

# An Explication and Classroom Field Study of the Virtual Human Interaction Lab's Expert (VHIL-E) LLM

Jeremy N. Bailenson,<sup>1</sup> Jonathan You,<sup>1</sup> David Markowitz,<sup>2</sup> Gustav Petersen,<sup>1</sup> Rabindra Ratan,<sup>3</sup>  
 Monique Santoso,<sup>1</sup> and Portia Wang<sup>1</sup>

## Abstract

VHIL-E is the Virtual Human Interaction Lab's Expert, a large language model (LLM) representing the lab's research, teaching, and outreach on virtual and augmented reality. We first motivate the project and then present best practices for implementing a retrieval-augmented generation (RAG), specifically the "Seven Cs": *collecting, cleaning, classifying, chunking, creating* embeddings, *correlating* embeddings into an index, and *connecting* the index to an LLM. We collected academic publications, transcribed public talks, dedicated interviews and news articles about the lab's research, and various curriculum materials from Virtual People, the lecture course taught about the lab's research for over two decades, resulting in over 2.3 million words broken into ~10,000 chunks. In study 1, we compared performance on a multiple-choice test of 231 questions between various implementations of the RAG system and traditional Base GPT models. Models generally performed well, between 83 and 90 percent, similar to student performance on course examinations. In study 2, an open-ended task implemented over 10 weeks in the Fall 2025 Virtual People course, students ( $N = 89$ ) used VHIL-E to query and understand the course materials and logged hallucinations, which were "egregiously wrong answers." Students then chose the single worst wrong answer over 8 weeks. They then compared the Base/RAG Hybrid—which prioritized the RAG but allowed ChatGPT to consult its general intelligence—to the RAG Constrained, which limited substantive information to the embedded index. Allowing VHIL-E access to GPT produced more than twice as many hallucinations as constraining it to the index. We discuss implications for scholars who build and use RAG-based LLM applications.

**Keywords:** virtual reality, augmented reality, large language models, RAG, artificial intelligence

**R**esearch labs accumulate a massive scope of knowledge over time, ranging from objective data to a lab's subjective worldview within their particular domain of study—in this case, the psychology of virtual reality (VR) and augmented reality (AR). The objective of this article is twofold, to both present the construction details and to report an initial field study of VHIL-E (Virtual Human Interaction Lab's Expert), which was built to represent two decades of the lab's research, outreach efforts, and pedagogical materials, and was tested in the Fall 2025 Virtual People course, in which students learn about the lab's work.

The first goal of this project is to explore incorporating VHIL-E as a tool for learning. Reviews and meta-analyses demonstrate the educational potential of LLMs,<sup>1,2</sup> but research also highlights the need to ground this work more in the learning sciences.<sup>3</sup> Moreover, egregiously incorrect responses can have an outsized impact on scholarship and education.<sup>4</sup> These so-called hallucinations can impede learning dramatically when undetected and uncorrected,<sup>5,6</sup> though they present a

unique pedagogical opportunity to educate students on the ability to assess source reliability and other skills.<sup>7,8</sup> We specifically built a retrieval-augmented generation (RAG) system because, in theory, it should reduce hallucinations by constraining the substantive information used by the LLM.

The second motivation is to assist with outreach. VHIL receives a high number of daily requests asking for information about VR and AR. We currently answer these queries by generating responses by hand. VHIL-E can help by not only replying faster but also more thoroughly, as current replies tend to be brief. Eventually, we plan to provide public access to VHIL-E directly to enhance the dissemination of scientific knowledge.<sup>9,10</sup>

The third goal is to assist with research.<sup>11,12</sup> Scholars increasingly incorporate LLMs into their workflows, but often without customized models. VHIL-E has already provided insights about VHIL research themes and trends over time and has highlighted differences by study sample size, variable type, public reaction, and effect sizes.

<sup>1</sup>Department of Communication, Stanford University, Stanford, CA, USA.

<sup>2</sup>Department of Communication, Michigan State University, East Lansing, Michigan, USA.

<sup>3</sup>Department of Media and Information, Michigan State University, East Lansing, Michigan USA.

Finally, the fourth goal is to study the process itself.<sup>13</sup> Building an embedded index from the ground up<sup>14</sup> has provided insights into best practices and has facilitated a generalized framework for creating an expert LLM for research labs called *The Seven Cs*, which we describe below.

