Project Notes:

Project Title:
Name:

Note Well: There are NO SHORT-cuts to reading journal articles and taking notes from them. Comprehension is paramount. You will most likely need to read it several times, so set aside enough time in your schedule.

Contents:

Knowledge Gaps:

This list provides a brief overview of the major knowledge gaps for this project, how they were resolved and where to find the information.

Knowledge Gap	Resolved By	Information is located	Date resolved
Tuning large language models	Watching YouTube videos	Video links in project logbooks	9/29/24
Privacy of data	Reading patents	Patents at bottom of project notes	10/10/24
Limitations and effectiveness of AI in math education	Journal articles	First few articles up to article 7 in project notes	9/27/24

Literature Search Parameters:

These searches were performed between (Start Date of reading) and XX/XX/2019.
List of keywords and databases used during this project.

Database/search engine	Keywords	Summary of search
Gordon Library	AI and (tutor* or teach*)	Found most of my articles from this and following up with references
Google patent search	Artificial intelligence in education	Found next to nothing useful
Google patent search	Artificial Intelligence	Found general patents related to data collection and privatization in AI

Tags:

Tag Name	

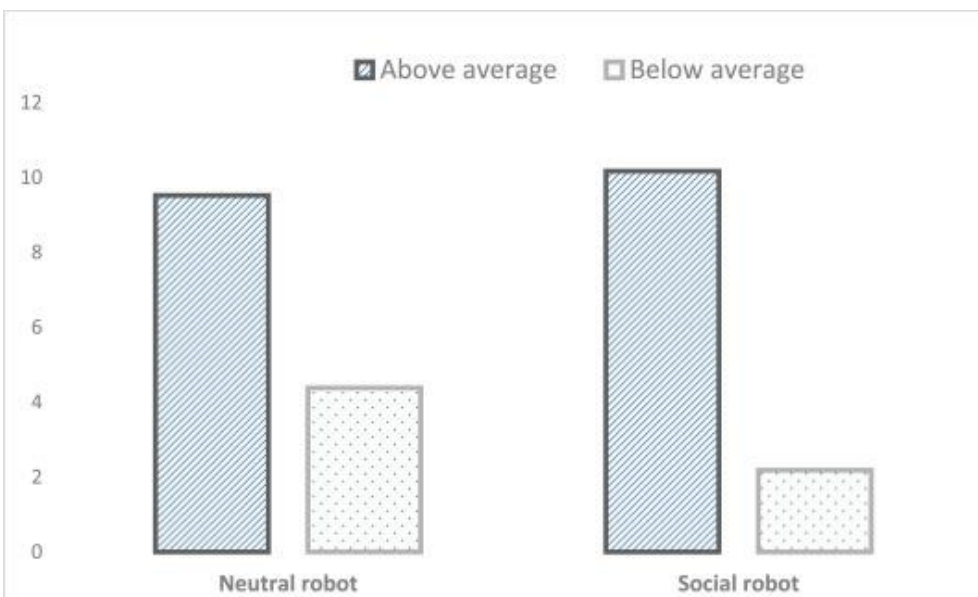
Article #1 Notes: Title

Article notes should be on separate sheets

KEEP THIS BLANK AND USE AS A TEMPLATE

Source Title	
Source citation (APA Format)	
Original URL	
Source type	
Keywords	
#Tags	
Summary of key points + notes (include methodology)	
Research Question/Problem/ Need	
Important Figures	
VOCAB: (w/definition)	
Cited references to follow up on	
Follow up Questions	

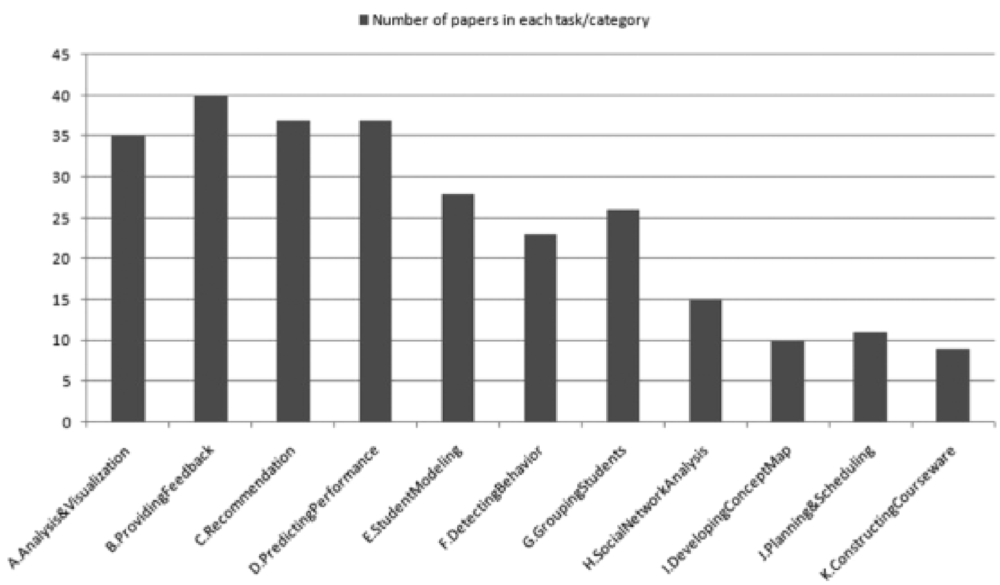
Article #1 Notes:

Source Title	Robot tutor and pupils' educational ability: Teaching the times tables
Source citation (APA Format)	Konijn, E. A., & Hoorn, J. F. (2020). Robot tutor and pupils' educational ability: Teaching the times tables. *Computers and Education*, *157*, 103970-. https://doi.org/10.1016/j.compedu.2020.103970
Original URL	https://doi.org/10.1016/j.compedu.2020.103970
Source type	Journal Article
Keywords	Robot tutor, tutoring, social robots
#Tags	
Summary of key points + notes (include methodology)	Physically present robots have been shown to be quite effective in tutoring, even compared to actual teachers. This study assessed the improvement of multiplication table recall in 86 elementary school students. It found that the robots were shown to be capable of improving students' scores regardless of their behavior (social or neutral), but ethics and other factors are still a concern.
Research Question/Problem/ Need	How effective are robot tutors in improving elementary school kids' multiplication skills?
Important Figures	 Y-axis is score improvement on test. Neutral robot was better for below average

	students while robot type didn't seem to matter much for advanced students. However, advanced students benefited the most from robots in general.
VOCAB: (w/definition)	Pedagogical – related to teaching/learning
Cited references to follow up on	Leyzberg, D., Spaulding, S., & Scassellati, B. (2014). Personalizing robot tutors to individuals' learning differences. In Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction (pp. 423–430). ACM. https://doi.org/10.1145/2559636.2559671.
Follow up Questions	1. How would students of different age groups compare in their response to robot tutoring? 2. How can teachers and robots work in tandem to maximize success in the classroom? 3. Why do students not respond any better to more supportive feedback from robots?

Article #2 Notes:

Source Title	Educational Data Mining: A Review of the State of the Art
Source citation (APA Format)	Romero, C. & Ventura, S. (2010). Educational Data Mining: A Review of the State of the Art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *40*(6), 601-618. https://doi.org/10.1109/TSMCC.2010.2053532
Original URL	https://doi.org/10.1109/TSMCC.2010.2053532
Source type	Journal article
Keywords	Educational Data Mining
#Tags	

Summary of key points + notes (include methodology)	The article went over EDM itself, including defining it, reviewing current research, and looking at promising future research. Past research in EDM mainly looked at predicting student performance. EDM has risen in popularity with the advent of LMSs (learning management systems) creating lots of data online about student-teacher interaction and education in general. Educational data mining has been used to visualize data to help educators and administrators, analysis of that data, provide feedback to teachers, and help students, as well as predicting their success, creating models of students, and detecting unwanted behaviors (cheating, dropping out, etc.). Also, grouping students to aid in personalization, social network analysis to help students find relevant information, concept maps to help teachers teach, creating course material, and scheduling. Future work in EDM would include making it more accessible, integrating it with LMSs, standardization, and fine-tuning data mining to be more education specific.
Research Question/Problem/ Need	How is educational data mining used and how does it impact students and teachers?
Important Figures	 There are a variety of use cases for educational data mining, y axis is number of papers relating to each topic, out of the 300 total that were reviewed
VOCAB: (w/definition)	Educational data mining – a field of data science that uses a variety of data analysis and machine learning techniques to solve problems in the world of educational research Learning management system – An online platform that connects students and teachers and allows teachers to assign materials and track student grades and progress (think Schoology or Canvas)
Cited references to follow up on	

Follow up Questions	How well does this analysis hold up in the current state of artificial intelligence? How much does EDM have on the seemingly less intuitive use cases such as scheduling and creating content maps (shouldn't teachers already be good at those things)? Why didn't constructing courseware have much research into it when it is likely a huge benefit to EDM as it would take lots of workload off teachers?

