More Sources, Better AI? When Context Helps and Harms AI-Assisted Discovery in Educational Data Archives

Xin Wei XWEI@DIGITALPROMISE.ORG
1001 Connecticut Avenue NW, Suite 935 Washington, D.C. 20036

Morgan Lee MPLEE@WPI.EDU
100 Institute Road, Worcester, MA 01609

Jie Min JIEM@STANFORD.EDU
560 Fremont Road, Stanford, CA 94305

Jeremy Roschelle JROSCHELLE@DIGITALPROMISE.ORG
702 Marshall Street, Suite 340, Redwood City, CA 94063

Abstract

Can AI support the research reasoning required to use unfamiliar educational datasets? We evaluated an AI-assisted discovery tool across two educational archives, ASSISTments and SEDA, using 50-item expert-designed assessments. We varied the sources available to the tool: dataset documentation, published papers, synthetic data samples, and their combination. Full triangulation across all three sources outperformed documentation alone in overall accuracy: 89% versus 67% on ASSISTments and 90% versus 73% on SEDA. However, more context was not always better. With codebook-oriented documentation, adding synthetic data without papers reduced appropriate-use accuracy from 80% to 40%, while adding papers without data reduced code-generation accuracy from 90% to 75%. Effective AI-assisted research support requires matching the information sources to how well a dataset's design fits the research questions.

Keywords: learning analytics, retrieval-augmented generation, educational data archives, knowledge representation, data discovery, information architecture

1. Introduction

Educational data archives could substantially advance research on teaching and learning, yet a persistent barrier limits their impact: the "semantic gap" between how researchers think about educational constructs and how databases encode them as operational variables. A researcher interested in student persistence must discover that the relevant variable is named attempt_count; one studying achievement disparities must navigate SEDA's complex naming conventions (e.g., mn_avg_ol) to locate district-level test score estimates. This vocabulary problem was identified decades ago in information retrieval (Furnas et al., 1987) and remains acute in educational data contexts.

The challenge is compounded by growing emphasis on making research data Findable, Accessible, Interoperable, and Reusable — The FAIR Guiding Principles (Wilkinson et al., 2016), now widely adopted by funding agencies and data repositories. In education, Bowers and Choi (2023) argue that FAIR data management is also an equity issue: without standardized, well-documented data infrastructure, under-resourced schools and districts are systematically disadvantaged in evidence-based decision-making. Yet educational archives

often fall short on the "Findable" dimension — not because metadata is absent, but because translating from a researcher's conceptual vocabulary to a dataset's operational vocabulary requires tacit knowledge that metadata alone rarely provides. As data sharing requirements expand (Institute of Education Sciences, 2024) while researchers remain largely unaware of them, untrained in how to share, and wary of sharing (Logan et al., 2021), the discoverability challenge will intensify.

This tacit knowledge required for effective data navigation is traditionally acquired through academic apprenticeship: working alongside experienced researchers who know a dataset's conventions, limitations, and analytical affordances. This apprenticeship model creates systematic inequities, disadvantaging early-career researchers, scholars at teaching-focused institutions, and those outside established research networks.

Large Language Models (LLMs) promise conversational interfaces that could bridge this gap, and Retrieval-Augmented Generation (RAG) has emerged as the dominant paradigm for grounding LLM outputs in retrieved documents to reduce hallucination (Lewis et al., 2020; Ji et al., 2023). Recent surveys document rapid growth in RAG applications across domains (Gao et al., 2023; Gupta et al., 2024), and the field has moved from naive single-source pipelines toward modular and agentic architectures that dynamically manage retrieval strategies and integrate diverse knowledge sources (Gao et al., 2023; Singh et al., 2025). Within education, LLMs have been discussed as supporting a range of student- and teacher-facing activities, including tutoring, automated assessment, and content generation (Kasneci et al., 2023), but their use for research data discovery remains largely unexplored.

Most systems retrieve from a single document collection, typically documentation or a knowledge base, and assume that the retrieved context provides sufficient grounding. Yet data discovery is inherently multifaceted: understanding a variable requires knowing its definition, how other researchers have employed it, and what the variable looks like in context. No single document type reliably provides all three. To capture these three facets, we apply cognitive psychology's knowledge taxonomy to RAG system design. Building on the distinction proposed by Anderson (1996) between declarative (factual) and procedural (how-to) knowledge, and on an account of schemata as organized knowledge structures (Rumelhart, 1980), we propose structural knowledge as a third critical dimension for data discovery. In educational data discovery, documentation provides declarative knowledge (variable definitions), published papers provide procedural knowledge (usage conventions), and synthetic data provides structural knowledge (concrete formatting and relationships).