### The Seven Cs

The Seven Cs are depicted as a process in Figure 1 and listed along with software choices and lessons learned in Table 1. We use a RAG, a pipeline for enhancing LLMs' knowledge base with domain-specific data. We chose a RAG over other approaches (e.g., fine-tuning) for its ability to provide accurate and up-to-date responses for specific domains by explicitly referencing external knowledge sources.<sup>14–16</sup>

The first C is *Collecting* various materials, which are described in detail in Table 2. While many academic labs' journal articles are accessible in central repositories like Google Scholar, additional resources (e.g., course materials, public talks, news articles) tend to be more dispersed. Algorithmic approaches can be productive, but in our experience, paywalls and variations on search terms require a fair amount of this to be done by hand. Our inclusion guidelines for academic publications were whether or not they were peer-reviewed. For the other materials, the lead author made decisions based on centrality (i.e., was the material focused mainly on the lab) and quality. Some materials were excluded as they were duplicates (e.g., news articles that were republished in multiple venues). All materials were either written by the authors or were available online.

The second C is *Cleaning* those materials so that they can be used properly. An initial challenge is handling extraneous information. For example, some journal articles include pointers to other similar articles in various places throughout the text. Furthermore, news articles often

include advertisements tailored semantically to the substance of the article. Another challenge is formatting artifacts that emerge, for example, when creating text versions of PDFs. While we were able to leverage some algorithms to clean the data, a majority of cleaning was actually done by hand to maximize accuracy.

*Classifying* is the process of labeling materials with meaningful metadata for the LLM. For example, academic papers had different tags (e.g., DOI, keywords) from interviews and news stories. These metadata helped the LLM connect chunks with attributes that could be queried. Having source data are particularly useful for students.

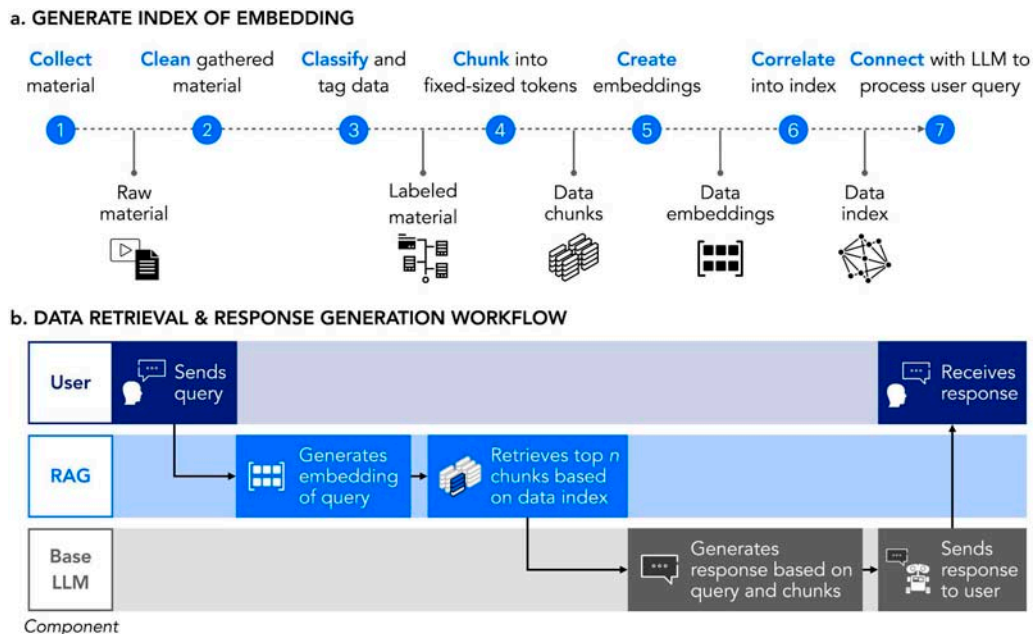
*Chunking* is the process of breaking the materials into roughly paragraph-sized bites (roughly 10,000 in our case). While some suggest the optimal chunk size can be around 512 tokens ( $\sim 500$  words),<sup>17</sup> our materials varied widely in format for testing and evaluation purposes. For example, the natural organizing length of utterances in a public talk is different from the stylized prose in a book. In the current study, we compared fixed chunk length to customized chunks that match the source materials, as noted in Table 2.

*Creating* embeddings is the process of changing a written chunk into a numerical vector to enable quantitative operations on the text. *Correlating* the embeddings into an index is the process of creating a searchable structure that allows similarity comparisons among the chunks.