Article #3 Notes: Title

Source Title	**Music teachers' labeling accuracy and quality ratings of lesson plans by artificial intelligence (AI) and humans**
Source citation (APA Format)	Cooper, P. K. (2024). Music teachers' labeling accuracy and quality ratings of lesson plans by artificial intelligence (AI) and humans. *International Journal of Music Education*, *0*(0). https://doi.org/10.1177/02557614241249163
Original URL	https://doi.org/10.1177/02557614241249163
Source type	Journal article
Keywords	Lesson plan, music education
#Tags	
Summary of key points + notes (include methodology)	A survey was sent out to US music teachers. Most of them had over ten years of experience and more than half of them had experience with AI, implying that they would be suitable participants in the study. Overall, the teachers were 55% accurate on average in labeling whether lesson plans were generated by AI or humans. This was not a statistically significant result. Also, AI generated content and human content were ranked similarly in usefulness. Using multiple regression, they did find that they could predict the accuracy of a teacher's guess based on their personal use of AI, their ratings of usefulness for both humans and AI, and how useful they thought AI would be in the future. The analysis found that overall, teachers were unsuccessful in predicting whether music lesson plans were generated by AI much better than chance.
Research Question/Problem/ Need	How well can AI generated music lesson plans be distinguished from human made ones?
Important Figures	

VOCAB: (w/definition)	Intelligent Tutor System – a program that individually tutors a student in a custom manner, similarly to a human tutor
Cited references to follow up on	
Follow up Questions	What, specifically, do these lesson plans entail? Can this be taken one step further, using AI to assign homework and tests? Can these lesson plans be modified on the fly as is so common in private lessons by AI?

Video #1 Notes:

Source Title	Introduction to Generative AI
Source citation (APA Format)	Google Cloud Tech. (2023, May 8). *Introduction to Generative AI* [Video]. YouTube. https://youtu.be/G2fqAlgmoPo?si=KH73Mt7LrbO5ryt5
Original URL	https://youtu.be/G2fqAlgmoPo?si=KH73Mt7LrbO5ryt5
Source type	YouTube video
Keywords	Generative AI
#Tags	
Summary of key points + notes (include methodology)	Generative AI is a type of artificial intelligence that creates content Artificial intelligence is a field of computer science, while machine learning is the subfield that involves creating models that can perform “intelligent” tasks Neural networks – layers of “neurons” (nodes) that make up a deep learning model and can use labeled or unlabeled data to process patterns (semisupervised learning) Transformers: consist of encoders and decoders that convert input data into relevant output data, transformers sometimes make hallucinations (incorrect outputs) Prompt design – creating a prompt that gives the desired output from a generative AI model Variety of model types: text-to-text, text-to-task, text-to-image, foundational Foundation models can be fine-tuned to a variety of tasks Generative AI Studio – Google's gen AI platform that allows developers to create generative AI. Has a library of pre-trained models and has tools for fine-tuning, deployment, and more Generative AI App Builder – create gen AI apps without code (drag and drop, might be limited)
Research Question/Problem/ Need	N/A
Important Figures	
VOCAB: (w/definition)	Neural networks, transformers (definitions in summary) Machine learning – models that can “learn” by changing their parameters and the connections in their neural networks to create more desirable outputs
Cited references to follow up	

on	
Follow up Questions	What are the different methods for fine tuning? How does prompt design work? Is that different from prompt engineering?

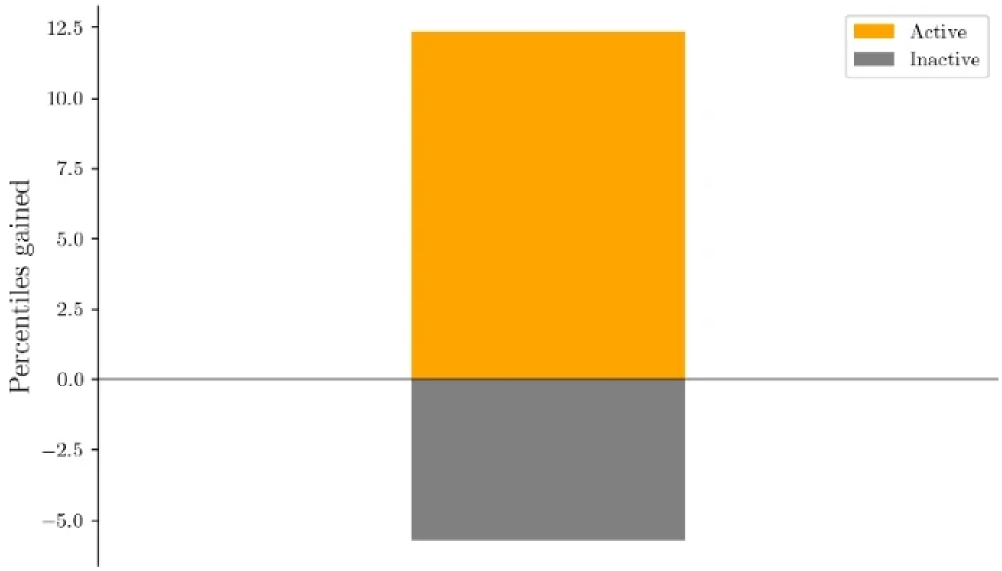
Video #2 Notes:

Source Title	Introduction to large language models
Source citation (APA Format)	Google Cloud Tech. (2023, May 8). *Introduction to Large Language Models* [Video]. YouTube. https://youtu.be/zizonToFXDs?si=SJnO9BN6-vkOfc0Q
Original URL	https://youtu.be/zizonToFXDs?si=SJnO9BN6-vkOfc0Q
Source type	YouTube video
Keywords	Large language model
#Tags	
Summary of key points + notes (include methodology)	LLMs are general models that can be fine tuned to specific use cases Tuned usually using domain specific data which is in a much smaller quantity than the data used to create the general model Parameters – the "knowledge" that the model gathered from the data. LLMs usually have many parameters I probably don't have the resources required to create a LLM (need tons of data), however, I could probably get the data required to tune one to educational purposes LLM performance is increasing over time LLM development does not require as much coding knowledge as regular machine learning development Prompt engineering differs from prompt design in that the goal is to improve the performance of the model, may use known effective keywords or give examples of the correct output
Research Question/Problem/ Need	N/A
Important Figures	
VOCAB: (w/definition)	Fine-tuning – modifying a pre-trained large language model to be more suited to a specific domain or field Parameters – act as guidelines that affect a model's output and are determined during the pre-training process. LLMs often have millions or even billions of parameters
Cited references to follow up	

on	
Follow up Questions	Need more information on fine-tuning.

Article #4 Notes:

Source Title	Effective learning with a personal AI tutor: A case study
Source citation (APA Format)	Baillifard, A., Gabella, M., Lavenex, P. B., Martarelli, C. S. (2024). Effective learning with a personal AI tutor: A case study. *Educ Inf Technol*. https://doi.org/10.1007/s10639-024-12888-5
Original URL	https://doi.org/10.1007/s10639-024-12888-5
Source type	Journal Article
Keywords	AI Tutor, AIEd, Learning Sciences, Personalization, Intelligent Tutoring Systems
#Tags	
Summary of key points + notes (include methodology)	Known benefits of AI in education: 1. Capable of predicting student outcomes and creating profiles of students 2. Good at testing students, taking workload off teachers 3. Can be personalized to help a wider range of students more effectively 4. Intelligent Tutoring Systems that simulate real 1:1 tutoring experiences Methodology: They used a tutoring app developed by MAGMA Learning that uses personalized tutoring. Also, they used GPT-3 to create a set of questions that would be used. A neural network was used to predict the probability of a student answering a question correctly (called the "grasp") and thus selected the best questions for the student to practice with. App was tested in an online neuroscience college course. Class was managed through an LMS called Moodle. There was also a parallel course taken by most of the same students at the same time, but they did not have the app for that course. 43 of the 51 students enrolled in the course did use the app, students could use the app as they pleased. 47 students took both final exams and 40 of those were using the app. They investigated performance on the final exam and use of the app. Also, performance on Moodle and the "grasp" compared to the final exam. They also compared these to the course with the final exam but no app used. Comparing active users to inactive users: Average increase in score of 0.71 for active user (test was on a scale from 1-6). They used different thresholds for "active" but found this to be the average increase, on average Comparing the two courses: Active users had an average grade increase of 12.4% in the course where the app was present compared to the parallel course. Inactive users had a decrease of 5.7%.

