Using Generative AI Tools

Imagination is All you Need

Michael A. Heroux

Generative AI Predictions in Quotes

*Like with all technological revolutions, I expect there to be significant impact on jobs, but exactly what that impact looks like is very difficult to predict…**I believe that there will be far greater jobs on the other side of this, and that the jobs of today will get better…***

Sam Altman

*Generative AI is the key to solving some of the world's biggest problems, such as climate change, poverty, and disease. **It has the potential to make the world a better place for everyone.***

Mark Zuckerberg

*Generative AI is the most powerful tool for creativity that has ever been created. **It has the potential to unleash a new era of human innovation.***

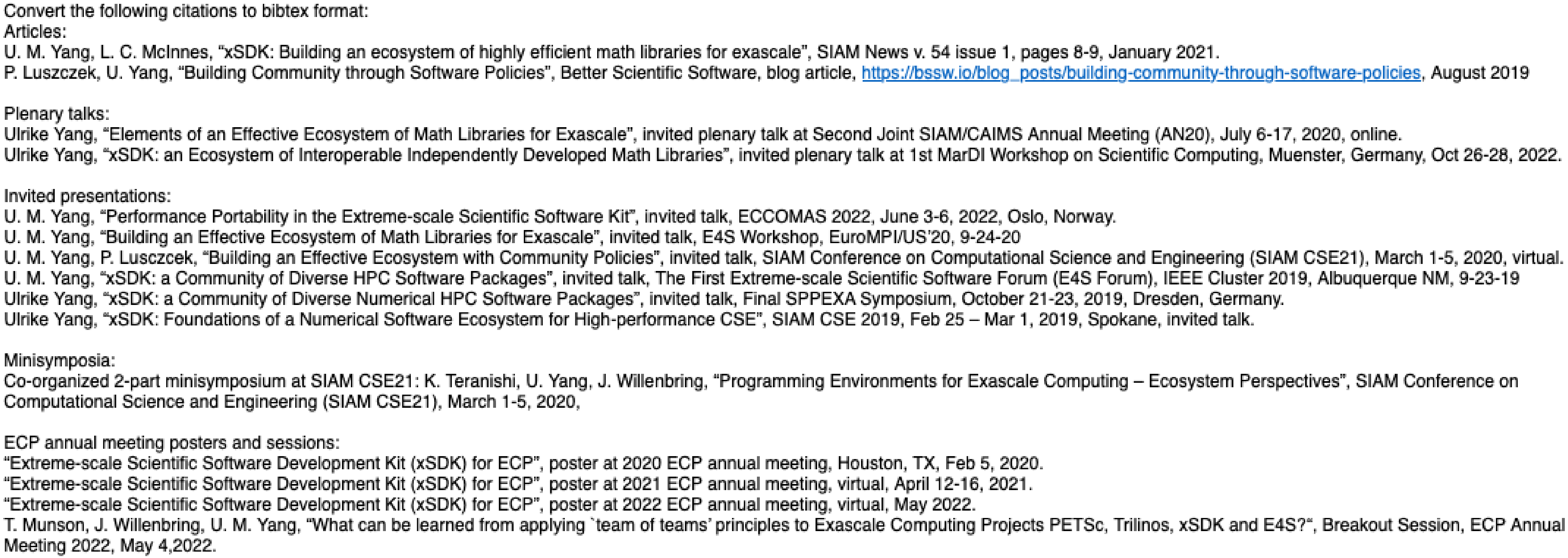
Elon Musk

What if the A.I. Boosters Are Wrong?

NY Times article on Daren Acemuglo predictions for modest growth

Theme 1: Helping You with Tedious Writing Tasks

OpenAI: BibTeX citations - Convert from MLA-style format



Convert the following citations to bibtex format:
Articles:
U. M. Yang, L. C. McInnes, “xSDK: Building an ecosystem of highly efficient math libraries for exascale”, SIAM News v. 54 issue 1, pages 8-9, January 2021.
P. Luszczek, U. Yang, “Building Community through Software Policies”, Better Scientific Software, blog article, https://bssw.io/blog_posts/building-community-through-software-policies, August 2019