Finally, *Connecting* means using APIs to send user queries to the index, and to return relevant chunks back to an LLM for processing.

### Study 1: Model Evaluation

The main research question for this study was to quantitatively examine how accurate VHIL-E was, both to provide a benchmark for its efficacy, as well as to iteratively test minor changes in parameters of the model.



**FIG. 1.** The Seven Cs. (a) shows the process of generating the embedded index and using it to process the user query. (b) shows the data retrieval and response generation workflow.

TABLE 1. SUMMARY OF THE SEVEN CS FOR VHIL-E RAG DEVELOPMENT: STAGES, IMPLEMENTATION TOOLS, AND LESSONS LEARNED

<i>Stage</i>	<i>Description</i>	<i>Tools</i>	<i>Most important lesson learned</i>
Collecting	Choosing the Type of Materials and Items of Each Type	Human Labor	The Expert themselves should be making decisions, it's okay to not be perfectly exhaustive
Cleaning	Removing formatting artifacts and information that could fuel hallucinations (i.e., reference sections, links to related stories).	Python Scripting	Script only works up to a point, human cleaning is critical to prevent artifacts
Classifying	Choosing the Tags that will be associated with all Chunks from the articles (i.e., authors, date, source, keywords).	YAML format readable for humans and AI	Consult domain experts to choose necessary tags based on returning information helpful for the user to have later on (e.g., sources)
Chunking	Breaking up articles into uniform (500 words) or customized small (200 to 400 words), medium (400 to 600 words), or large (500 to 800 words) chunks	Python Scripting	Get to know the cadence of your source type before choosing chunk length to maximize semantic fit
Creating Embeddings	Transforming chunks into vectors	Python Scripting, OpenAI text-embedding-3-large	Check "Leaderboard" for Embedding models before choosing to maximize performance
Correlating Index	Creating a searchable matrix of embeddings based on similarity	FAISS index	We chose open-source software, but lots of options exist
Connecting to LLM	Deploying the index with an API to a server so that a LLM can connect to it	FastAPI, Google Run Cloud, OpenAI APIs	Consider speed, accuracy, and amount of information per answer when choosing an LLM.

### Methods

We administered a test with 231 multiple-choice questions (MCQs) from the Virtual People course between the years of 2007 and 2024. The test covered themes related to the interdisciplinary study of VR, such as tracking and rendering, avatar psychology, training, and mental health applications. Approximately half of the test items had four response options; the rest had five or six options. This deterministic, quantitative assessment tool was useful for testing ground truth across many variations of VHIL-E without requiring human coders.

We tested four overall strategies by implementing various versions of the LLM. The *Base GPT* simply relied on the typical user input and asked it to "Answer the following question." However, given that many of the training materials from the RAG are likely in the general ChatGPT database, we also wanted to test the Base GPT when prompted to focus on the lab itself, and in the *Base GPT as VHIL Rep* condition, we asked it to: "Answer the following question as a representative of Stanford's Virtual Human Interaction Lab." The next strategy, *RAG Constrained*, limited the substantive information to the embedded index, prompted as: "When answering the following question, use only the retrieved passages from the VHIL-E index which were returned and no other data source. Do not use any other knowledge." Finally, we wanted to test a strategy that prioritized the embedded index, but also allowed information from the general ChatGPT database. In the *Base/RAG Hybrid* condition we prompted: "When answering the

following question, you can use either the retrieved passages from the VHIL-E index or knowledge from other data sources. Prioritize the retrieved passages from the VHIL-E index. If the VHIL-E index conflicts with other sources, choose the index."

For the two conditions with the embedded index, we tested the *Number of Chunks Returned*, either 3, 5, or 7. We also varied *Chunk Length*, either uniform chunks of 500 tokens, or assigned small (200–400 tokens), medium (400–600 tokens), or large (500–800 tokens) chunks based on the type of training materials (i.e., books versus talk transcriptions).<sup>17</sup> The customized strategy was designed to align chunk boundaries with typical paragraph lengths and structural markers (e.g., section headings), consistent with findings that retrieval systems benefit from segmentations that follow natural document structure.<sup>16,17</sup> We then compared the uniform and customized chunking strategies.