We therefore hypothesize that information architecture (the deliberate distribution of these complementary knowledge sources across the model's information environment) will improve AI-assisted data discovery. This hypothesis has an equity dimension: effective information architecture depends on domain expertise and curation rather than expensive computation, and so may benefit resource-constrained institutions. We address two research questions:

RQ1: How does information architecture, operationalized as different combinations of documentation, research papers, and synthetic data, affect AI-assisted data discovery?

RQ2: Under what conditions do different knowledge sources help or harm performance across datasets and discovery tasks?

2. Methodology

2.1. Datasets

We selected two educational data archives representing contrasting documentation contexts. ASSISTments 2019-20 (Prihar and Gonsalves, 2021) contains log-level data from a mathematics intelligent tutoring system, including student interaction sequences, problem-level features, and performance indicators. Its documentation consists of a relatively brief codebook that defines variables but provides limited guidance on analytical conventions or data structure. SEDA 5.0 (Fahle et al., 2024) is the Stanford Education Data Archive, containing aggregated mathematics and reading achievement data at district and school levels.

We characterized the two archives' primary documentation using observable documentation features rather than post hoc performance outcomes. Specifically, we compared documentation length, file-structure coverage, variable definitions, data construction procedures, methodological guidance, and naming conventions. The ASSISTments documentation is a 14-page dataset description organized around the platform's ten-table standard dataset format. It defines tables, shared identifiers, and individual variables, and provides limited examples of cross-table linking, but offers relatively little extended methodological guidance about analytical conventions, data-generating processes, or appropriate variable use.

In contrast, the SEDA 5.0 documentation is a 95-page technical manual covering data construction, cleaning, cutscore and mean estimation, suppression rules, and detailed definitions and notation. We therefore describe the two archives as representing contrasting documentation contexts: ASSISTments as primarily codebook-oriented documentation and SEDA as technical-methodological documentation. These documentation differences are relevant for understanding baseline performance differences across datasets. However, the two datasets also differ in construct alignment — the degree to which researchers' likely questions match what the data was designed to measure.

2.2. Evaluation Framework

We developed 50-item assessments (10 items per domain) for each dataset across five capability domains reflecting distinct stages of the research discovery workflow:

1. **Variable Identification:** Can the system correctly identify which variables correspond to given research constructs?

2. **Methodological Understanding:** Does the system understand the dataset's collection methodology, sampling, and measurement approach?

3. **Usage Appropriateness:** Can the system determine whether a variable is appropriate for a given analytical purpose?

4. **Code Generation:** Can the system produce executable R code for data manipulation tasks?

5. **Hypothesis Generation:** Can the system formulate testable hypotheses using available variables?

Items were developed iteratively by domain experts. Each item was scored on a three-level scale (1 = correct; 0.5 = partial; 0 = incorrect). Two independent raters blind-scored all responses with high inter-rater reliability (Cohen's $\kappa = 0.87$). Disagreements were resolved through discussion between raters. Each configuration was run three times, with the same 50 items administered in identical order across all models and configurations. Although the generated text varied across runs, the expert scores were identical, indicating that run-to-run variation did not materially affect our quality measures for these items. Each item was presented to the model as a standalone query with no follow-up or iterative prompting. The ASSISTments and SEDA assessments were developed independently to reflect each dataset's unique content and structure; items were not designed to be parallel in difficulty across datasets, which means cross-dataset comparisons reflect both documentation and construct alignment differences as well as potential item difficulty differences.

2.3. Material Curation

Documentation consisted of official codebooks for each dataset (Prihar and Gonsalves, 2021; Fahle et al., 2024). Research papers included all publications using each dataset identified through citation tracking (ASSISTments: 5 papers; SEDA: 23 papers). The difference in corpus size reflects the datasets' research communities. Synthetic data samples were 50-row random samples from each dataset, preserving variable names, data types, and representative values to convey structural and relational information without revealing sensitive individual-level data.