	Active users of Moodle were not shown to do any better on the final exam than non-Moodle users. There was also a strong correlation between grasp prediction and exam grade. Overall, app usage and grade were positively correlated.
Research Question/Problem/ Need	How well does AI in education work with known learning sciences?
Important Figures	 Active users did better on the final exam than inactive users. Inactive users even seemed to do worse on the course that offered the AI app as a study option.
VOCAB: (w/definition)	Retrieval practice – recalling information from memory without having it available to help you remember it Natural language processing – using machine learning to allow computers to understand and create human language
Cited references to follow up on	
Follow up Questions	How can be sure it was causation and not just correlation (ambitious students would happen to use the app more and have higher grades)? How were the neural networks trained? How important is the personalization aspect of the AI tutor itself, rather than just having the tutor?

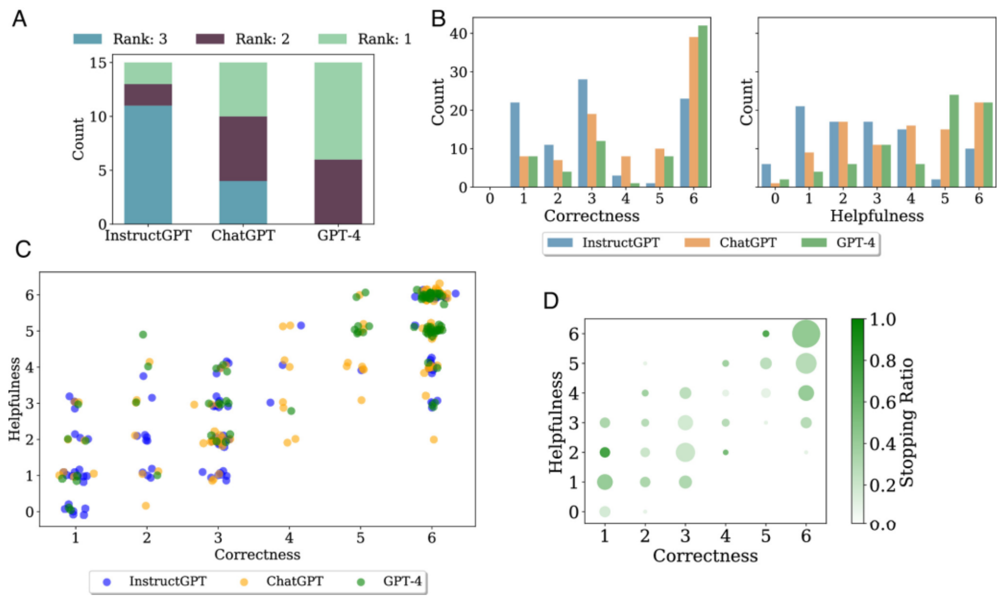
Article #5 Notes:

Source Title	Editorial Note: From Conventional AI to Modern AI in Education: Re-examining AI and Analytic Techniques for Teaching and Learning
Source citation (APA Format)	Xie, H., Hwang, G. J., & Wong, T. L. (2021). Editorial Note: From Conventional AI to Modern AI in Education: Re-examining AI and Analytic Techniques for Teaching and Learning. *Educational Technology & Society*, *24*(3), 85-88. https://doi.org/10.30191/ETS.202107_24(3).0006
Original URL	https://doi.org/10.30191/ETS.202107_24(3).0006 https://link.gale.com/apps/doc/A668399451/AONE?u=mlin_c_worpoly&sid=bookmark-AONE&xid=6fa17e23 (need this link because otherwise I can't access)
Source type	Journal Article
Keywords	Modern AI, AI transformation, Deep neural networks, Analytic techniques
#Tags	
Summary of key points + notes (include methodology)	It's an editorial note for the issue itself which discusses technology in education Modern AI uses deep neural networks, while traditional AI uses statistical models. There is limited research on modern AI's use in education as most of it in the past has been with traditional AI. Teachers and AI developers don't know much about each other's domains, so it is hard to connect the two for effective AI education. Precision education is the next big step in the use of AI in education, along with more general predictions, and using AI for new apps.
Research Question/Problem/ Need	How effective has artificial intelligence been in education thus far?
Important Figures	
VOCAB: (w/definition)	Convolution neural network – neural networks that are better at using image and audio inputs (What are Convolutional Neural Networks? | IBM) Generative adversarial network – made up of a generator and discriminator neural network, the generator attempts to create data identical to training data until it can fool the discriminator; used for unsupervised learning (Generative Adversarial Network (GAN) - GeeksforGeeks) Precision education – identifying students who are at risk of failure, dropping out, etc., and giving them the guidance and resources needed to succeed accordingly

Cited references to follow up on	Chen, X., Xie, H., & Hwang, G. J. (2020a). A Multi-perspective study on Artificial Intelligence in Education: grants, conferences, journals, software tools, institutions, and researchers. Computers and Education: Artificial Intelligence, 1, 100005. - like the current article in that the references in this one may be more useful than the information itself; may lead to studies or developments more pertinent to my topic Yang, S. J., Ogata, H., Matsui, T., & Chen, N. S. (2021). Human-centered artificial intelligence in education: Seeing the invisible through the visible. Computers and Education: Artificial Intelligence, 2, 100008. Wang, J., Xie, H., Wang, F. L., Lee, L. K., & Au, O. T. S. (2021). Top-n personalized recommendation with graph neural networks in MOOCs. Computers and Education: Artificial Intelligence, 2, 100010. Almohammadi, K., Hagras, H., Alghazzawi, D., & Aldabbagh, G. (2016). Users- centric adaptive learning system based on interval type-2 fuzzy logic for massively crowded E-learning platforms. Journal of Artificial Intelligence and Soft Computing Research, 6(2), 81-101.
Follow up Questions	Three years later, how true do these gaps hold up? Is the gap between AI experts and educators specific to the education field only, or does this problem exist across many domains? Why has predicting students' risk or classifying students been so researched compared to other needs in education?

Article #6 Notes:

