lecture (Fig. 5). We
503 also show that these knowledge estimates can generalize to held-out questions (Fig. 6). Finally,
504 we use our framework to construct visual maps that provide snapshot estimates of how much
505 participants know about any concept within the scope of our text embedding model, and how
506 much their knowledge of those concepts changes with training (Fig. 8).

507 We view our work as making several contributions to the study of how people acquire con-
508 ceptual knowledge. First, from a methodological standpoint, our modeling framework provides
509 a systematic means of mapping out and characterizing knowledge in maps that have infinite (ar-
510 bitrarily many) numbers of coordinates, and of “filling out” those maps using relatively small
511 numbers of multiple choice quiz questions. Our experimental finding that we can use these maps
512 to predict responses to held-out questions has several psychological implications as well. For ex-
513 ample, concepts that are assigned to nearby coordinates by the text embedding model also appear
514 to be “known to a similar extent” (as reflected by participants’ responses to held-out questions;
515 Fig. 6). This suggests that participants also *conceptualize* similarly the content reflected by nearby
516 embedding coordinates. How participants’ knowledge falls off with spatial distance is captured
517 by the knowledge maps we infer from their quiz responses (e.g., Figs. 7, 8). In other words, our
518 study shows that knowledge about a given concept implies knowledge about related concepts,
519 and we also show how estimated knowledge falls off with distance in text embedding space.

520 In our study, we characterize the “coordinates” of participants’ knowledge using a relatively
521 simple “bag of words” text embedding model [LDA; 7]. More sophisticated text embedding mod-
522 els, such as transformer-based models [16, 48, 59, 62] can learn complex grammatical and semantic
523 relationships between words, higher-order syntactic structures, stylistic features, and more. We
524 considered using transformer-based models in our study, but we found that the text embeddings
525 derived from these models were surprisingly uninformative with respect to differentiating or oth-
526 erwise characterizing the conceptual content of the lectures and questions we used. We suspect
527 that this reflects a broader challenge in constructing models that are high-resolution within a given
528 domain (e.g., the domain of physics lectures and questions) *and* sufficiently broad so as to enable
529 them to cover a wide range of domains. For example, we found that the embeddings derived even

from much larger and more modern models like BERT [16], GPT [62], LLaMa [59], and others that are trained on enormous text corpora, end up yielding poor resolution within the content space spanned by individual course videos (Supp. Fig. 6). Whereas the LDA embeddings of the lectures and questions are “near” each other (i.e., 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 (top row of Supp. Fig. 6). The LDA embeddings of the questions for each lecture and the corresponding lecture’s trajectory are also similar. For example, as shown in Fig. 2C, the LDA embeddings for *Four Fundamental Forces* questions (blue dots) appear closer to the *Four Fundamental Forces* lecture trajectory (blue line), whereas the LDA embeddings for *Birth of Stars* questions (green dots) appear closer to the *Birth of Stars* lecture trajectory (green line). The BERT embeddings of the lectures and questions do not show this property (Supp. Fig. 6). We also examined per-question “content matches” between individual questions and individual moments of each lecture (Figs. 4, 6). 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. Finally, we also compared the pairwise correlations between embeddings of 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 *Four Fundamental Forces* material are highly correlated with the embeddings of the *Four Fundamental Forces* lecture, but not with the embeddings of *Birth of Stars*, questions about *Birth of Stars*, or general physics knowledge questions. We see a similar pattern with the LDA

embeddings of the *Birth of Stars* questions (Fig. 3, Supp. Fig. 2). In contrast, the BERT embeddings are all highly correlated with each other (Supp. Fig. 6). 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 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 to many different content areas will be an important challenge for future work in this domain.

Another application for large language models that does *not* require explicitly modeling the content of individual lectures or questions is to leverage the models’ abilities to generate text. For example, generative text models like ChatGPT [48] and LLaMa [59] are already being used to build a new generation of interactive tutoring systems [e.g., 39]. Unlike the approach we have taken here, these generative text model-based systems do not explicitly model what learners know, or how their knowledge changes over time with training. One could imagine building a hybrid system that combines the best of both worlds: a large language model that can *generate* text, combined with a smaller model that can *infer* what learners know and how their knowledge changes over time. Such a hybrid system could potentially be used to build the next generation of interactive tutoring systems that are able to adapt to learners’ needs in real time, and that are able to provide more nuanced feedback about what learners know and what they do not know.

At the opposite end of the spectrum from large language models, one could also imagine

586 simplifying some aspects of our LDA-based approach by computing simple word overlap metrics.
587 For example, the Jaccard similarity between text A and B is computed as the number of unique
588 words in the intersection of words from A and B divided by the number of unique words in the
589 union of words from A and B . In a supplementary analysis (Supp. Fig. 5), we compared the
590 LDA-based question-lecture matches we reported in Figure 4 with the Jaccard similarities between
591 each question and each sliding window of text from the corresponding lecture. As shown in
592 Supplementary Figure 5, this simple word-matching approach does not appear to capture the same
593 level of specificity as the LDA-based approach. Whereas the LDA-based approach often yields a
594 clear peak in the time series of correlations between each question and the corresponding lecture,
595 the Jaccard similarity-based approach does not. Furthermore, these LDA-based matches appear
596 to capture conceptual overlaps between the questions and lectures (Supp. Tab. 3), whereas simple
597 word matching does not. For example, one of the example questions examined in Supplementary
598 Figure 5 asks “Which of the following occurs as a cloud of atoms gets more dense?” The LDA-based
599 matches identify lecture timepoints where the relevant *topics* are discussed (e.g., when words like
600 “cloud,” “atom,” “dense,” etc., are mentioned *together*). The Jaccard similarity-based matches,
601 on the other hand, are strong when *any* of these words are mentioned, even if they do not occur
602 together.

603 We view our approach as occupying a sort of “sweet spot,” between much larger language
604 models and simple word matching-based approaches, that enables us to capture the relevant
605 conceptual content of course materials at an appropriate semantic scale. Our approach enables us
606 to accurately and consistently identify each question’s content in a way that also matches up with
607 what is presented in the lectures. In turn, this enables us to construct accurate predictions about
608 participants’ knowledge of the conceptual content tested by held-out questions (Fig. 6).