2.4. Knowledge Basis

These sources are supplied to the model via Retrieval-Augmented Generation (RAG). Originally proposed by Lewis et al. (2020), RAG architectures *augment* generative modules with a module that retrieves relevant information from a corpus in order to ground generation in retrieved text rather than in parametric memory alone. Our modeling configurations (described below) differ only in which sources are retrievable. To make the types of knowledge described in section 2.2 concrete, consider a single construct—student persistence in ASSISTments. The codebook supplies *declarative* knowledge that names and defines the relevant variable, `attempt_count` (Variable Identification); prior papers supply the *procedural* convention of reading repeated attempts as persistence, a link the codebook omits (Usage Appropriateness); and a synthetic sample supplies the *structural* knowledge—the integer type and the identifiers joining the ten tables—needed to write correct code (Code Generation). These mappings denote each source's predominant, not exclusive, contribution. We therefore interpret our configurations as testing each source's *marginal* contribution given documentation, consistent with our finding that source value is interactive rather than additive.

2.5. Architectural Configurations

We tested four RAG configurations representing systematic combinations of knowledge types:

- **Configuration A (Baseline):** Documentation only, testing whether declarative knowledge alone suffices for discovery.
- **Configuration B:** Documentation + research papers, adding procedural knowledge about usage conventions.
- **Configuration C:** Documentation + synthetic data, adding structural knowledge about data formatting and relationships.
- **Configuration D (Full Triangulation):** All three sources combined.

We compared two baseline systems to contextualize whether documentation-only performance was broadly similar across AI environments: GPT-4o (chunked RAG with standard embedding-based retrieval, 4,096-token chunks, top-k = 5) and NotebookLM (long-context RAG by Google that processes entire document sets within its extended context window). Configurations B through D used NotebookLM to hold the AI environment constant while varying knowledge sources. Thus, model comparisons should be interpreted as contextual rather than as a fully crossed test of model-by-source interactions.

2.6. Statistical Analysis

We used the Friedman test (the nonparametric analog of repeated-measures ANOVA), treating the 50 items as repeated observations. We reported Kendall's W as the effect size and computed overall and within each capability domain. When the Friedman test was significant, we used Nemenyi tests for post-hoc pairwise comparisons, evaluated at $\alpha = .05$; the Nemenyi procedure controls the familywise error rate across all pairs. Our focal contrasts were full triangulation versus baseline (D vs. A) and papers versus data (C vs. B).

3. Results

3.1. ASSISTments

Table 1 shows that the full triangulation model achieved the highest overall performance (89%), compared with GPT-4o baseline (67%), NotebookLM baseline (79%), doc + papers (75%), and doc + data (65%) (RQ1). The Friedman test indicated significant overall differences across models, $\chi^2(4) = 22.19$, $p < .001$, Kendall's W = 0.11. Adjusted Nemenyi post-hoc comparisons showed that the full triangulation model significantly outperformed both GPT-4o baseline (p = .033) and doc + data (p = .047); no other pairwise comparisons reached statistical significance. Kendall's W was largest for Code (W = 0.45), suggesting the strongest model differentiation in that domain (Table 1). Because domain-specific analyses were based on only 10 items per domain, these domain-level findings should be interpreted as exploratory.

3.2. SEDA

On SEDA, configuration again differed significantly ($\chi^2(4) = 37.02$, $p < .001$, Kendall's W = 0.19), with the strongest differentiation in Code (W = 0.72; table 2) (RQ1). Performance

Domain	GPT-4o	NLM A	NLM B	NLM C	NLM D	W
Hypotheses	9/10	8/10	7/10	6/10	8.25/10	0.14
Code	5/10	9/10	7.5/10	8.5/10	10/10	0.45
Usage	6/10	8/10	7.5/10	4/10	9/10	0.27
Methodology	6.5/10	6.5/10	7.5/10	6/10	8.5/10	0.21
Variable ID	7/10	8/10	8/10	8/10	8.5/10	0.07
Total	33.5/50	39.5/50	37.5/50	32.5/50	44.25/50	0.11
% Mean (SD)	0.67 (0.31)	0.79 (0.35)	0.75 (0.35)	0.65 (0.39)	0.89 (0.23)	

Table 1: Quality scores by domain and RAG configuration with ASSISTments data

fell into two tiers. The lower tier—comprising the two baselines and doc + papers configuration—showed no significant differences among them (73–81%). The upper tier—doc + data and full triangulation—achieved 92% and 90% respectively and significantly exceeded the lower tier (all $p < .001$). Notably, adding papers provided no benefit over baseline documentation, consistent with the interpretation that SEDA's comprehensive documentation already incorporates the procedural knowledge that papers would otherwise supply.