Finally, for every possible test combination described above, we compared LLM models, *gpt-4o* and *gpt-5*, from OpenAI. These are high-performance models that perform well on academic assessment tasks and have been used to extract or examine content from academic research texts in prior work.<sup>18</sup>

### Results

Table 3 presents the results across test combinations. To increase reliability, we repeated each test three times, and as the very small standard deviations show, the specific tests

TABLE 2. COLLECTED MATERIALS: CONTENT CATEGORIES, QUANTITIES, AND DESCRIPTIONS BY MATERIAL TYPE (2003–2025)

<i>Material</i>	<i>Number of items</i>	<i>Number of chunks</i>	<i>Chunk size</i>	<i>Total words</i>	<i>Average word count per Item</i>	<i>Year range</i>	<i>Description</i>
Books	2	553	Large	162484	81242	2010–2018	Trade books about VR and the lab's research, also used as textbooks in the Virtual People course.
Academic Publications	261	5761	Medium	1504120	5763	2003–2025	Peer-reviewed publications from VHIL scholars examining VR and AR across various academic disciplines
Dissertations	14	1057	Large	302595	21614	2007–2025	Doctoral dissertations and other theses completed by VHIL graduate researchers. Note that these were not included in the evaluation tests in the current article.
Keynote Talks by VHIL Members	57	1132	Small	419234	7355	2009–2025	Transcriptions of talks given to public audiences in academia, industry, conferences and other venues
News	87	580	Small	102896	1183	2008–2025	Communications material where VHIL's work was featured or discussed
Articles	55	308	Small	54208	986	2008–2024	News articles where the primary focus was on the lab's research
One-on-One Interviews	6	101	Small	18968	3161	2014–2023	News articles where a lab member was interviewed by a journalist about research
Stanford Press Releases	15	101	Small	16698	1113	2012–2025	Stanford press releases about particular research findings by VHIL
Op Eds written by VHIL	11	70	Small	13022	1184	2008–2023	Opinion pieces written by VHIL authors to highlight particular arguments and research findings
Virtual People Q&A Lectures	14	782	Medium	169469	12105	2021–2022	Recordings of class sessions in which students asked the teaching staff about previous readings and lectures
Virtual People Lecture Slides	17	63	Small	6037	355	2024	Written content of lecture slides from the most recent iteration of Virtual People

TABLE 3. COMPARING MULTIPLE-CHOICE EXAM PERFORMANCE BY MODEL, STRATEGY, NUMBER OF CHUNKS RETURNED, CHUNK LENGTH, AND TEST SCORE/STANDARD DEVIATIONS (ACROSS THREE TRIALS)

<i>Model</i>	<i>Strategy</i>	<i>Num. chunks returned</i>	<i>Chunk length</i>	<i>Test score</i>	<i>Std. Dev.</i>
gpt-4o	Base/RAG Hybrid	3	Uniform	0.8687	0.0050
gpt-4o	Base/RAG Hybrid	5	Uniform	0.8716	0.0025
gpt-4o	Base/RAG Hybrid	7	Uniform	0.8759	0.0025
gpt-4o	RAG Constrained	3	Uniform	0.8485	0.0075
gpt-4o	RAG Constrained	5	Uniform	0.8528	0.0087
gpt-4o	RAG Constrained	7	Uniform	0.8341	0.0050
gpt-4o	Base GPT	N/A	N/A	0.6898	0.0175
gpt-5	Base/RAG Hybrid	3	Uniform	0.8975	0.0050
gpt-5	Base/RAG Hybrid	5	Uniform	0.8947	0.0066
gpt-5	Base/RAG Hybrid	7	Uniform	0.9004	0.0043
gpt-5	RAG Constrained	3	Uniform	0.8413	0.0066
gpt-5	RAG Constrained	5	Uniform	0.8398	0.0150
gpt-5	RAG Constrained	7	Uniform	0.8644	0.0050
gpt-5	Base GPT	N/A	N/A	0.8831	0.0043
gpt-4o	Base GPT as VHIL Rep	N/A	N/A	0.8110	0.0050
gpt-5	Base GPT as VHIL Rep	N/A	N/A	0.8817	0.0109
gpt-4o	Base/RAG Hybrid	3	Customized	0.8600	0.0066
gpt-4o	Base/RAG Hybrid	5	Customized	0.8600	0.0025
gpt-4o	Base/RAG Hybrid	7	Customized	0.8615	0.0043
gpt-4o	RAG Constrained	3	Customized	0.8600	0.0050
gpt-4o	RAG Constrained	5	Customized	0.8571	0.0115
gpt-4o	RAG Constrained	7	Customized	0.8543	0.0025
gpt-4o	Base GPT	N/A	N/A	0.6696	0.0109
gpt-5	Base/RAG Hybrid	3	Customized	0.8961	0.0075
gpt-5	Base/RAG Hybrid	5	Customized	0.9004	0.0043
gpt-5	Base/RAG Hybrid	7	Customized	0.9004	0.0115
gpt-5	RAG Constrained	3	Customized	0.8557	0.0066
gpt-5	RAG Constrained	5	Customized	0.8629	0.0066
gpt-5	RAG Constrained	7	Customized	0.8672	0.0025
gpt-5	Base GPT	N/A	N/A	0.8687	0.0164