609 One limitation of our approach is that topic models contain no explicit internal representations
610 of more complex aspects of “knowledge,” like knowledge graphs, dependencies or associations
611 between concepts, causality, and so on. These representations might (in principle) be added
612 as extensions to our approach to more accurately and precisely capture, characterize, and track
613 learners’ knowledge. However, modeling these aspects of knowledge will likely require substantial

614 additional research effort.

615 Within the past several years, the global pandemic forced many educators to suddenly adapt to
616 teaching remotely [31, 45, 56, 63]. This change in world circumstances is happening alongside (and
617 perhaps accelerating) geometric growth in the availability of high-quality online courses from plat-
618 forms such as Khan Academy [32], Coursera [64], EdX [34], and others [53]. Continued expansion
619 of the global internet backbone and improvements in computing hardware have also facilitated
620 improvements in video streaming, enabling videos to be easily shared and viewed by increasingly
621 large segments of the world’s population. This exciting time for online course instruction provides
622 an opportunity to re-evaluate how we, as a global community, educate ourselves and each other.
623 For example, we can ask: what defines an effective course or training program? Which aspects of
624 teaching might be optimized and/or augmented by automated tools? How and why do learning
625 needs and goals vary across people? How might we lower barriers to receiving a high-quality
626 education?

627 Alongside these questions, there is a growing desire to extend existing theories beyond the
628 domain of lab testing rooms and into real classrooms [30]. In part, this has led to a recent
629 resurgence of “naturalistic” or “observational” experimental paradigms that attempt to better
630 reflect more ethologically valid phenomena that are more directly relevant to real-world situations
631 and behaviors [46]. In turn, this has brought new challenges in data analysis and interpretation. A
632 key step towards solving these challenges will be to build explicit models of real-world scenarios
633 and how people behave in them (e.g., models of how people learn conceptual content from real-
634 world courses, as in our current study). A second key step will be to understand which sorts
635 of signals derived from behaviors and/or other measurements [e.g., neurophysiological data; 3,
636 17, 43, 47, 50] might help to inform these models. A third major step will be to develop and
637 employ reliable ways of evaluating the complex models and data that are a hallmark of naturalistic
638 paradigms.

639 Beyond specifically predicting what people *know*, the fundamental ideas we develop here also
640 relate to the notion of “theory of mind” of other individuals [23, 28, 42]. Considering others’ unique
641 perspectives, prior experiences, knowledge, goals, etc., can help us to more effectively interact and

642 communicate [51, 55, 58]. One could imagine future extensions of our work (e.g., analogous to
643 the knowledge and learning maps shown in Fig. 8), that attempt to characterize how well-aligned
644 different people’s knowledge bases or backgrounds are. In turn, this might be used to model how
645 knowledge (or other forms of communicable information) flows not just between teachers and
646 students, but between friends having a conversation, individuals on a first date, participants at
647 a business meeting, doctors and patients, experts and non-experts, political allies or adversaries,
648 and more. For example, the extent to which two people’s knowledge maps “match” or “align” in
649 a given region of text embedding space might serve as a predictor of how effectively they will be
650 able to communicate about the corresponding conceptual content.

651 Ultimately, our work suggests a rich new line of questions about the geometric “form” of
652 knowledge, how knowledge changes over time, and how we might map out the full space of
653 what an individual knows. Our finding that detailed estimates about knowledge may be obtained
654 from short quizzes shows one way that traditional approaches to evaluation in education may be
655 extended. We hope that these advances might help pave the way for new approaches to teaching
656 or delivering educational content that are tailored to individual students’ learning needs and goals.

657 Materials and methods

658 Participants

659 We enrolled a total of 50 Dartmouth undergraduate students in our study. Participants received
660 optional course credit for enrolling. We asked each participant to complete a demographic survey
661 that included questions about their age, gender, native spoken language, ethnicity, race, hearing,
662 color vision, sleep, coffee consumption, level of alertness, and several aspects of their educational
663 background and prior coursework.

664 Participants’ ages ranged from 18 to 22 years (mean: 19.52 years; standard deviation: 1.09
665 years). A total of 15 participants reported their gender as male and 35 participants reported their
666 gender as female. A total of 49 participants reported their native language as “English” and 1

667 reported having another native language. A total of 47 participants reported their ethnicity as
668 “Not Hispanic or Latino” and three reported their ethnicity as “Hispanic or Latino.” Participants
669 reported their races as White (32 participants), Asian (14 participants), Black or African American
670 (5 participants), American Indian or Alaska Native (1 participant), and Native Hawaiian or Other
671 Pacific Islander (1 participant). (Note that some participants selected multiple racial categories.)

672 A total of 49 participants reporting having normal hearing and 1 participant reported having
673 some hearing impairment. A total of 49 participants reported having normal color vision and 1
674 participant reported being color blind. Participants reported having had, on the night prior to
675 testing, 2–4 hours of sleep (1 participant), 4–6 hours of sleep (9 participants), 6–8 hours of sleep (35
676 participants), or 8+ hours of sleep (5 participants). They reported having consumed, on the same
677 day and leading up to their testing session, 0 cups of coffee (38 participants), 1 cup of coffee (10
678 participants), 3 cups of coffee (1 participant), or 4+ cups of coffee (1 participant).

679 No participants reported that their focus was currently impaired (e.g., by drugs or alcohol).
680 Participants reported their current level of alertness, and we converted their responses to numerical
681 scores as follows: “very sluggish” (-2), “a little sluggish” (-1), “neutral” (0), “fairly alert” (1), and
682 “very alert” (2). Across all participants, a range of alertness levels were reported (range: -2–1;
683 mean: -0.10; standard deviation: 0.84).