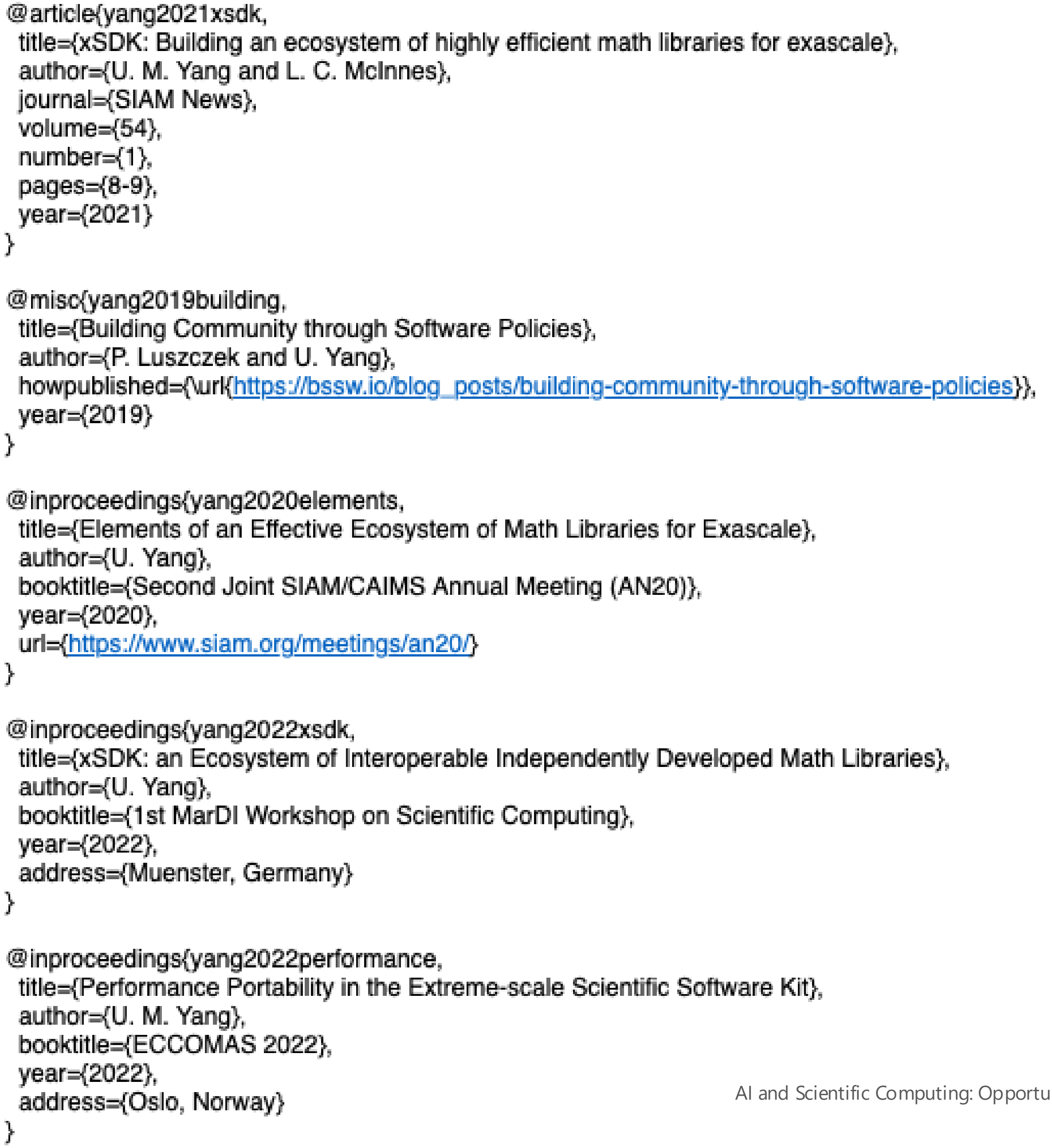
Plenary talks:
Ulrike Yang, “Elements of an Effective Ecosystem of Math Libraries for Exascale”, invited plenary talk at Second Joint SIAM/CAIMS Annual Meeting (AN20), July 6-17, 2020, online.
Ulrike Yang, “xSDK: an Ecosystem of Interoperable Independently Developed Math Libraries”, invited plenary talk at 1st MarDI Workshop on Scientific Computing, Muenster, Germany, Oct 26-28, 2022.

Invited presentations:
U. M. Yang, “Performance Portability in the Extreme-scale Scientific Software Kit”, invited talk, ECCOMAS 2022, June 3-6, 2022, Oslo, Norway.
U. M. Yang, “Building an Effective Ecosystem of Math Libraries for Exascale”, invited talk, E4S Workshop, EuroMPI/US’20, 9-24-20
U. M. Yang, P. Lusczcek, “Building an Effective Ecosystem with Community Policies”, invited talk, SIAM Conference on Computational Science and Engineering (SIAM CSE21), March 1-5, 2020, virtual.
U. M. Yang, “xSDK: a Community of Diverse HPC Software Packages”, invited talk, The First Extreme-scale Scientific Software Forum (E4S Forum), IEEE Cluster 2019, Albuquerque NM, 9-23-19
Ulrike Yang, “xSDK: a Community of Diverse Numerical HPC Software Packages”, invited talk, Final SPPEXA Symposium, October 21-23, 2019, Dresden, Germany.
Ulrike Yang, “xSDK: Foundations of a Numerical Software Ecosystem for High-performance CSE”, SIAM CSE 2019, Feb 25 – Mar 1, 2019, Spokane, invited talk.