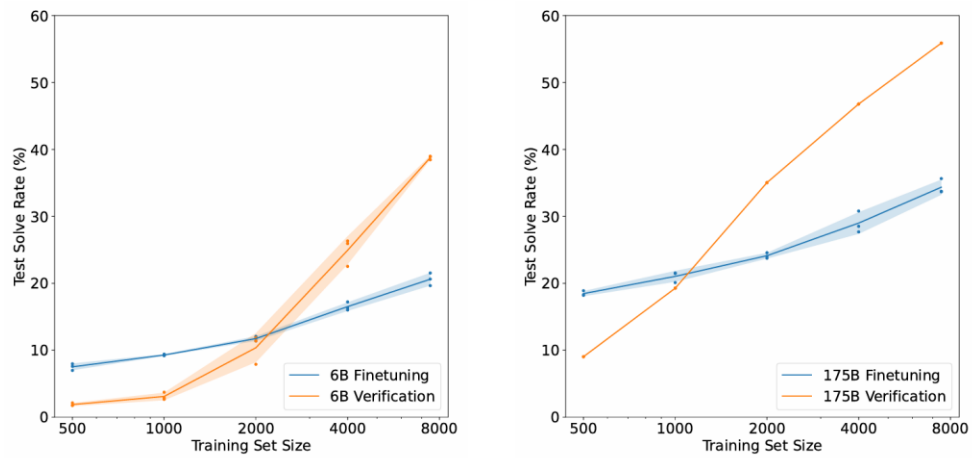
Source Title	Evaluating language models for mathematics through interactions
Source citation (APA Format)	Collins, K. M., Jiang, A. Q., Frieder, S., Wong, L., Zilka, M., Bhatt, U., Lukasiewicz, T., Wu, Y., Tenenbaum, J. B., Hart, W., Gowers, T., Li, W., Weller, A., & Jamnik, M. (2024). Evaluating language models for mathematics through interactions. *Proceedings of the National Academy of Sciences of the United States of America*, *121*(24). https://doi.org/10.1073/pnas.2318124121
Original URL	https://doi.org/10.1073/pnas.2318124121
Source type	Journal article
Keywords	Dynamic (evaluation) - observing how people and language models interact over the course of an entire "conversation" rather than a snapshot evaluation
#Tags	
Summary of key points + notes (include methodology)	First, they developed a platform called CheckMate that allowed people to interact with LLM chatbots and rate them individually or comparatively. They tested how people used InstructGPT, ChatGPT, and GPT-4. For individually, they were allowed to use a model to help solve a math problem and then rate each step of the process. Comparatively, they ranked the different models without knowing which was which. Participants' experience ranged from undergraduate students to college professors, however data on participants was not collected beyond this. Specifically, they asked participants to prove undergraduate level theorems and allowed them to use AI any way they wished, as they wanted to see how people naturally used it. They were asked to rate perceived helpfulness along with mathematical correctness. They used dynamic evaluation – observing a model's entire interaction with a person – rather than static evaluation. GPT-4 was ranked the highest overall and received the highest helpfulness and correctness ratings. Models built for chatting (GPT-4 and ChatGPT) were ranked much better than those that aren't (InstructGPT). The correlation between helpfulness and correctness was decent but not 100% - sometimes it could be helpful but wrong (contains decent ideas) or correct but unhelpful (verbosity). Currently, measuring helpfulness and correctness cannot be done computationally or automatically, they must be determined by humans. This is one of the reasons humans were used to test these models. GPT-4 often struggled with arithmetic mistakes. In general, LLMs were found to struggle with **algebra, being too wordy, and reliance on memorized solutions**.
Research Question/Problem/ Need	How good are LLMs at assisting people with undergraduate level math problems?

<table>
<tr><td>Important Figures</td><td>

This figure shows that overall, GPT-4 was the best in every aspect.</td></tr>
<tr><td>VOCAB: (w/definition)</td><td>Taxonomize – to arrange a set into a classification</td></tr>
<tr><td>Cited references to follow up on</td><td>M. Lee, P. Liang, Q. Yang, “CoAuthor: Designing a human–AI collaborative writing dataset for exploring language model capabilities” in Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (2022), pp. 1–19.</td></tr>
<tr><td>Follow up Questions</td><td>Because they make so many arithmetic mistakes, could models like GPT-4 still be effective in teaching younger students?
If these models were helpful only some of the time, is it up to the human to determine when to use them, or is it feasible to improve them so that they are always helpful?
What does static evaluation of a LLM look like?</td></tr>
</table>

Article #7 Notes:

Source Title	**Training Verifiers to Solve Math Word Problems**
Source citation (APA Format)	Cobbe, K., Kosaraju, V., Bavarain, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C., Schulman, J. (2021). Training Verifiers to Solve Math Word Problems. arxiv.org/abs/2110.14168.
Original URL	https://doi.org/10.48550/arXiv.2110.14168
Source type	Article
Keywords	verifier
#Tags	
Summary of key points + notes (include methodology)	The researchers created GSM8K, a dataset of grade-school level math problems. It has natural language solutions rather than equations which would allow for better evaluation of large language models. They found that language models did not have high levels of accuracy on this dataset despite the problems lacking complexity. They hypothesized that using verifiers would increase this accuracy. They started by finetuning the models with various training set sizes. Unsurprisingly, models with more parameters and larger training sets had higher accuracy, when ran for 2 epochs. However, when allowed to run for more, allowing models to output 100 answers would cause their accuracy to eventually decrease (due to overfitting). However, this accuracy was still much higher than models that output only 1 answer. Next, they trained verifiers to output the probability that a model was correct. These models were trained on problems and solutions, but solutions could be labeled as correct even if the reasoning was wrong, as long as the final answer was right. They trained verifiers by finetuning a “generator” on the training data for 2 epochs, generating 100 solutions and labeling them as correct or incorrect, and then training a verifier on these solutions. They kept it at 2 epochs to maintain diversity in the data. They noted that it should be possible to combine the generator and verifier. They found that with high enough data sets, the models did

	much better with verifiers for both 6B and 175B parameter models.
Research Question/Problem/ Need	Can using verifiers increase performance of large language models on math problems?
Important Figures	 Verifiers drastically increase performance with enough data. A 6B parameter model with a verifier did better than the 175B parameter one without a verifier. Interestingly, verifiers decrease performance with small training sets
VOCAB: (w/definition)	Epoch – a complete pass of a model through an entire dataset
Cited references to follow up on	
Follow up Questions	Why were verifiers worse for smaller training sets? How did they fine tune the models? Why didn't they combine the verifier and generator if they said it was possible to do it?

Article #8 Notes:

Source Title	The comparison of general tips for mathematical problem solving generated by generative AI with those generated by human teachers
Source citation (APA Format)	Jia, J., Wang, T., Zhang, Y., & Wang, G. (2024). The comparison of general tips for mathematical problem solving generated by generative AI with those generated by human teachers. *Asia Pacific Journal of Education*, *44*(1), 8–28. https://doi.org/10.1080/02188791.2023.2286920
Original URL	https://doi.org/10.1080/02188791.2023.2286920
Source type	Journal article
Keywords	Intelligent tutoring system, large language models, prompt engineering
#Tags	
Summary of key points + notes (include methodology)	They decided to use prompt engineering on ChatGPT to see if it could effectively generate tips for solving math problems. They used zero-shot, one-shot, and few-shot learning with and without CoT. However, they only used each one twice (one tip for a geometry problem and one for an algebra problem, for a total of 12 tips). They also had teachers generate tips for the same problems to compare them. Then, they developed a rubric to score these tips and had people score AI-generated tips and teacher-made tips. They created an online survey which got 121 responses, most of which were from people with teaching experience. Participants had to score the 12 tips given per problem and decide which ones were made by ChatGPT (6 were per problem). On average, teacher-made tips were scored better. They ran t-tests on both the ratings from the geometry problem and those from the algebra problem and had a p-value of less than 0.05 on both, suggesting that teacher-made tips were better

than ChatGPT's tips.
However, they also found that participants could not differentiate the source of the tips.

Research Question/Problem/ Need

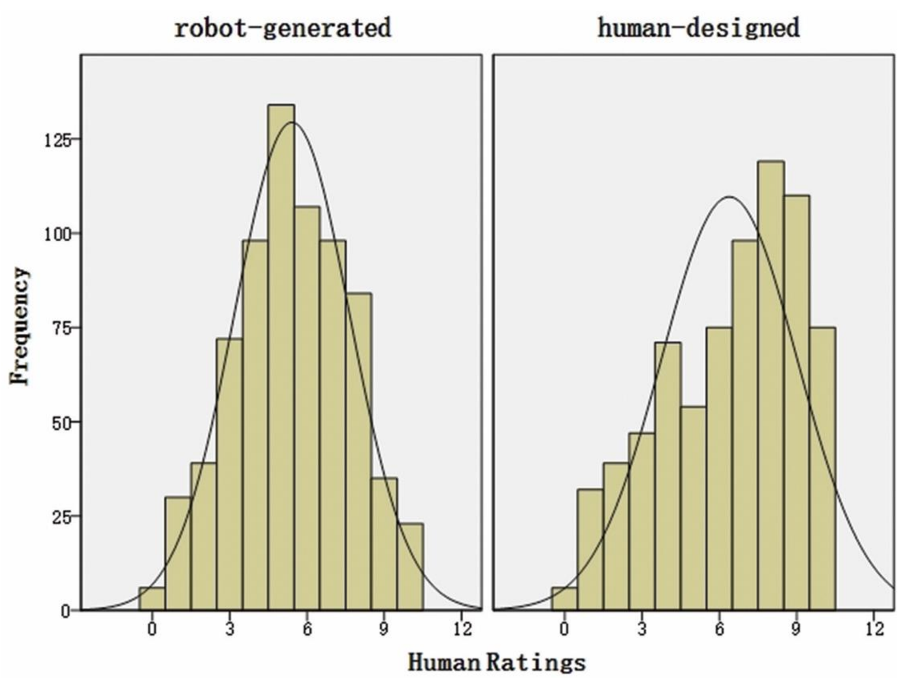
How do general tips for solving math problems generated by large language models differ from those generated by humans?