684 Participants reported their undergraduate major(s) as “social sciences” (28 participants), “nat-
685 ural sciences” (16 participants), “professional” (e.g., pre-med or pre-law; 8 participants), “mathe-
686 matics and engineering” (7 participants), “humanities” (4 participants), or “undecided” (3 partici-
687 pants). Note that some participants selected multiple categories for their undergraduate major(s).
688 We also asked participants about the courses they had taken. In total, 45 participants reported hav-
689 ing taken at least one Khan Academy course in the past, and 5 reported not having taken any Khan
690 Academy courses. Of those who reported having watched at least one Khan Academy course,
691 7 participants reported having watched 1–2 courses, 11 reported having watched 3–5 courses, 8
692 reported having watched 5–10 courses, and 19 reported having watched 10 or more courses. We
693 also asked participants about the specific courses they had watched, categorized under different
694 subject areas. In the “Mathematics” area, participants reported having watched videos on AP

695 Calculus AB (21 participants), Precalculus (17 participants), Algebra 2 (14 participants), AP Cal-
696 culus BC (12 participants), Trigonometry (11 participants), Algebra 1 (10 participants), Geometry
697 (8 participants), Pre-algebra (7 participants), Multivariable Calculus (5 participants), Differential
698 Equations (5 participants), Statistics and Probability (4 participants), AP Statistics (2 participants),
699 Linear Algebra (2 participants), Early Math (1 participant), Arithmetic (1 participant), and other
700 videos not listed in our survey (5 participants). In the “Science and engineering” area, participants
701 reported having watched videos on Chemistry, AP Chemistry, or Organic Chemistry (21 partic-
702 ipants); Physics, AP Physics I, or AP Physics II (18 participants); Biology, AP Biology; or High
703 school Biology (15 participants); Health and Medicine (1 participant); or other videos not listed
704 in our survey (5 participants). We also asked participants whether they had specifically seen the
705 videos used in our experiment. Of the 45 participants who reported having taken at least
706 one Khan Academy course in the past, 44 participants reported that they had not watched the *Four*
707 *Fundamental Forces* video, and 1 participant reported that they were not sure whether they had
708 watched it. All participants reported that they had not watched the *Birth of Stars* video. When
709 we asked participants about non-Khan Academy online courses, they reported having watched
710 or taken courses on Mathematics (15 participants), Science and engineering (11 participants), Test
711 preparation (9 participants), Economics and finance (3 participants), Arts and humanities (2 partic-
712 ipants), Computing (2 participants), and other categories not listed in our survey (17 participants).
713 Finally, we asked participants about in-person courses they had taken in different subject areas.
714 They reported taking courses in Mathematics (38 participants), Science and engineering (37 par-
715 ticipants), Arts and humanities (34 participants), Test preparation (27 participants), Economics
716 and finance (26 participants), Computing (14 participants), College and careers (7 participants), or
717 other courses not listed in our survey (6 participants).

718 **Experiment**

719 We hand-selected two course videos from the Khan Academy platform: *Four Fundamental Forces*
720 (an introduction to gravity, electromagnetism, the weak nuclear force, and the strong nuclear force;
721 duration: 10 minutes and 29 seconds) and *Birth of Stars* (an introduction to how stars are formed;

722 duration: 7 minutes and 57 seconds). All participants viewed the videos in the same order (i.e.,
723 *Four Fundamental Forces* followed by *Birth of Stars*).

724 We then hand-created 39 multiple-choice questions: 15 about the conceptual content of *Four*
725 *Fundamental Forces* (i.e., Lecture 1), 15 about the conceptual content of *Birth of Stars* (i.e., Lecture 2),
726 and 9 questions that tested for general conceptual knowledge about basic physics (covering material
727 that was not presented in either video). To help broaden the set of lecture-specific questions,
728 our team worked through each lecture in small segments to identify what each segment was
729 “about” conceptually, and then write a question about that concept. The general physics questions
730 were drawn our team’s prior coursework and areas of interest, along with internet searches and
731 brainstorming with the project team and other members of J.R.M.’s lab. Although we attempted to
732 design the questions to test “conceptual knowledge,” we note that estimating the specific “amount”
733 of conceptual understanding that each question “requires” to answer is somewhat subjective, and
734 might even come down to the “strategy” a given participant uses to answer the question at that
735 particular moment. The full set of questions and answer choices may be found in Supplementary
736 Table 1. The final set of questions (and response options) was reviewed and approved by J.R.M.
737 before we collected or analyzed the text or experimental data.

738 Over the course of the experiment, participants completed three 13-question multiple-choice
739 quizzes: the first before viewing Lecture 1, the second between Lectures 1 and 2, and the third
740 after viewing Lecture 2 (see Fig. 1). The questions appearing on each quiz, for each participant,
741 were randomly chosen from the full set of 39, with the constraints that (a) each quiz contained
742 exactly 5 questions about Lecture 1, 5 questions about Lecture 2, and 3 questions about general
743 physics knowledge, and (b) each question appear exactly once for each participant. The orders of
744 questions on each quiz, and the orders of answer options for each question, were also randomized.
745 We obtained informed consent from all participants, and our experimental protocol was approved
746 by the Committee for the Protection of Human Subjects at Dartmouth College. We used this
747 experiment to develop and test our computational framework for estimating knowledge and
748 learning.

749 **Analysis**

750 **Statistics**

751 All of the statistical tests performed in our study were two-sided. The 95% confidence intervals
752 we reported for each correlation were estimated by generating 10,000 bootstrap distributions of
753 correlation coefficients by sampling (with replacement) from the observed data.

754 **Constructing text embeddings of multiple lectures and questions**

755 We adapted an approach we developed in prior work [25] to embed each moment of the two
756 lectures and each question in our pool in a common representational space. Briefly, our approach
757 uses a topic model [Latent Dirichlet Allocation; 7] trained on a set of documents, to discover a set
758 of k “topics” or “themes.” Formally, each topic is defined as a distribution of weights over words
759 in the model’s vocabulary (i.e., the union of all unique words, across all documents, excluding
760 “stop words.”). Conceptually, each topic is intended to give larger weights to words that are
761 semantically related (as inferred from their tendency to co-occur in the same document). After
762 fitting a topic model, each document in the training set, or any *new* document that contains at
763 least some of the words in the model’s vocabulary, may be represented as a k -dimensional vector
764 describing how much the document (most probably) reflects each topic. To select an appropriate
765 k for our model, as a starting point, we identified the minimum number of topics that yielded
766 at least one “unused” topic (i.e., in which all words in the vocabulary were assigned uniform
767 weights) after training. This indicated that the number of topics was sufficient to capture the set
768 of latent themes present in the two lectures (from which we constructed our document corpus, as
769 described below). We found this value to be $k = 15$ topics. We found that with a limited number
770 of additional adjustments following [8], such as removing corpus-specific stop-words, the model
771 yielded (subjectively) sensible and coherent topics. The distribution of weights over words in
772 the vocabulary for each discovered topic is shown in Supplementary Figure 1, and each topic’s
773 top-weighted words may be found in Supplementary Table 2.