Domain	GPT-4o	NLM A	NLM B	NLM C	NLM D	W
Hypotheses	6.5/10	7/10	7.5/10	7/10	7/10	0.13
Code	5.5/10	5/10	5/10	10/10	9/10	0.72
Usage	8/10	8.5/10	8/10	9/10	9.5/10	0.17
Methodology	9/10	10/10	10/10	10/10	10/10	0.20
Variable ID	7.5/10	8/10	10/10	10/10	9.5/10	0.26
TOTAL	36.5/50	38.5/50	40.5/50	46/50	45/50	0.19
% Mean (SD)	0.73 (0.31)	0.77 (0.29)	0.81 (0.27)	0.92 (0.19)	0.90 (0.20)	

Table 2: Quality scores by domain and RAG configuration with SEDA 5.0 data

3.3. Cross-Dataset Comparison

The utility of added knowledge sources varies with dataset characteristics (RQ2). For Code generation, adding synthetic data improved 8 of 10 SEDA items but zero ASSISTments items showed a "data only helped" pattern. For Usage, adding data without papers sharply reduced ASSISTments performance (-0.40 per item) (Figure 1) but was mildly beneficial for SEDA (+0.05 per item). These opposing patterns were initially attributed to documentation quality: SEDA's extensive documentation makes added data safely interpretable, while ASSISTments' codebook-oriented documentation leaves interpretive gaps. However, the two datasets also differ in construct alignment: SEDA's variables were designed to measure the constructs its usage items address (achievement gaps, educational opportunity), while ASSISTments' usage items require connecting behavioral logs to psychological constructs beyond the platform's design purpose (self-regulation, cognitive effort).

3.4. Model Comparison

At baseline the two systems were comparable on SEDA (73% vs. 77%), with NotebookLM ahead on ASSISTments (67% vs. 79%). Because the augmented configurations were implemented in NotebookLM, these comparisons do not constitute a fully crossed model comparison. They do, however, suggest that information architecture may be at least as consequential as model choice for this class of discovery tasks (RQ2).

Usage item: "Describe the limitations of using opportunity zone status to compute achievement gaps."

Documentation + data (C), scored 0: "Not Found. The sources outline that `opportunity_zone` is a categorical status (Yes, No, or Contiguous), but they do not discuss the limitations of using this status to compute achievement gaps."

Full triangulation (D), scored 1: "A major limitation of using opportunity zone status (found in `ddets`) to compute achievement gaps is that the dataset has no knowledge of individual students' demographic information, race, gender, home environments, socioeconomic status, or state test scores. Therefore, researchers can only calculate coarse, district-level differences rather than individual achievement gaps."

Figure 1: One ASSISTments *Usage Appropriateness* item under two configurations. Documentation alone also answered correctly (score 1); adding a synthetic data sample without papers (C) caused the model to retreat to a non-answer, whereas full triangulation (D) recovered a correct, usable judgment. The failure is not a hallucinated fact but a shift in reasoning—relevant-but-incomplete context made the model more conservative.

4. Discussion

4.1. Theoretical Contributions

Our results support the hypothesis that information architecture shapes AI-assisted data discovery: performance depended on how the available knowledge sources were composed, not on the base model alone. This connects our study to two existing ideas and shows what they look like in practice. First, recent RAG research treats retrieval as a designed, reconfigurable component whose composition drives performance, rather than a fixed pipeline attached to a model (Gao et al., 2023; Singh et al., 2025); our finding that a source's value depends on which other sources accompany it, helping in some combinations and harming in others, is a concrete instance of that view. Second, Simon's (1956) principle that intelligent behavior arises from the fit between an agent's capabilities and the structure of its environment has a direct analog here: a discovery system's performance depends on the information environment assembled around the model, so a well-composed environment can partly compensate for limitations in model sophistication.

Most notably, adding individual knowledge sources to documentation can *degrade* performance when complementary sources are absent. Adding synthetic data without papers reduced ASSISTments Usage performance from 80% to 40%, a substantial decline that full triangulation reversed to 90%. Figure 1 illustrates this on a single item: documentation alone answered correctly, adding a data sample without papers produced a non-answer, and full triangulation restored a correct response. Similarly, adding papers without data reduced ASSISTments Code performance by an average of 0.15 points per item, hurting 2 of 10 items while helping none; full triangulation achieved 100% on the Code domain. Our results suggest that relevant but incomplete context can be worse than less context—not because the model generates incorrect facts, but because partial context changes how the model reasons.