The prompts given for each strategy are as follows. Base GPT: “Answer the following question.” Base GPT as VHIL Representative: “Answer the following question as a representative of Stanford’s Virtual Human Interaction Lab.” RAG Constrained: “When answering the following question, use only the retrieved passages from the VHIL-E index which were returned and no other data source. Do not use any other knowledge.” Base/RAG Hybrid: “When answering the following question, you can use either the retrieved passages from the VHIL-E index or knowledge from other data sources. Prioritize the retrieved passages from the VHIL-E index. If the VHIL-E index conflicts with other sources, choose the index.”

performed similarly when repeated. We decided not to implement a Monte Carlo method to test significance, as we did not want to assign statistical significance to very small effect sizes, given how similarly the models performed. Other than the Base GPT condition using gpt-4o, which consistently scored below 70 percent, the rest of the combinations fell in the range of 83–90 percent. In general, the models performed well on the MCQs, fairly similar to actual students across the two decades of the course; the mean score on midterms and finals is typically between 85 percent and 90 percent. As a further demonstration of reliability, repeating the tests with small changes to parameters such as the number of chunks returned resulted in fairly consistent results.

**Study 2: Longitudinal Classroom Field Study**

The research question for this study was to examine whether or not limiting the LLM’s access to the embedded index, as opposed to letting it be influenced by the general model, would result in fewer hallucinations as assessed by a large team of expert coders.

We designed an in-class study to test performance on open-ended responses, and tested two of the strategies from the four described above—Base/RAG Hybrid and RAG-Constrained. Because of classroom field study limitations, we were constrained to only two conditions, which we chose to causally isolate the influence of possible hallucinated materials from the general GPT model. Both strategies used GPT-5, and the RAG used uniform chunk length and returned 5 chunks.

To maximize ecological validity, the study was conducted with students enrolled in the Virtual People course, for which the vast majority of the curriculum is represented in the RAG. In other words, this controlled field study constituted an optimal setting to examine VHIL-E’s ability to accurately represent the course material.

*Method*

Participants. Participants were 89 university students enrolled in the 10-week Virtual People course, taught in Fall Quarter 2025. At the beginning of the course, students were invited to participate in an Institutional Review Board-

approved (IRB) study of how AI and VR use influenced learning. All students who were part of the course took part in all the activities; however, we only included the data of those who consented to participate in this study by signing a consent form that was authorized by the university's IRB and a second oversight organization for students. To eliminate any potential perceptions of coercion, the procedure ensured that researchers and teaching team members were unaware of the identities of participating students until after the course finished by using a third-party arbiter.

There were 168 students in the course, 153 students consented to participate in the research study, and 95 were present for the final hallucination exercise. Students ( $M = 34$ ,  $F = 61$ ) were between 18 and 29 years old ( $n_{18\sim 23} = 93$ ,  $n_{24\sim 29} = 2$ ) and identified as Asian or Asian-American ( $n = 44$ ), White ( $n = 23$ ), African, African-American, or Black ( $n = 10$ ), Hispanic or Latinx ( $n = 6$ ), Indigenous/Native American, Alaska Native, First Nations ( $n = 1$ ), Native Hawaiian or other Pacific Island ( $n = 2$ ), middle Eastern ( $n = 2$ ), multiracial ( $n = 6$ ), and declined to or did not respond ( $n = 1$ ).

**Procedure.** Once a week for eight consecutive weeks (weeks 2 through 9 of the course), students were instructed to use the Base/RAG Hybrid model to examine the course materials. Specifically, we asked them to continually ask the Base/RAG Hybrid model questions about that week's assigned readings until it produced a verifiably wrong answer. They then logged the question and the wrong answer, and indicated why the answer was wrong. At the end of the course in week 10, we had each student choose the "most egregious wrong answer" from among their eight over the course. They then indicated whether the answer was an egregious hallucination with a yes or no response.