774 As illustrated in Figure 2A, we start by building up a corpus of documents using overlapping

775 sliding windows that span each video’s transcript. Khan Academy provides professionally created,
776 manual transcriptions of all videos for closed captioning. However, such transcripts would not
777 be readily available in all contexts to which our framework could potentially be applied. Khan
778 Academy videos are hosted on the YouTube platform, which additionally provides automated
779 captions. We opted to use these automated transcripts [which, in prior work, we have found to be
780 of sufficiently near-human quality to yield reliable data in behavioral studies; 65] when developing
781 our framework in order to make it more directly extensible and adaptable by others in the future.

782 We fetched these automated transcripts using the `youtube-transcript-api` Python pack-
783 age [15]. The transcripts consisted of one timestamped line of text for every few seconds (mean:
784 2.34 s; standard deviation: 0.83 s) of spoken content in the video (i.e., corresponding to each indi-
785 vidual caption that would appear on-screen if viewing the lecture via YouTube, and when those
786 lines would appear). We defined a sliding window length of (up to) $w = 30$ transcript lines, and
787 assigned each window a timestamp corresponding to the midpoint between the timestamps for its
788 first and last lines. This w parameter was chosen to match the same number of words per sliding
789 window (rounded to the nearest whole word, and before preprocessing) as the sliding windows
790 we defined in our prior work [25; i.e., 185 words per sliding window].

791 These sliding windows ramped up and down in length at the beginning and end of each
792 transcript, respectively. In other words, each transcript’s first sliding window covered only its first
793 line, the second sliding window covered the first two lines, and so on. This ensured that each line
794 from the transcripts appeared in the same number (w) of sliding windows. We next performed a
795 series of standard text preprocessing steps: normalizing case, lemmatizing, removing punctuation
796 and removing stop-words. We constructed our corpus of stop words by augmenting the Natural
797 Language Toolkit [NLTK; 4] English stop word list with the following additional words, selected
798 using one of the approaches suggested by [8]: “actual,” “actually,” “also,” “bit,” “could,” “e,”
799 “even,” “first,” “follow,” “following,” “four,” “let,” “like,” “mc,” “really,” “saw,” “see,” “seen,”
800 “thing,” and “two.” This yielded sliding windows with an average of 73.8 remaining words, and
801 lasting for an average of 62.22 seconds. We treated the text from each sliding window as a single
802 “document,” and combined these documents across the two videos’ windows to create a single

803 training corpus for the topic model.

804 After fitting a topic model to the two videos' transcripts, we could use the trained model to
805 transform arbitrary (potentially new) documents into k -dimensional topic vectors. A convenient
806 property of these topic vectors is that documents that reflect similar blends of topics (i.e., documents
807 that reflect similar themes, according to the model) will yield similar coordinates (in terms of
808 correlation, cosine similarity, Kullback-Leibler divergence, Euclidean distance, or other geometric
809 measures). In general, the similarity between different documents' topic vectors may be used to
810 characterize the similarity in conceptual content between the documents.

811 We transformed each sliding window's text into a topic vector, and then used linear interpolation
812 (independently for each topic dimension) to resample the resulting time series to one vector
813 per second. We also used the fitted model to obtain topic vectors for each question in our pool (see
814 Supp. Tab. 1). Taken together, we obtained a *trajectory* for each video, describing its path through
815 topic space, and a single coordinate for each question (Fig. 2C). Embedding both videos and all of
816 the questions using a common model enables us to compare the content from different moments
817 of videos, compare the content across videos, and estimate potential associations between specific
818 questions and specific moments of video.

819 **Estimating dynamic knowledge traces**

820 We used the following equation to estimate each participant's knowledge about timepoint t of a
821 given lecture, $\hat{k}(t)$:

$$\hat{k}(f(t, L)) = \frac{\sum_{i \in \text{correct}} \text{ncorr}(f(t, L), f(i, Q))}{\sum_{j=1}^N \text{ncorr}(f(t, L), f(j, Q))}, \quad (1)$$

822 where

$$\text{ncorr}(x, y) = \frac{\text{corr}(x, y) - \text{mincorr}}{\text{maxcorr} - \text{mincorr}}, \quad (2)$$

823 and where mincorr and maxcorr are the minimum and maximum correlations between any lecture
824 timepoint and question, taken over all timepoints in the given lecture, and all five questions *about*
825 that lecture appearing on the given quiz. We also define $f(s, \Omega)$ as the s^{th} topic vector from the set

826 of topic vectors Ω . Here t indexes the set of lecture topic vectors, L , and i and j index the topic
827 vectors of questions used to estimate the knowledge trace, Q . Note that “correct” denotes the set
828 of indices of the questions the participant answered correctly on the given quiz.

829 Intuitively, $\text{ncorr}(x, y)$ is the correlation between two topic vectors (e.g., the topic vector from one
830 timepoint in a lecture, x , and the topic vector for one question, y), normalized by the minimum and
831 maximum correlations (across all timepoints t and questions Q) to range between 0 and 1, inclusive.
832 Equation 1 then computes the weighted average proportion of correctly answered questions about
833 the content presented at timepoint t , where the weights are given by the normalized correlations
834 between timepoint t ’s topic vector and the topic vectors for each question. The normalization step
835 (i.e., using ncorr instead of the raw correlations) ensures that every question contributes some
836 non-negative amount to the knowledge estimate.