The modest aggregate Kendall's W values (0.11–0.19) reflect the concentration of architecture effects in specific domains and items rather than weak effects overall. Code generation showed the strongest differentiation, while ceiling effects in Methodology (6 of

10 ASSISTments items and all 10 SEDA items invariant across configurations) reduced detectable variation. These results suggest that domain-level evaluation is necessary; aggregate metrics would hide both the large Code gains and the Usage losses that together drive our main result. This is consistent with broader RAG evaluation work that assesses retrieval and generation along multiple dimensions rather than relying on a single aggregate accuracy score (Gao et al., 2023); our results extend that logic to the level of individual capability domains.

Framed as context engineering, our results show that performance depends on the full information environment a model is given: well-composed context helps, while poorly composed context can be worse than minimal context (Anthropic, 2025).

4.2. Practical Implications

These findings suggest practical ways to broaden access to complex educational archives. Multi-source discovery tools can lower the apprenticeship barrier that disadvantages novice researchers and under-sourced institutions, and because information architecture appears to matter as much as model sophistication, they can be built with widely available LLMs and careful knowledge curation rather than specialized compute. None of the configurations we tested requires fine-tuning, training compute, or dedicated hardware, so the binding constraint is not computational but informational: the quality and curation of the available knowledge sources (documentation, research literature, and data samples) and the domain expertise to assemble them. This is timely because federal funding agencies increasingly require data management and sharing plans that promote data accessibility and reuse (Institute of Education Sciences, 2024). Without tools that help researchers navigate archives independently, expanded sharing risks an "availability without accessibility" paradox—data that is technically open but practically unusable. This concern is concrete for emerging multi-platform research infrastructures. SafeInsights, an NSF-funded national hub that securely connects many digital learning platforms (including ASSISTments) so researchers can study learning across them while protecting student privacy (SafeInsights, 2024), will give researchers access to numerous unfamiliar archives at once. In such settings, AI-assisted discovery support and well-composed information architecture are what turn technically available data into practically usable data.

The results also carry implications for the organizations that build and publish archives. First, archives should prioritize synthetic data generation. Sample datasets yielded substantial gains, particularly for code generation, at minimal cost, making them a high-return investment. Second, archives should invest in comprehensive documentation, which improves baseline performance and is valuable independent of other sources. Third, archives should anticipate uses beyond a dataset's design purpose. When the anticipated construct gap is large, variable definitions are not enough, and archives need curated examples—paper collections or explicit construct-mapping sections—that show how variables have actually been used, not only what they are.

For practitioners deploying discovery tools, full triangulation is the safest default: it was the only configuration never observed to harm performance, and it was best or tied for best on both archives. When full triangulation is infeasible, the best single addition depends on the documentation. With technical-methodological documentation (SEDA),

documentation alone already performs well and papers add little, so a synthetic data sample is the higher-return addition. With codebook-oriented documentation (ASSISTments), no single addition is safe: adding a synthetic data sample without papers sharply reduced usage-appropriateness accuracy. Here the choice is all-or-nothing, full triangulation or documentation alone, rather than a partial augmentation. In short, with rich documentation a single data sample is a safe, high-return addition, but with thin documentation a source should be added only together with its complement.

4.3. Limitations

Several limitations bound our conclusions. First, we study only two U.S.-based archives, although they were selected to contrast documentation quality, they also differ in subject matter, granularity, data structure, and research community size. Because assessment items were developed independently for each dataset and were not designed to be parallel in difficulty, cross-dataset differences in augmentation effects may partly reflect item difficulty differences alongside documentation and construct-alignment differences.

Second, Configurations B-D were implemented only in NotebookLM, so architecture effects cannot be fully separated from model-specific effects; replicating B-D on a second long-context system is the key next step. We also characterize documentation quality through observable features rather than a validated instrument, leaving open which specific features most affect baseline performance.

Finally, discrete evaluation items do not fully capture iterative, exploratory discovery, and our outcome is expert-scored output quality rather than whether novice users actually discover variables more effectively with these tools. Planned user studies will measure real-world discovery efficiency, confidence calibration, and differential benefits by researcher experience.

5. Conclusion

This study suggests that the semantic gap in educational data discovery is partly an information architecture problem, not solely a model capability problem. Our results indicate that context composition is interactive rather than additive: adding individual knowledge sources can both help and harm performance depending on the information environment. By applying cognitive psychology's knowledge taxonomy to RAG design (triangulating declarative, procedural, and structural knowledge) and identifying when each source helps or harms, we offer a practical route toward more accessible and equitable data-driven educational research. Future work should test whether these architecture effects, and the role of construct alignment, transfer to additional archives and whether they improve discovery outcomes for actual researchers.