Important Figures

Table 2. Evaluation indicators and scores for general tips.

Primary Indicator	Secondary Indicator	Description	Score
Correctness	Correctness	The general tip correctly reflects the key steps or thinking process when solving the problem, without any logical or mathematical errors.	1
Expressiveness	Expressiveness	The general tip is clear, concise and coherent, without any grammatical errors, and uses the appropriate language and symbols.	1
Comprehensiveness	Comprehensiveness	If the content of general tip covers two aspects: method and knowledge that requires. It gets 1 point if it only covers one aspect, and 2 points if it covers both aspects.	2
Proceduralness	Clarify the problem	The general tip can help learners understand what the problem is asking.	1
	Make a plan	The general tip can help learners take into account all the given conditions and find the connection between the known conditions and the unknowns.	1
	Implement the plan	The general tip can help learners transform the original problem into an equivalent problem or a solved problem.	1
	Review	The general tip can summarize the knowledge points and methods used in this problem at the end, connecting common sense and theory.	1
Heuristicness	Heuristicness	The general tip can reflect the students' subjectivity and initiative, guide them to think and explore independently, rather than directly provide the answer or too simple hints. It can help learners apply the method to other problems and build their own knowledge network.	2
Total			10

The rubric that they had participants use to grade tips



Teacher-designed tips were rated higher on average than ChatGPT's tips

VOCAB: (w/definition)

Zero-shot learning – allowing the model to generate responses without any examples given
One-shot learning – allowing the model to generate responses with one specific

	example given Few-shot learning – using a few examples to allow the model to generate responses Chain of thought (CoT) - guiding the model to reason through to the desired output step-by-step (can be one-shot, few-shot, or even zero-shot)
Cited references to follow up on	
Follow up Questions	How can ChatGPT’s tips be improved to be as good as or even exceed teacher tips? Which prompt engineering method was the best at generating tips? How come people couldn’t differentiate the source of the tips if they could differentiate their quality?

Article #9 Notes:

Source Title	Dropout Prediction in MOOCs: Using Deep Learning for Personalized Intervention
Source citation (APA Format)	Xing, W., & Du, D. (2019). Dropout Prediction in MOOCs: Using Deep Learning for Personalized Intervention. *Journal of Educational Computing Research*, *57*(3), 547-570. https://doi.org/10.1177/0735633118757015
Original URL	https://doi.org/10.1177/0735633118757015
Source type	Journal article
Keywords	MOOC, dropout, deep learning,
#Tags	
Summary of key points + notes (include methodology)	MOOCs can have up to 90% attrition rates. This is usually just written off as a tradeoff for scale, but researchers wanted to look into solving it. They wanted to use deep learning because it would be impossible to manually look after all of these students. For their methodology they only investigated a single 8-week MOOC hosted on Canvas with 11 modules, 3617 students, 14 discussion forums, and 12 multiple choices. They tracked the various features listed in table 1. They started with 3 algorithms: K-nearest neighbors, support vector machines, and decision tree. Then they created a deep learning network (70% training data and 30% testing) which is different because it does automatic feature extraction and tuning. This allowed them to determine which students were most likely to drop out every week and thus plan personalized intervention. Deep learning model performed the best KNN doesn't give a probability so it couldn't be used Higher probability would indicate to teachers to give more intervention
Research Question/Problem/ Need	How can students that are at risk of dropping out receive personalized intervention?

Important Figures

Table 1. Features and Description.

Feature	Description
Number of announcements	Number of times students view the announcements
Number of assignments	Number of times students access the assignments
Number of calendar	Number of times students view the calendar
Number of module pages	Number of times students access the module pages
Number of courses	Number of times students access the courses
Number of discussion	Number of times students access the discussion forum
Number of files	Number of times students access the files
Number of gradebooks	Number of times students check the gradebooks
Number quizzes	Number of times students access the quizzes
Number of submissions	Number of times students submit assignments
Number of Wikis	Number of times students access the wikis pages
Number of actives days	Number of days student interacts with the course
Dropout week	The week when the student last visits the course. This is for algorithms to predict.

The features they tracked in the study.

VOCAB: (w/definition)

MOOC – massive online open course – free online courses that anyone can enroll into
Attrition – gradual dropping out
K-nearest neighbors (KNN) – an algorithm that classifies data into categories based on a given number of dimensions
Decision tree – consists of a root node (features), branches (rules for classification), and leaf nodes (classification)
Support vector machine (SVM) - creates a plane in feature space to separate

Cited references to follow up on

Follow up Questions

Can they evaluate the effectiveness of personalized intervention?
Are dropout rates affected by the content of the MOOC?
How well would educators be able to understand and use these?

Article #10 Notes:

Source Title	Enhancing Math Learning with AI: ChatGPT's Impact on Number Base Conversion Comprehension
Source citation (APA Format)	Gadapa, S. P., Daud, S. B. M., Hui, B. T. C., & Raju, M. R. T. (2024). Enhancing Math Learning with AI: ChatGPT's Impact on Number Base Conversion Comprehension. *International Journal of Academic Research in Progressive Education and Development*, *13*(3), 992–1008. http://dx.doi.org/10.6007/IJARPED/v13-i3/21642
Original URL	http://dx.doi.org/10.6007/IJARPED/v13-i3/21642
Source type	Journal article
Keywords	ChatGPT, Student Performance, Wilcoxon-singed Rank Test, Man-Whitney U Test, Number Base Conversions
#Tags	
Summary of key points + notes (include methodology)	First, they generated various levels of questions using ChatGPT. The experimental group of students had a ChatGPT assessment in between the pre- and post-assessments while the control group did not. Each group had 170 randomly selected students. They collected data on the students' answers and demographics. For statistical testing, they ran many tests (to determine which test should be used) on the data but finally a Wilcoxon signed-rank test. It determined that there was a significant difference for the experimental group (but not for the control group which was to be expected). They also found no significant difference between male and female students.
Research Question/Problem/ Need	How well can questions generated by ChatGPT impact students' skill in number base conversions?
Important Figures	
VOCAB: (w/definition)	Shapiro-Wilk test – a test that determines if data follows a normal distribution or not
Cited references to follow up on	Luan, H., Geczy, P., Lai, H., Gobert, J., Yang, S. J. H., Ogata, H., Baltes, J., Guerra, R., Li, P., & Tsai, C.-C. (2020). Challenges and Future Directions of Big Data and Artificial Intelligence in Education. Frontiers in Psychology, 11. https://doi.org/10.3389/fpsyg.2020.580820
Follow up Questions	Are ChatGPT questions any better or worse than regular ones made by teachers?

	If not, how could these questions be improved? Would ChatGPT's effectiveness be maintained for more complex questions?

Patent #1 Notes:

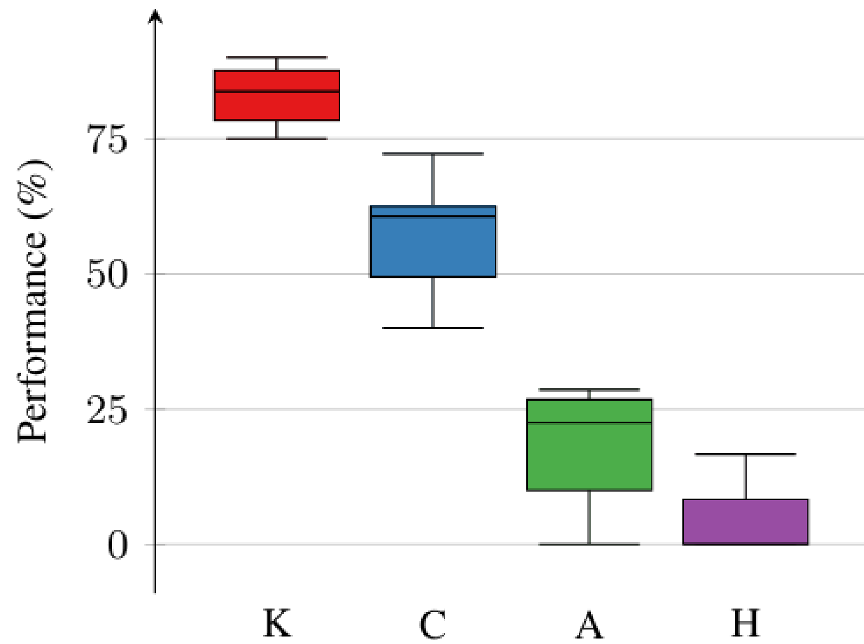
Source Title	Methods and systems for secure data analysis and machine learning
Source citation (APA Format)	Carley, D. N., (2022). *Methods and systems for secure data analysis and machine learning* (U.S. Patent No. 20220067181A1). U.S. Patent and Trademark Office. https://patents.google.com/patent/US20220067181A1/en
Original URL	https://patents.google.com/patent/US20220067181A1/en
Source type	Patent
Keywords	
#Tags	
Summary of key points + notes (include methodology)	The goal was to use systems that would allow for security in machine learning development. To allow this, the system included keeping labeled data confidential so that a user or device would only have access to subsets of the data at any given time. Also, the parameters would be kept completely confidential from the users as well. Most importantly, a model would be encrypted after training so that it can be safely released for public use.
Research Question/Problem/ Need	How can machine learning models be made more secure?
Important Figures	
VOCAB: (w/definition)	Artifacts – outputs of a model during various stages
Cited references to follow up on	
Follow up Questions	Is it possible to encrypt the data during the initial phases? Why is it important to encrypt the model? Are these protections foolproof and if not, how can they be improved?