837 **GLMM METHODS SECTION PLACEHOLDER**

838 In the set of analyses reported in Figure 6, we assessed whether estimates of participants’ knowledge
839 at the embedding coordinates of individual quiz questions could reliably predict their ability to
840 correctly answer those questions. In essence, we treated each question a given participant answered
841 on a given quiz as a “lecture” consisting of a single timepoint, and used Equation 1 to estimate
842 the participant’s knowledge for its embedding coordinate based on their performance on all *other*
843 questions they answered on that same quiz (“All questions”; Fig. 6, top row). We also computed
844 analogous knowledge estimates for each lecture-related question based on all other questions
845 (answered by the same participant, on the same quiz) about the same lecture (“Within-lecture”;
846 Fig. 6, middle rows) and based on all questions about the other of the two lectures (“Across-lecture”;
847 Fig. 6, bottom rows).

848 In each version of this analysis (i.e., row in Fig. 6), and separately for each of the three quizzes
849 (i.e., column in Fig. 6), we then fit a generalized linear mixed model (GLMM) with a logistic link
850 function to the set of knowledge estimates for all questions answered by all participants on the
851 given quiz. Our models were implemented in R using the `lme4` package [2].

$$\text{accuracy} \sim \text{knowledge} + (\text{knowledge} | \text{participant}) + (\text{knowledge} | \text{question})$$

852 note that qID is implicitly nested within levels of lecture

$$\begin{aligned} \text{accuracy}_i &\sim \text{Binomial}(n = 1, \text{prob}_{\text{accuracy}=\text{correct}} = \widehat{P}) \\ \log \left[\frac{\widehat{P}}{1 - \widehat{P}} \right] &= \alpha_{j[i], k[i]} \\ \begin{pmatrix} \alpha_j \\ \gamma_{1j} \end{pmatrix} &\sim N \left(\begin{pmatrix} \gamma_0^\alpha + \gamma_{1k[i]}^\alpha (\text{knowledge}) \\ \mu_{\gamma_{1j}} \end{pmatrix}, \begin{pmatrix} \sigma_{\alpha_j}^2 & \rho_{\alpha_j \gamma_{1j}} \\ \rho_{\gamma_{1j} \alpha_j} & \sigma_{\gamma_{1j}}^2 \end{pmatrix} \right), \text{ for participant } j = 1, \dots, J \\ \begin{pmatrix} \alpha_k \\ \gamma_{1k} \end{pmatrix} &\sim N \left(\begin{pmatrix} \mu_{\alpha_k} \\ \mu_{\gamma_{1k}} \end{pmatrix}, \begin{pmatrix} \sigma_{\alpha_k}^2 & \rho_{\alpha_k \gamma_{1k}} \\ \rho_{\gamma_{1k} \alpha_k} & \sigma_{\gamma_{1k}}^2 \end{pmatrix} \right), \text{ for question } k = 1, \dots, K \end{aligned}$$

853 **Estimating the “smoothness” of knowledge**

854 In the analysis reported in Figure 7A, we show how participants’ ability to correctly answer
 855 quiz questions changes as a function of distance from a given correctly or incorrectly answered
 856 reference question. We used a bootstrap-based approach to estimate the maximum distances over
 857 which these proportions of correctly answered questions could be reliably distinguished from
 858 participants’ overall average proportion of correctly answered questions.

859 For each of 10,000 iterations, we drew a random subsample (with replacement) of 50 partic-
 860 ipants from our dataset. Within each iteration, we first computed the 95% confidence interval
 861 (CI) of the across-subsample-participants mean proportion correct on each of the three quizzes,
 862 separately. To compute this interval for each quiz, we repeatedly (1,000 times) subsampled par-
 863 ticipants (with replacement, from the outer subsample for the current iteration) and computed
 864 the mean proportion correct of each of these inner subsamples. We then identified the 2.5th and
 865 97.5th percentiles of the resulting distributions of 1,000 means. These three intervals (one for each
 866 quiz) served as our thresholds for confidence that the proportion correct within a given distance
 867 from a reference question was reliably different (at the $p < 0.05$ significance level) from the average

868 proportion correct across all questions on the given quiz.

869 Next, for each participant in the current subsample, and for each of the three quizzes they
870 completed (separately), we iteratively treated each of the 15 questions appearing on the given
871 quiz as the “reference” question. We constructed a series of concentric 15-dimensional “spheres”
872 centered on the reference question’s embedding space coordinate, where each successive sphere’s
873 radius increased by 0.01 (correlation distance) between 0 and 2, inclusive (i.e., tiling the range
874 of possible correlation distances with 201 spheres in total). We then computed the proportion
875 of questions enclosed within each sphere that the participant answered correctly, and averaged
876 these per-radius proportion correct scores across reference questions that were answered correctly,
877 and those that were answered incorrectly. This resulted in two number-of-spheres sequences of
878 proportion-correct scores for each subsample participant and quiz: one derived from correctly
879 answered reference questions, and one derived from incorrectly answered reference questions.

880 We computed the across-subsample-participants mean proportion correct for each radius value
881 (i.e., sphere) and “correctness” of reference question. This yielded two sequences of proportion-
882 correct scores for each quiz, analogous to the blue and red lines displayed in Figure 7A, but for
883 the present subsample. For each quiz, we then found the minimum distance from the reference
884 question (i.e., sphere radius) at which each of these two sequences of per-radius proportion correct
885 scores intersected the 95% confidence interval for the overall proportion correct (i.e., analogous to
886 the black error bands in Fig. 7A).

887 This resulted in two “intersection” distances for each quiz (for correctly answered and incor-
888 rectly answered reference questions). Repeating this full process for each of the 10,000 bootstrap
889 iterations output two distributions of intersection distances for each of the three quizzes. The
890 means and 95% confidence intervals for these distributions are plotted in Figure 7B.

891 **Creating knowledge and learning map visualizations**

892 An important feature of our approach is that, given a trained text embedding model and partic-
893 ipants’ quiz performance on each question, we can estimate their knowledge about *any* content
894 expressible by the embedding model—not solely the content explicitly probed by the quiz ques-

tions, or even appearing in the lectures. To visualize these estimates (Fig. 8, Supp. Figs. 7, 8, 9, 10, and 11), we used Uniform Manifold Approximation and Projection [UMAP; 40, 41] to construct a 2D projection of the text embedding space. Whereas our main analyses used a 15-topic embedding space, we used a 100-topic embedding space for these visualizations. This change in the number of topics overcame an undesirable behavior in the UMAP embedding procedure, whereby embedding coordinates for the 15-topic model tended to be “clumped” into separated clusters, rather than forming a smooth trajectory through the 2D space. When we increased the number of topics to 100, the embedding coordinates in the 2D space formed a smooth trajectory through the space, with substantially less clumping (Fig. 8). Creating a “map” by sampling this 100-dimensional space at high resolution to obtain an adequate set of topic vectors spanning the embedding space would be computationally intractable. However, sampling a 2D grid is trivial.