Data and Code Availability.

The evaluation items, model prompts, scoring rubrics, and analysis scripts are available upon request. Synthetic data samples are released in full; the ASSISTments and SEDA archives are available from their original providers under their respective terms.

Acknowledgments

This work was supported by the Institute of Education Sciences (IES), U.S. Department of Education, through SEERNet Grant R305N210034 and National Science Foundation (NSF), through Grant 2153481. Any opinions, findings, and conclusions or recommendations expressed are those of the authors and do not necessarily reflect the views of the IES, the U.S. Department of Education, or NSF.

References

John R. Anderson. ACT: A simple theory of complex cognition. *American Psychologist*, 51 (4):355–365, April 1996. ISSN 1935-990X, 0003-066X. doi: 10.1037/0003-066X.51.4.355.

Anthropic. Effective context engineering for AI agents. https://www.anthropic.com/engineering/effective-context-engineering-for-ai-agents, September 2025.

Alex J. Bowers and Yeonsoo Choi. Building School Data Equity, Infrastructure, and Capacity Through FAIR Data Standards: Findable, Accessible, Interoperable, and Reusable. *Educational Researcher*, 52(7):450–458, October 2023. ISSN 0013-189X, 1935-102X. doi: 10.3102/0013189X231181103.

Erin M. Fahle, Jim Saliba, Demetra Kalogrides, Benjamin R. Shear, Sean F. Reardon, and Andrew D. Ho. Stanford education data archive: Technical documentation (version 5.0). Technical report, Stanford University, 2024. URL https://purl.stanford.edu/cs829jn7849.

G. W. Furnas, T. K. Landauer, L. M. Gomez, and S. T. Dumais. The vocabulary problem in human-system communication. *Communications of the ACM*, 30(11):964–971, November 1987. ISSN 0001-0782, 1557-7317. doi: 10.1145/32206.32212.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. Retrieval-Augmented Generation for Large Language Models: A Survey, 2023.

Shailja Gupta, Rajesh Ranjan, and Surya Narayan Singh. A Comprehensive Survey of Retrieval-Augmented Generation (RAG): Evolution, Current Landscape and Future Directions, 2024.

Institute of Education Sciences. Education research grants program request for applications, 2024.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, 55(12):1–38, December 2023. ISSN 0360-0300, 1557-7341. doi: 10.1145/3571730.

Enkelejda Kasneci, Kathrin Sessler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier,

Stephan Krusche, Gitta Kutyniok, Tilman Michaeli, Claudia Nerdel, Jürgen Pfeffer, Oleksandra Poquet, Michael Sailer, Albrecht Schmidt, Tina Seidel, Matthias Stadler, Jochen Weller, Jochen Kuhn, and Gjergji Kasneci. ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103:102274, April 2023. ISSN 10416080. doi: 10.1016/j.lindif.2023.102274.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, pages 9459–9474, Red Hook, NY, USA, December 2020. Curran Associates Inc. ISBN 978-1-7138-2954-6.

Jessica A. R. Logan, Sara A. Hart, and Christopher Schatschneider. Data Sharing in Education Science. *AERA Open*, 7:23328584211006475, January 2021. ISSN 2332-8584, 2332-8584. doi: 10.1177/23328584211006475.

Ethan Prihar and Manuel Gonsalves. ASSISTments 2020-2021 School Year Dataset, November 2021.

David E. Rumelhart. Schemata: The Building Blocks of Cognition. In *Theoretical Issues in Reading Comprehension*. Routledge, 1980.

SafeInsights. SafeInsights: A national research infrastructure for education. https://www.safeinsights.org/, 2024. NSF Cooperative Agreement No. 2153481.

H. A. Simon. Rational choice and the structure of the environment. *Psychological Review*, 63(2):129–138, 1956. ISSN 1939-1471, 0033-295X. doi: 10.1037/h0042769.

Aditi Singh, Abul Ehtesham, Saket Kumar, Tala Talaei Khoei, and Athanasios V. Vasilakos. Agentic Retrieval-Augmented Generation: A Survey on Agentic RAG, 2025.

Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino Da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J.G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A.C 'T Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene Van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan Van Der Lei, Erik Van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1):160018, March 2016. ISSN 2052-4463. doi: 10.1038/sdata.2016.18.