Following this, they entered the same question again into the RAG Constrained model, and after the new answer, students were asked to "Indicate whether it was an egregious hallucination" with a yes or no response. It is important to note that these students had been studying the course materials, which were used to train VHIL-E, for over 2 months at that point and were sufficiently qualified to identify errors.

Of the 95 students who participated, 6 responses were removed due to incorrectly prompting the Base/RAG Hybrid model instead of the RAG Constrained model. If participants submitted two responses ( $n = 4$ ), their second submission was removed. This resulted in a total of 89 submissions.

## Results

We counted the number of egregious responses for each strategy. When using the Base/RAG Hybrid model, participants reported 62 hallucinations (69.7 percent) and 27 non-hallucinations (30.3 percent). However, when using the RAG-constrained model, participants reported 29 hallucinations (32.6 percent) and 60 non-hallucinations (67.4 percent) to the same questions. A McNemar's test was conducted to examine changes in hallucination status. The proportion of participants reporting hallucinations significantly decreased in the RAG-constrained model,  $\chi^2(1) = 24.98$ ,  $OR = 9.25$ , 95 percent CI (3.32, 35.73), Cohen's  $g = 0.40$ ,  $p < 0.001$ , compared to the Base/RAG hybrid model.

## Discussion

We present the methodology to build an LLM to represent an academic lab's research, teaching, and outreach. In a longitudinal classroom field study, VHIL-E performed similarly to commercially available LLMs on MCQs but substantially reduced hallucinations in an open-ended task.

### *Theoretical implications and pedagogical implications*

Human performance on MCQs may be driven by recognition processes, or surface-level learning, while answering open-ended questions often requires active retrieval and language production processes that engage deeper levels of comprehension.<sup>19</sup> Similarly, research shows that LLM accuracy on MCQs drops significantly (up to  $\sim 40$  percentage points) when the correct answer is replaced with "none of the other answers", suggesting a strong reliance on pattern matching as opposed to reasoning.<sup>20</sup> Along with the argument that answering MCQs is not representative of how people typically use LLMs,<sup>21</sup> this underscores the importance of using open-ended questions for evaluating LLMs.

LLM hallucinations are extremely difficult to eliminate<sup>22</sup> and can be crippling in a scientific context, where credibility is a critical attribute for scholars. Indeed, one of the major goals of building VHIL-E was to prevent the use of blatantly incorrect scientific details that LLMs tend to portray confidently. The current study shows that when constraining the substantive information to a RAG, hallucinations decrease substantially.

### *Limitations and future research directions*

There are many limitations to this article. As we continue to improve VHIL-E, we will augment and improve the collected materials, for example, removing chunks that lead to bad responses. Also, the assessment reported in this article is preliminary, and we will continue to understand not only when VHIL-E is wrong but also why it is wrong, with the goals of building a better system. In our user study, we are relying on students themselves to identify hallucinations as opposed to expert coders. In future work, we plan on exploring the semantic properties of these wrong answers. In the current article, while we review and are influenced by theoretical work in psychology and other fields, the advances are more pragmatic from the current work than theoretical.

One of the most unexpected utilities of VHIL-E is the chunk-retrieval process. In the next iteration of the course, we plan on testing a novel RAG strategy. When a user asks a question, instead of using an LLM to further process the chunks that are returned into a summary answer, we simply provide the prose of the chunks themselves. Consequently, we can ensure materials are accurate and allow learners to embark on a constructivist process with those sources.<sup>23,24</sup> Preliminary observations show this approach to be promising for learning.

## Acknowledgements

The authors would like to thank Michael Bernstein, Michael Casale, Cyan DeVeaux, Jeff Hancock, Arjun Nagendran, Daniel Pimentel, Anna Carolina Queiroz, Tara Srirangarajan, Ruth Starkman, Yujie Tao, Kathy Yu, and Rui Zhu for advice, assistance or feedback on this project.