At a high level, the UMAP algorithm obtains low-dimensional embeddings by minimizing the cross-entropy between the pairwise (clustered) distances between the observations in their original (e.g., 100-dimensional) space and the pairwise (clustered) distances in the low-dimensional embedding space (in our approach, the embedding space is 2D). In our implementation, pairwise distances in the original high-dimensional space were defined as 1 minus the correlation between each pair of coordinates, and pairwise distances in the low-dimensional embedding space were defined as the Euclidean distance between each pair of coordinates.

In our application, all of the coordinates we embedded were topic vectors, whose elements are always non-negative and sum to one. Although UMAP is an invertible transformation at the embedding locations of the original data, other locations in the embedding space will not necessarily follow the same implicit “rules” as the original high-dimensional data. For example, inverting an arbitrary coordinate in the embedding space might result in negative-valued vectors, which are incompatible with the topic modeling framework. To protect against this issue, we log-transformed the topic vectors prior to embedding them in the 2D space. When we inverted the embedded vectors (e.g., to estimate topic vectors or word clouds, as in Fig. 8C), we passed the inverted (log-transformed) values through the exponential function to obtain a vector of non-negative values, and normalized them to sum to one.

923 After embedding both lectures' topic trajectories and the topic vectors of every question, we
924 defined a rectangle enclosing the 2D projections of the lectures' and quizzes' embeddings. We then
925 sampled points from a regular 100×100 grid of coordinates that evenly tiled this enclosing rectangle.
926 We sought to estimate participants' knowledge (and learning, i.e., changes in knowledge) at each
927 of the resulting 10,000 coordinates.

928 To generate our estimates, we placed a set of 39 radial basis functions (RBFs) throughout the
929 embedding space, centered on the 2D projections for each question (i.e., we included one RBF for
930 each question). At coordinate x , the value of an RBF centered on a question's coordinate μ , is given
931 by:

$$\text{RBF}(x, \mu, \lambda) = \exp\left\{-\frac{\|x - \mu\|^2}{\lambda}\right\}. \quad (3)$$

932 The λ term in the RBF equation controls the "smoothness" of the function, where larger values
933 of λ result in smoother maps. In our implementation we used $\lambda = 50$. Next, we estimated the
934 "knowledge" at each coordinate, x , using:

$$\hat{k}(x) = \frac{\sum_{i \in \text{correct}} \text{RBF}(x, q_i, \lambda)}{\sum_{j=1}^N \text{RBF}(x, q_j, \lambda)}. \quad (4)$$

935 Intuitively, Equation 4 computes the weighted proportion of correctly answered questions, where
936 the weights are given by how nearby (in the 2D space) each question is to the x . We also defined
937 *learning maps* as the coordinate-by-coordinate differences between any pair of knowledge maps.
938 Intuitively, learning maps reflect the *change* in knowledge across two maps.

939 Author contributions

940 Conceptualization: P.C.F., A.C.H., and J.R.M. Methodology: P.C.F., A.C.H., and J.R.M. Software:
941 P.C.F. Validation: P.C.F. Formal analysis: P.C.F. Resources: P.C.F., A.C.H., and J.R.M. Data curation:
942 P.C.F. Writing (original draft): J.R.M. Writing (review and editing): P.C.F., A.C.H., and J.R.M. Visu-
943 alization: P.C.F. and J.R.M. Supervision: J.R.M. Project administration: P.C.F. Funding acquisition:
944 J.R.M.

945 **Data availability**

946 All of the data analyzed in this manuscript may be found at <https://github.com/ContextLab/efficient-learning-khan>.
947

948 **Code availability**

949 All of the code for running our experiment and carrying out the analyses may be found at
950 <https://github.com/ContextLab/efficient-learning-khan>.

951 **Acknowledgements**

952 We acknowledge useful discussions, assistance in setting up an earlier (unpublished) version of
953 this study, and assistance with data collection efforts from Will Baxley, Max Bluestone, Daniel
954 Carstensen, Kunal Jha, Caroline Lee, Lucy Owen, Xinming Xu, and Kirsten Ziman. Our work was
955 supported in part by NSF CAREER Award Number 2145172 to J.R.M. The content is solely the
956 responsibility of the authors and does not necessarily represent the official views of our supporting
957 organizations. The funders had no role in study design, data collection and analysis, decision to
958 publish, or preparation of the manuscript.

959 **References**

- 960 [1] Ashby, F. G. and Maddox, W. T. (2005). Human category learning. *Annual Review of Psychology*,
961 56:149–178.
962 [2] Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models
963 using lme4. *Journal of Statistical Software*, 67(1):1–48.
964 [3] Bevilacqua, D., Davidesco, I., Wan, L., and Chaloner, K. (2019). Brain-to-brain synchrony and