Patent #2 Notes:

Source Title	Distributed labeling for supervised learning
Source citation (APA Format)	Bhowmick, A., Rogers, R. M., Vaishampayan, U. S., Vyrros, A. H., (2020). *Distributed labeling for supervised learning* (U.S. Patent No. 20200104705A1). U.S. Patent and Trademark Office. https://patents.google.com/patent/US20200104705A1/en
Original URL	https://patents.google.com/patent/US20200104705A1/en
Source type	Patent
Keywords	
#Tags	
Summary of key points + notes (include methodology)	They developed a technique to crowdsource labeling of data to be used for machine learning models while maintaining the privacy of the data. It starts with sending out unlabeled data to people's mobile devices that would give them back labels. Then, they would receive these labels encoded and determine the most frequent ones. They would add each element in the original data set with its most commonly proposed label to a training data set. Finally, it would train a model on this new data set.
Research Question/Problem/ Need	How can labels be created for unlabeled data, when necessary, without sacrificing privacy and still being feasible on a large scale?
Important Figures	
VOCAB: (w/definition)	Recurrent neural network – a neural network that feeds the output from the previous step as the input into the next one
Cited references to follow up on	
Follow up Questions	How will they know that these are reliable sources of labels? What is the necessity of privatizing data? What are the limitations of crowdsourced labels?

Article #11 Notes:

Source Title	Investigating the Effectiveness of ChatGPT in Mathematical Reasoning and Problem Solving: Evidence from the Vietnamese National High School Graduation Examination
Source citation (APA Format)	Dao, X.-Q., & Le, N.-B. (2023, October 10). *Investigating the Effectiveness of ChatGPT in Mathematical Reasoning and Problem Solving: Evidence from the Vietnamese National High School Graduation Examination*. ArXiv. https://arxiv.org/pdf/2306.06331
Original URL	https://doi.org/10.48550/arXiv.2306.06331
Source type	Arxiv article
Keywords	ChatGPT · large language model · natural language processing · Vietnamese high school graduation examination
#Tags	
Summary of key points + notes (include methodology)	Their goal was to determine ChatGPT's effectiveness at high school level math skills. Specifically, they evaluated its performance on the VNHSGE (VietNamese High School Graduation Examination) dataset. The dataset was also separated by which year the question came from. VNHSGE consists of 250 multiple choice questions covering high school math topics like algebra, geometry, and calculus. The researchers divided the questions into four levels of difficulty (knowledge, comprehension, application, and high application). They gave ChatGPT each question as well as instructions on how to format the answer. As expected, they found that it performed worse on the harder questions. The exact percentages are shown in the figure.
Research Question/Problem/ Need	How effective is ChatGPT at various math tests?

<table>
<tr><td>Important Figures</td><td>

ChatGPT's performance on each difficulty of question.</td></tr>
<tr><td>VOCAB: (w/definition)</td><td></td></tr>
<tr><td>Cited references to follow up on</td><td></td></tr>
<tr><td>Follow up Questions</td><td>How can it improve at the more difficult problems?
How do these scores compare to a teacher, or someone expected to be knowledgeable at these skills?
How did they separate the question difficulties?</td></tr>
</table>

Article #12 Notes

Source Title	Learning Relation-Enhanced Hierarchical Solver for Math Word Problems
Source citation (APA Format)	Lin, X., Huang, Z., Zhao, H., Chen, E., Liu, Q., Lian, D., Li, X., & Wang, H. (2023). Learning Relation-Enhanced Hierarchical Solver for Math Word Problems. *IEEE Transactions on Neural Networks and Learning Systems*, *35*(10), 13830-13844. http://dx.doi.org/10.1109/TNNLS.2023.3272114
Original URL	http://dx.doi.org/10.1109/TNNLS.2023.3272114
Source type	Journal Article
Keywords	
#Tags	
Summary of key points + notes (include methodology)	The main idea was that because humans solve math word problems much more efficiently than language models, they should take a more human approach to solving them. This was characterized by a few main factors. For example, humans gather meaning from a problem phrase by phrase, while machine learning models may go word by word by default. Also, humans can mentally group together similar problems, making them easier to solve, while AI models usually do not. They developed a hierarchical math solver (HMS) that derives semantics from each clause of a problem as they relate to the total problem. Then, they make a relation enhanced math solver (RHMS) that determines the similarity between math problems based on the structure. Both the HMS and RHMS proved to be effective when tested on large math datasets.
Research Question/Problem/ Need	How can math word problems be solved more efficiently by artificial intelligence?
Important Figures	
VOCAB: (w/definition)	GAT – graph attention network – a neural network that works with data structured as graphs
Cited references to follow up on	
Follow up Questions	How much more resource intensive is the RHMS than the HMS? Was any testing done to show that there was a statistically significant difference between RHMS and HMS performance?

	How were they able to make the models perform well on many different data sets?

Article #13 Notes

Source Title	Learning Fine-Grained Expressions to Solve Math Word Problems
Source citation (APA Format)	Huang, D., Shi, S., Lin, C., & Yin, J. (2017). Learning Fine-Grained Expressions to Solve Math Word Problems. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 805-814. https://doi.org/10.18653/v1/D17-1084
Original URL	Learning Fine-Grained Expressions to Solve Math Word Problems - ACL Anthology
Source type	Journal Article
Keywords	
#Tags	
Summary of key points + notes (include methodology)	The main challenge the researchers tackled was deriving math concepts from natural language. Problems may use different words and contexts but still require the same math concept to be applied. First, they used their training data to create a few templates that the model could then use to solve any problem. That way, the model would be able to map the problem to a given template and then place the numbers in and solve. A template can just be thought of as a system of equations. When tested on a public dataset Dolphin18K, they got an accuracy of 28%. This may seem low, but at the time it was quite competitive as one state-of-the-art system at the time only reached an accuracy of 18%, for example.
Research Question/Problem/ Need	Systems that automatically solve math word problems have very low accuracy.
Important Figures	
VOCAB: (w/definition)	Template – in the context of this article, a template was a system of equations with coefficients as variables that could be substituted by the numbers in the problem

Cited references to follow up on	
Follow up Questions	Would this concept still apply to problems with many steps? How complex were the problems in the dataset? Are these methods outdated compared to new methods and technology?