## References

1. Deng R, Jiang M, Yu X, et al. Does ChatGPT enhance student learning? A systematic review and meta-analysis of experimental studies. *Comput Educ* 2025;227:105224; doi: 10.1016/j.compedu.2024.105224
2. Wang J, Fan W. The effect of ChatGPT on students' learning performance, learning perception, and higher-order thinking: Insights from a meta-analysis. *Humanit Soc Sci Commun* 2025;12(1); doi: 10.1057/s41599-025-04787-y
3. Weidlich J, Gašević D, Drachsler H, et al. ChatGPT in education: An effect in search of a cause. *Computer Assisted Learn* 2025;41(5).
4. Elsayed H. The impact of hallucinated information in large language models on student learning outcomes: A critical examination of misinformation risks in AI-assisted education. *Northern Rev Algorithmic Res Theoretical Computation Complexity* 2024;9(8):11–23.
5. Ciubotaru BI. The hallucination problem in generative artificial intelligence: Accuracy and trust in digital learning. In: *Proceedings of the International Conference on Virtual Learning*, vol. 20. 2025, pp. 35–45.
6. Maynez J, Narayan S, Bohnet B, et al. On faithfulness and factuality in abstractive summarization. *ArXiv* 2020.
7. Torres RC. AI hallucination in the context of education: Exploring college students' use of generative AI for academic tasks. *Proceedings of the 2025 16th International Conference on E-Education, E-Business, E-Management and E-Learning (IC4E)*. IEEE, 2025, pp. 445–449.
8. Guo Q, Zhen J, Wu F, et al. Can students make STEM progress with the large language models (LLMs)? An empirical study of LLMs integration within middle school science and engineering practice. *J Educ Comput Res* 2025;63(2): 372–405.
9. Hughes RC, van Heerden A. PLOS-LLM: Can and should AI enable a new paradigm of scientific knowledge sharing? *PLOS Digit Health* 2024;3(4):e0000501.
10. Liao Z, Antoniak M, Cheong I, et al. LLMs as research tools: A large scale survey of researchers' usage and perceptions. *ArXiv* 2024.
11. Markowitz DM, Bailenson JN. A looking glass into a research wonderland: Decades of virtual reality scholarship explicated via natural language processing. *Cyberpsychol Behav Soc Netw* 2025;28(4):227–232.
12. Riva G, Mantovani F, Wiederhold BK, et al. Psychomatics—A multidisciplinary framework for understanding artificial minds. *Cyberpsychol Behav Soc Netw* 2025;28(7):515–523.
13. Zhang Y, Adila D, Shin C, et al. Personalize your LLM: Fake it then align it. *ArXiv* 2025.
14. Ovadia O, Brief M, Mishaeli M, et al. Fine-tuning or retrieval? Comparing knowledge injection in LLMs. In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 2024, pp. 237–250.
15. Soudani H, Kanoulas E, Hasibi F. Fine tuning vs. retrieval augmented generation for less popular knowledge. In: *Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*. 2024, pp. 12–22.
16. Robertson SE, Walker S. (1999). Okapi/Keenbow at TREC-8, Vol. 8. TREC, pp. 151–162. Available from: <https://trec.nist.gov/pubs/trec8/papers/okapi.pdf>
17. Lensu U. Impact of chunking granularity on accuracy and token consumption in retrieval-augmented generation for question-answering. *Master's Thesis*. 2025.
18. Chen D, Fisch A, Weston J, et al. Reading wikipedia to answer open-domain questions. *arXiv Preprint* 2017.
19. Ozuru Y, Briner S, Kurby CA, et al. Comparing comprehension measured by multiple-choice and open-ended questions. *Can J Exp Psychol* 2013;67(3):215–227.
20. Bedi S, Jiang Y, Chung P, et al. Fidelity of medical reasoning in large language models. *JAMA Netw Open* 2025;8(8): e2526021.
21. Balepur N, Rudinger R, Boyd-Graber JL. Which of these best describes multiple choice evaluation with LLMs? A) Forced B) Flawed C) Fixable D) All of the above. In: *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics* 2025;3394–3418.
22. Banerjee S, Agarwal A, Singla S. LLMs will always hallucinate, and we need to live with this. In: *Intelligent Systems Conference*. Springer Nature: Cham, Switzerland; 2025, pp. 624–648.
23. Fiorella L, Mayer RE. Eight ways to promote generative learning. *Educ Psychol Rev* 2016;28(4):717–741.
24. Gabbard RB. Constructivism, hypermedia, and the world wide web. *Cyberpsychology and Behavior* 2000;3(1): 103–110.

Address correspondence to:  
 Dr. Jeremy N. Bailenson  
 Department of Communication  
 Stanford University 450 Jane Stanford Way  
 Stanford, CA 94305-2050  
 USA

E-mail: Bailenson@Stanford.edu