- 965 learning outcomes vary by student-teacher dynamics: evidence from a real-world classroom
966 electroencephalography study. *Journal of Cognitive Neuroscience*, 31(3):401–411.
- 967 [4] Bird, S., Klein, E., and Loper, E. (2009). *Nature language processing with Python: analyzing text*
968 *with the natural language toolkit*. Reilly Media, Inc.
- 969 [5] Blaye, A., Bernard-Peyron, V., Paour, J.-L., and Bonthoux, F. (2006). Category flexibility in chil-
970 dren: distinguishing response flexibility from conceptual flexibility; the protracted development
971 of taxonomic representations. *European Journal of Developmental Psychology*, 3(2):163–188.
- 972 [6] Blei, D. M. and Lafferty, J. D. (2006). Dynamic topic models. In *Proceedings of the International*
973 *Conference on Machine Learning*, pages 113–120, New York, NY. Association for Computing
974 Machinery.
- 975 [7] Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine*
976 *Learning Research*, 3:993–1022.
- 977 [8] Boyd-Graber, J., Mimno, D., and Newman, D. (2014). Care and feeding of topic models:
978 problems, diagnostics, and improvements. In Airolid, E. M., Blei, D. M., Erosheva, E. A., and
979 Fienberg, S. E., editors, *Handbook of Mixed Membership Models and Their Applications*. CRC Press.
- 980 [9] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A.,
981 Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child,
982 R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M.,
983 Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei,
984 D. (2020). Language models are few-shot learners. *arXiv*, 2005.14165.
- 985 [10] Caramazza, A. and Mahon, B. Z. (2003). The organization of conceptual knowledge: the
986 evidence from category-specific semantic deficits. *Trends in Cognitive Sciences*, 7(8):354–361.
- 987 [11] Cer, D., Yang, Y., Kong, S. Y., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-
988 Cespedes, M., Yuan, S., Tar, C., Sung, Y.-H., Strope, B., and Kurzweil, R. (2018). Universal
989 sentence encoder. *arXiv*, 1803.11175.

- 990 [12] Constantinescu, A. O., O'Reilly, J. X., and Behrens, T. E. J. (2016). Organizing conceptual
991 knowledge in humans with a gridlike code. *Science*, 352(6292):1464–1468.
- 992 [13] Deacon, D., Grose-Fifer, J., Yang, C. M., Stanick, V., Hewitt, S., and Dynowska, A. (2004).
993 Evidence for a new conceptualization of semantic representation in the left and right cerebral
994 hemispheres. *Cortex*, 40(3):467–478.
- 995 [14] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990).
996 Indexing by latent semantic analysis. *Journal of the American Society for Information Science*,
997 41(6):391–407.
- 998 [15] Depoix, J. (2018). YouTube transcript API. <https://github.com/jdepoix/youtube-transcript-api>.
- 1000 [16] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). BERT: pre-training of deep
1001 bidirectional transformers for language understanding. *arXiv*, 1810.04805.
- 1002 [17] Dikker, S., Wan, L., Davidesco, I., Kaggen, L., Oostrik, M., McClintock, J., Rowland, J.,
1003 Michalareas, G., van Bavel, J. J., Ding, M., and Poeppel, D. (2017). Brain-to-brain synchrony
1004 tracks real-world dynamic group interactions in the classroom. *Current Biology*, 27(9):1375–1380.
- 1005 [18] Estes, W. K. (1986a). Array models for category learning. *Cognitive Psychology*, 18(4):500–549.
- 1006 [19] Estes, W. K. (1986b). Memory storage and retrieval processes in category learning. *Journal of*
1007 *Experimental Psychology: General*, 115:155–174.
- 1008 [20] Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical*
1009 *Transactions of the Royal Society A*, 222(602):309–368.
- 1010 [21] Gallagher, J. J. (2000). Teaching for understanding and application of science knowledge.
1011 *School Science and Mathematics*, 100(6):310–318.
- 1012 [22] Gluck, M. A., Shohamy, D., and Myers, C. E. (2002). How do people solve the “weather
1013 prediction” task? individual variability in strategies for probabilistic category learning. *Learning*
1014 *and Memory*, 9:408–418.

- 1015 [23] Goldstein, T. R. and Winner, E. (2012). Enhancing empathy and theory of mind. *Journal of*
1016 *Cognition and Development*, 13(1):19–37.
- 1017 [24] Hall, R. and Greeno, J. (2008). *21st century education: A reference handbook*, chapter Conceptual
1018 learning, pages 212–221. Sage Publications.
- 1019 [25] Heusser, A. C., Fitzpatrick, P. C., and Manning, J. R. (2021). Geometric models reveal behav-
1020 ioral and neural signatures of transforming experiences into memories. *Nature Human Behaviour*,
1021 5:905–919.
- 1022 [26] Huebner, P. A. and Willits, J. A. (2018). Structured semantic knowledge can emerge au-
1023 tomatically from predicting word sequences in child-directed speech. *Frontiers in Psychology*,
1024 9:doi.org/10.3389/fpsyg.2018.00133.
- 1025 [27] Hulbert, J. C. and Norman, K. A. (2015). Neural differentiation tracks improved recall of com-
1026 peting memories following interleaved study and retrieval practice. *Cerebral Cortex*, 25(10):3994–
1027 4008.
- 1028 [28] Kanske, P., Böckler, A., and Singer, T. (2015). Models, mechanisms and moderators dissociating
1029 empathy and theory of mind. In *Social Behavior From Rodents to Humans*, pages 193–206. Springer.
- 1030 [29] Katona, G. (1940). *Organizing and memorizing: studies in the psychology of learning and teaching*.
1031 Columbia University Press.
- 1032 [30] Kaufman, D. M. (2003). Applying educational theory in practice. *British Medical Journal*,
1033 326(7382):213–216.
- 1034 [31] Kawasaki, H., Yamasaki, S., Masuoka, Y., Iwasa, M., Fukita, S., and Matsuyama, R. (2021).
1035 Remote teaching due to COVID-19: an exploration of its effectiveness and issues. *International*
1036 *Journal of Environmental Research and Public Health*, 18(5):2672.
- 1037 [32] Khan, S. (2004). *The Khan Academy*. Salman Khan.
- 1038 [33] Kintsch (1970). *Learning, memory, and conceptual processes*. Wiley.