Article #14 Notes

Source Title	The Ultimate Guide to Fine-Tuning LLMs from Basics to Breakthroughs: An Exhaustive Review of Technologies, Research, Best Practices, Applied Research Challenges and Opportunities Venkatesh Balavadhani Parthasarathy, Ahtsham Zafar, Aafaq Khan, and Arsalan Shahid
Source citation (APA Format)	Parthasarathy, V. B., Zafar, A., Khan, A., & Shahid, A. (2024, October 30). *The Ultimate Guide to Fine-Tuning LLMs from Basics to Breakthroughs: An Exhaustive Review of Technologies, Research, Best Practices, Applied Research Challenges and Opportunities*. ArXiv. https://arxiv.org/pdf/2408.13296
Original URL	2408.13296
Source type	Arxiv article
Keywords	
#Tags	
Summary of key points + notes (include methodology)	There are 3 types of fine-tuning: supervised, unsupervised, and instructional. Supervised uses labeled data and is better when a specific task is in mind, while unsupervised is when unlabeled data is used to improve language capabilities in a domain. Instructional uses prompt engineering. Retrieval augmented generation (RAG) – incorporation of one's own data into prompts for LLMs. It is much more cost effective because it doesn't require all the hassles of fine-tuning and is good for question-and-answer use cases. RAG requires much less data. Parameter-efficient fine-tuning (PEFT) - less intensive than total fine-tuning because it involves adding adaptive layers to the neural network rather than editing every single one. Low rank adaptation (LoRA) - transforming the model into one with lower number of parameters. Allows for less resource-intensive tuning.
Research Question/Problem/ Need	The goal was to determine the best types of parameter-efficient fine-tuning for large language models, as well as creating a guide on which ones should be used.
Important Figures	

VOCAB: (w/definition)	Parameter-efficient fine-tuning – only tweaking a smaller number of parameters in a language model during fine-tuning rather than the entire model
Cited references to follow up on	
Follow up Questions	Are different methods of fine-tuning better suited to different tasks? What kind of parameter-efficient fine-tuning is best for abstract question answering? Which methods of fine-tuning are the most resource efficient?

Article #15 Notes

Source Title	Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks
Source citation (APA Format)	Lewis, P., Perez, E., Piktus, A., Petroni, P., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschef, T., Riedel, S., & Kiela, D. (2021, April 12). *The Ultimate Guide to Fine-Tuning LLMs from Basics to Breakthroughs: An Exhaustive Review of Technologies, Research, Best Practices, Applied Research Challenges and Opportunities*. ArXiv. https://arxiv.org/pdf/2005.11401
Original URL	[2005.11401] Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks
Source type	Arxiv article
Keywords	
#Tags	
Summary of key points + notes (include methodology)	Retrieval augmented generation (RAG) uses input to get some form of stored information (usually text documents) and then uses that to get an output. Used as fine tuning when you have your own data. They experimented with RAG on open-domain question answering, abstractive question answering, jeopardy question generation, and fact answering. RAG outperformed other models in all tasks tested.
Research Question/Problem/ Need	The problem was that LLMs have lots of knowledge due to the amount of data they are trained on but usually cannot use the knowledge in meaningful ways besides regurgitating it.
Important Figures	
VOCAB: (w/definition)	Retrieval augmented generation – uses input to get information from a given source, usually a text document, then uses that source to form an output
Cited references to follow up on	
Follow up Questions	How did they score performance, especially in more complicated tasks like jeopardy question generation? Would retrieval augmented generation be suited towards non-NLP tasks?

	Is RAG-token model which can pull from different documents for each token better than RAG-sequence model which only uses one document?

Article #16 Notes

Source Title	Sequence to Sequence Learning with Neural Networks
Source citation (APA Format)	Sutskever, I., Vinyals, O., & Le, Q. V. (2014, December 14). *Sequence to Sequence Learning with Neural Networks*. ArXiv. https://arxiv.org/pdf/1409.3215
Original URL	arXiv:1409.3215v3 [cs.CL] 14 Dec 2014
Source type	Arxiv article
Keywords	
#Tags	
Summary of key points + notes (include methodology)	Deep Neural Networks are very powerful but can't handle tasks that are sequential problems as their inputs and outputs are of unspecified dimensionality. Their plan was to test a model on English to French translation. They started with a Recurrent Neural Network which can map input sequences to output sequences provided that they are the same length. They used an LSTM to map an entire input sequence to a vector which would then be mapped to an output using another LSTM. This allowed recurrent neural networks to be used because the size of the vectors would be known. Their model received a BLEU score of 34.81 on the task of translation from English to French on the dataset used. BLEU is a common metric used to score machine learning models. For context, a different model that they were comparing to only had a score of 33.
Research Question/Problem/ Need	Deep Neural Networks are powerful but limited to labeled data; they can't be used for sequential tasks.
Important Figures	
VOCAB: (w/definition)	Long Short-Term Memory – LSTM – a type of recurrent neural network that can hold information for a longer period of time
Cited references to follow up on	

Follow up Questions	How was the LSTM able to translate long sentences despite memory constraints? How can the semantics of a sentence be captured in just a single vector? Would it be easy to use the model to translate French to English?

Article #17 Notes

Source Title	AI Chatbots as Math Algorithm Problem Solvers: A Critical Evaluation of Its Capabilities and Limitations
Source citation (APA Format)	Dahal, N., Luitel, B., C., Lamichhane, B., R., & Pant, B., P. (2023). AI Chatbots as Math Algorithm Problem Solvers: *Proceedings of the 28th Asian Technology Conference in Mathematics*, 429–438. https://www.researchgate.net/publication/375522509_AI_Chatbots_as_Math_Algorithm_Problem_Solvers_A_Critical_Evaluation_of_Its_Capabilities_and_Limitations
Original URL	(PDF) AI Chatbots as Math Algorithm Problem Solvers: A Critical Evaluation of Its Capabilities and Limitations
Source type	Conference paper
Keywords	
#Tags	
Summary of key points + notes (include methodology)	Advanced Chatbot language models like ChatGPT and Bard can solve and explain basic math problems. They can also generate problems for educational purposes. They often use embed code as a response which can then be used to create an answer. However, a limitation of many of these models is that their solutions often come with drawbacks and are poorly explained. For example, even the tool WolframAlpha might not show all the steps required to solve a problem. They treat WolframAlpha like a chatbot in this article even though it differs heavily from models such as ChatGPT and Bard which are much more focused on natural language. Another drawback was that models tend to perform better on problems that are more theoretical examples and struggle with application of math concepts in the real world.
Research Question/Problem/Need	How well is AI currently equipped to handle skills necessary for aiding in math education?

Important Figures	
VOCAB: (w/definition)	Multimodal – a model that can work with various types of data (text, image, audio, etc.)
Cited references to follow up on	
Follow up Questions	Why did they lump in WolframAlpha with chatbots? Why do natural language models struggle with geometry specifically? Why would they struggle with things that require real world knowledge if they are mostly trained on real world data rather than pure math data?

Article #18 Notes

Source Title	ChatGLM-Math: Improving Math Problem-Solving in Large Language Models with a Self-Critique Pipeline
Source citation (APA Format)	Xu, Y., Liu, X., Liu, X., Hou, Z., Li, Y., Zhang, X., Wang, Z., Zeng, A., Du, Z., Zhao, W., Tang, J., & Dong, Y. (2024, April 3). *ChatGLM-Math: Improving Math Problem-Solving in Large Language Models with a Self-Critique Pipeline*. ArXiv. https://arxiv.org/pdf/2404.02893

Original URL

[2404.02893] ChatGLM-Math: Improving Math Problem-Solving in Large Language Models with a Self-Critique Pipeline

Source type

ArXiv article

Keywords

#Tags

Summary of key points + notes (include methodology)

Machine learning models have already been used to generate feedback. This paper aimed to create a math-critique model to generate feedback on a large language model and thus improve its performance.
One of the main problems with training LLMs to solve math problems is that the standard method of supervised fine-tuning may increase its math domain ability, but this would come at the cost of general language abilities.
They used rejective fine tuning and direct performance optimization. Rejective fine tuning in this case was allowing the model to create responses, then scoring those responses with the math-critique model, then getting rid of low-scoring responses and fine-tuning with high-scoring ones. For direct performance optimization, they compared pairs of correct and incorrect answers to further tune the model. This was done after direct performance optimization.
They tested these methods with many different models in both English and Chinese, including GPT-3.5-Turbo, Claude-2, and ChatGLM-3

Research Question/Problem/ Need

Large language models struggle with math problem solving because it differs from standard language usage.

Important Figures

Models	#params	Chinese					English			General	
		MathUserEval			Ape210k	Cmath	GSM8k	MATH	Hungarian	AlignBench	MT-Bench
		Overall	Elementary	Advanced						Language	
GPT-4-1106-Preview [34]	N/A	**5.73**	**5.07**	**6.81**	84.2	**89.3**	**93.6**	**53.6**	**92**	8.29	**9.32**
GPT-4-0613 [34]	N/A	4.14	3.34	5.33	83.6	86.5	91.4	45.8	68	7.59	9.18
GPT-3.5-Turbo-0613 [34]	N/A	3.42	3.04	4.07	70.4	76.8	78.2	28.0	41	6.82	8.36
Claude-2 [1]	N/A	3.29	2.63	4.35	72.8	80.5	88.0	-	55	6.78	8.06
GLM-4	N/A	5.11	4.86	5.43	**93.5**	89.0	91.8	49.0	75	**8.38**	8.62
Skywork-13B-Math [54]	13B	2.66	2.75	2.54	74.4	77.3	72.3	17.0	39	5.58	4.12
InternLM2-Chat [43]	20B	3.25	3.00	3.68	72.0	80.7	79.6	34.8	48	7.68	8.21
Math-InternLM2 [43]	20B	3.17	3.08	3.37	75.2	78.5	82.6	37.7	66	6.53	6.09
Yi-Chat [56]	34B	2.64	2.49	2.87	65.1	77.7	76.0	15.9	39	6.18	6.54
DeepSeek-Chat [12]	67B	3.24	2.76	3.84	76.7	80.3	**84.1**	32.6	58	7.11	**8.35**
MetaMath (EN) [57]	70B	-	-	-	-	-	82.3	26.0	35	-	4.28
Qwen-Chat [3]	72B	3.87	3.99	3.67	77.1	**88.1**	76.4	31.8	52	7.29	6.43
ChatGLM3-32B-SFT-2312*	32B	3.25	3.03	3.60	78.0	79.8	75.8	29.0	39	7.37	8.05
+ RFT	32B	4.01	3.86	4.26	87.0	85.3	82.4	39.5	58	7.42	8.03
+ RFT, DPO	32B	**4.23**	**4.01**	**4.59**	**89.4**	85.6	82.6	**40.6**	**73**	**7.80**	8.08

* ChatGLM3-32B-SFT-2312 is a newer version of the ChatGLM series and not identical to the model discussed in [19], despite sharing the same model size.

Scoring of inference of each model after training with the described methods. Each dataset has a different manner of scoring, so they are best compared within a column.

VOCAB: (w/definition)

RLHF – reinforcement learning from human feedback – using humans to score responses that are then used to further tune the model

Cited references to follow up on	
Follow up Questions	Do these results show that it is worth using the self-critique pipeline? Why weren't the numbers of parameters of some of the models known? Why did ChatGLM score the highest on most of the datasets?

Article #19 Notes

Source Title	SciInstruct: a Self-Reflective Instruction Annotated Dataset for Training Scientific Language Models
Source citation (APA Format)	Zhang, D., Hu, Z., Zhoubian, Z., Du, Z., Yang, K., Wang, Z., Yue, Y., Dong, Y., & Tang, J. (2024, November 18). *SciInstruct: a Self-Reflective Instruction Annotated Dataset for Training Scientific Language Models.* ArXiv. https://arxiv.org/pdf/2401.07950
Original URL	[2401.07950] SciInstruct: a Self-Reflective Instruction Annotated Dataset for Training Scientific Language Models
Source type	ArXiv article
Keywords	

#Tags

Summary of key points + notes (include methodology)

They needed a large dataset full of varying scientific questions to train large language models. Chain of thought reasoning has been used to improve LLM performance on reasoning tasks, but for scientific data, chain of thought examples are not abundant.
They started with 257,143 data points which were question-answer pairs. They used GPT-4 to generate intermediate steps by prompting it to give steps that would get to the answer. Then, they had other models label the accuracy to filter out inaccurate steps.
To test Sci-Instruct, they chose ChatGLM3, Llama3-8B-Instruct, and Mistral-7B. They used the Sci-Instruct dataset to fine-tune each of these models so that they could test its accuracy. Then, they used various evaluation datasets to test their abilities. These models were able to outperform others, even if they had more parameters.

Research Question/Problem/ Need

LLMs are useful in science domains but are limited by a lack of scientific reasoning.

Important Figures

Model	CEval-Hard	CEval-Sci	MMLU-Sci	SciEval	SciBench	GPQA_Diamond	Avg. Sci	Avg. {Sci+Math}
(API, parameter details unknown)								
GPT-4	54.96	60.55	-	**73.93**	**28.52**	39.70	-	-
GPT-3.5-turbo	41.37	46.83	-	66.97	12.17	-	-	-
Claude-v1.3	39.14	44.64	-	63.45	-	-	-	-
(# parameter = 6B$\sim$7B)								
LLaMA-2-7B	28.29†	30.00†	30.41	28.37	0.40	-	-	-
Galactica-6.7B	11.84†	11.44†	30.68	50.87	-	-	-	-
ChatGLM2-6B	29.61†	45.71†	37.09†	53.02†	1.54†	-	-	-
ChatGLM2-6B-Base	32.90†	40.95†	38.06†	50.38†	1.20†	-	-	-
ChatGLM3-6B	36.84†	38.57†	41.78†	56.56†	2.40†	28.70	34.14	29.73
ChatGLM3-6B-Base	45.40†	54.29†	40.16†	61.69†	2.40†	24.75	38.12	40.34
SciGLM (ChatGLM3-6B-Base)	**51.97**	**60.00**	45.34	62.09	**3.77**	25.25	**41.40**	**45.32**
Llama3-8B-Instruct (zero-shot)	26.32†	27.62†	26.90†	**71.38**†	1.03†	27.27†	30.09	28.58
Llama3-8B-Instruct (few-shot)	25.66†	23.33†	**52.67**†	71.38†	3.60†	31.31†	34.66	37.92
+ SciInstruct	32.24	34.76	40.86	66.47	3.60	29.29	34.54	36.04
Mistral-7B: MetaMATH (zero-shot)	9.87†	8.57†	28.25†	63.61†	4.63†	27.78†	23.79	25.57
Mistral-7B: MetaMATH (few-shot)	9.21†	19.52†	44.74†	63.61†	6.17†	29.29†	28.76	33.92
+ SciInstruct	30.92	38.10	42.16	64.16	6.23	27.27	34.81	37.91
(# parameter = 12B$\sim$13B)								
LLaMA-2-13B	19.74†	19.05†	35.85	36.96	1.37	26.20	22.59	22.13
Vicuna-13B	-	-	32.13	53.93	-	-	-	-
(# parameter = 30B$\sim$32B)								
Galactica-30B	-	-	35.53	54.96	-	-	-	-
ChatGLM3-32B-Base	53.95†	64.29†	**50.30**†	67.38†	4.29†	22.22	43.74	48.62
SciGLM (ChatGLM3-32B-Base)	**56.58**	**66.19**	49.38	**69.84**	**5.15**	**25.76**	**45.48**	**51.47**

Bolded numbers within a column means that model performed best compared to others with similar numbers of parameters

VOCAB: (w/definition)

Lean – a popular syntax that is used to write math proofs and theorems with formality and logic

Cited references to follow up on

Follow up Questions

How did they ensure accuracy of labeling of intermediate steps?
What's the difference between ChatGLM3 and SciGLM(ChatGLM3) if they're both tuned off SciInstruct?
How is small size considered a downside if it's just experimental?

Article #20 Notes

Source Title	Dual Instruction Tuning with Large Language Models for Mathematical Reasoning
Source citation (APA Format)	Zhao, T., & Zhou, Y. (2024, March 27). *Dual Instruction Tuning with Large Language Models for Mathematical Reasoning.* ArXiv. https://doi.org/10.48550/arXiv.2403.18295
Original URL	[2403.18295] Dual Instruction Tuning with Large Language Models for Mathematical Reasoning
Source type	ArXiv article
Keywords	
#Tags	
Summary of key points + notes (include methodology)	They used dual instruction tuning, meaning they would tune the model's generations in both directions of the sequence. They used the existing dataset MathInstruct, and applied what they called intermediate reasoning state prediction. This would involve masking certain parts of the data and then having the model fill in the gaps with its own generations. These generations would then be added to the dataset. This required the models to

	use context from previous steps to get closer to an answer. They also applied instruction reconstruction, which involved doing the same thing as intermediate reasoning state prediction, but from backwards reasoning. They trained various models on this chain of thought data and found that it mainly resulted in improvements on more challenging datasets. Additionally, they calculated loss, or error in expected output on average.
Research Question/Problem/ Need	Although Chain-of-thought is a powerful method to improve LLM reasoning skills, it still has limitations with the steps sometimes being missing, inaccurate, or unnecessary.
Important Figures	
VOCAB: (w/definition)	Ablation – removing components from a model one at a time to see what causes changes
Cited references to follow up on	
Follow up Questions	Wouldn't this result in inaccurate steps if the models made bad generations? Why was it not as effective at improving performance on simple datasets? Why did they test with a task that was not math related?