- 1039 [34] Kolowich, S. (2013). How EdX plans to earn, and share, revenue from its free online courses.
- 1040 *The Chronicle of Higher Education*, 21:1–5.
- 1041 [35] Landauer, T. K. and Dumais, S. T. (1997). A solution to Plato’s problem: the latent semantic
- 1042 analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*,
- 1043 104:211–240.
- 1044 [36] Lee, H. and Chen, J. (2022). Predicting memory from the network structure of naturalistic
- 1045 events. *Nature Communications*, 13(4235):doi.org/10.1038/s41467–022–31965–2.
- 1046 [37] MacLellan, E. (2005). Conceptual learning: the priority for higher education. *British Journal of*
- 1047 *Educational Studies*, 53(2):129–147.
- 1048 [38] Manning, J. R. (2021). Episodic memory: mental time travel or a quantum “memory wave”
- 1049 function? *Psychological Review*, 128(4):711–725.
- 1050 [39] Manning, J. R., Menjunatha, H., and Kording, K. (2023). Chatify: A Jupyter extension
- 1051 for adding LLM-driven chatbots to interactive notebooks. <https://github.com/ContextLab/chatify>.
- 1053 [40] McInnes, L., Healy, J., and Melville, J. (2018a). UMAP: Uniform manifold approximation and
- 1054 projection for dimension reduction. *arXiv*, 1802(03426).
- 1055 [41] McInnes, L., Healy, J., Saul, N., and Großberger, L. (2018b). UMAP: Uniform Manifold
- 1056 Approximation and Projection. *Journal of Open Source Software*, 3(29):861.
- 1057 [42] Meltzoff, A. N. (2011). Social cognition and the origins of imitation, empathy, and theory of
- 1058 mind. In *The Wiley-Blackwell Handbook of Childhood Cognitive Development*. Wiley-Blackwell.
- 1059 [43] Meshulam, M., Hasenfratz, L., Hillman, H., Liu, Y. F., Nguyen, M., Norman, K. A., and Hasson,
- 1060 U. (2020). Neural alignment predicts learning outcomes in students taking an introduction to
- 1061 computer science course. *Nature Communications*, 12(1922):doi.org/10.1038/s41467–021–22202–3.
- 1062 [44] Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word represen-
- 1063 tations in vector space. *arXiv*, 1301.3781.

- 1064 [45] Moser, K. M., Wei, T., and Brenner, D. (2021). Remote teaching during COVID-19: implications
1065 from a national survey of language educators. *System*, 97:102431.
- 1066 [46] Nastase, S. A., Goldstein, A., and Hasson, U. (2020). Keep it real: rethinking the primacy of
1067 experimental control in cognitive neuroscience. *NeuroImage*, 15(222):117254–117261.
- 1068 [47] Nguyen, M., Chang, A., Micciche, E., Meshulam, M., Nastase, S. A., and Hasson, U. (2022).
1069 Teacher-student neural coupling during teaching and learning. *Social Cognitive and Affective*
1070 *Neuroscience*, 17(4):367–376.
- 1071 [48] OpenAI (2023). ChatGPT. <https://chat.openai.com>.
- 1072 [49] Piantadosi, S. T. and Hill, F. (2022). Meaning without reference in large language models.
1073 *arXiv*, 2208.02957.
- 1074 [50] Poulsen, A. T., Kamronn, S., Dmochowski, J., Parra, L. C., and Hansen, L. K. (2017). EEG
1075 in the classroom: synchronised neural recordings during video presentation. *Scientific Reports*,
1076 7:43916.
- 1077 [51] Ratka, A. (2018). Empathy and the development of affective skills. *American Journal of*
1078 *Pharmaceutical Education*, 82(10):doi.org/10.5688/ajpe7192.
- 1079 [52] Reilly, D. L., Cooper, L. N., and Elbaum, C. (1982). A neural model for category learning.
1080 *Biological Cybernetics*, 45(1):35–41.
- 1081 [53] Rhoads, R. A., Berdan, J., and Toven-Lindsey, B. (2013). The open courseware movement in
1082 higher education: unmasking power and raising questions about the movement’s democratic
1083 potential. *Educational Theory*, 63(1):87–110.
- 1084 [54] Scott, P., Asoko, H., and Leach, J. (2007). *Handbook of research on science education*, chapter
1085 Student conceptions and conceptual learning in science. Routledge.
- 1086 [55] Shao, Y. N., Sun, H. M., Huang, J. W., Li, M. L., Huang, R. R., and Li, N. (2018). Simulation-
1087 based empathy training improves the communication skills of neonatal nurses. *Clinical Simula-*
1088 *tion in Nursing*, 22:32–42.

- 1089 [56] Shim, T. E. and Lee, S. Y. (2020). College students' experience of emergency remote teaching
1090 during COVID-19. *Children and Youth Services Review*, 119:105578.
- 1091 [57] Simon, M. A., Tzur, R., Heinz, K., and Kinzel, M. (2004). Explicating a mechanism for
1092 conceptual learning: elaborating the construct of reflective abstraction. *Journal for Research in
1093 Mathematics Education*, 35(5):305–329.
- 1094 [58] Stepien, K. A. and Baernstein, A. (2006). Education for empathy. *Journal of General Internal
1095 Medicine*, 21:524–530.
- 1096 [59] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B.,
1097 Goyal, N., Hambro, E., Azhar, F., Rodriguz, A., Joulin, A., Grave, E., and Lample, G. (2023).
1098 LLaMA: open and efficient foundation language models. *arXiv*, 2302.13971.
- 1099 [60] Tulchinskii, E., Kuznetsov, K., Kushnareva, L., Cherniavskii, D., Barannikov, S., Pio-
1100 ntkovskaya, I., Nikolenko, S., and Burnaev, E. (2023). Intrinsic dimension estimation for robust
1101 detection of AI-generated texts. *arXiv*, 2306.04723.
- 1102 [61] van Paridon, J., Liu, Q., and Lupyan, G. (2021). How do blind people know that blue is cold?
1103 distributional semantics encode color-adjective associations. *Proceedings of the Annual Meeting of
1104 the Cognitive Science Society*, 43(43).
- 1105 [62] Viswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and
1106 Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing
1107 Systems*.
- 1108 [63] Whalen, J. (2020). Should teachers be trained in emergency remote teaching? Lessons learned
1109 from the COVID-19 pandemic. *Journal of Technology and Teacher Education*, 28(2):189–199.
- 1110 [64] Young, J. R. (2012). Inside the Coursera contract: how an upstart company might profit from
1111 free courses. *The Chronicle of Higher Education*, 19(7):1–4.
- 1112 [65] Ziman, K., Heusser, A. C., Fitzpatrick, P. C., Field, C. E., and Manning, J. R. (2018). Is

¹¹¹³ automatic speech-to-text transcription ready for use in psychological experiments? *Behavior*
¹¹¹⁴ *Research Methods*, 50:2597–2605.