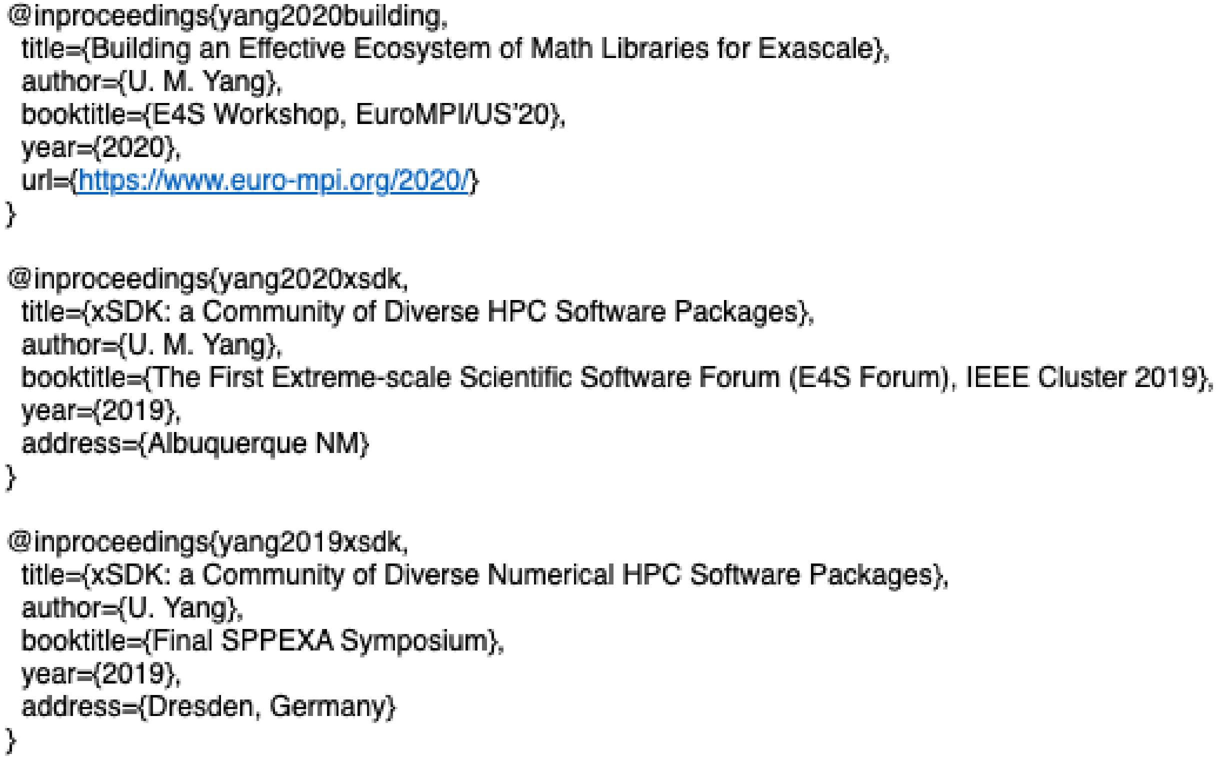
Minisymposia:
Co-organized 2-part minisymposium at SIAM CSE21: K. Teranishi, U. Yang, J. Willenbring, “Programming Environments for Exascale Computing – Ecosystem Perspectives”, SIAM Conference on Computational Science and Engineering (SIAM CSE21), March 1-5, 2020,

ECP annual meeting posters and sessions:
“Extreme-scale Scientific Software Development Kit (xSDK) for ECP”, poster at 2020 ECP annual meeting, Houston, TX, Feb 5, 2020.
“Extreme-scale Scientific Software Development Kit (xSDK) for ECP”, poster at 2021 ECP annual meeting, virtual, April 12-16, 2021.
“Extreme-scale Scientific Software Development Kit (xSDK) for ECP”, poster at 2022 ECP annual meeting, virtual, May 2022.
T. Munson, J. Willenbring, U. M. Yang, “What can be learned from applying `team of teams’ principles to Exascale Computing Projects PETSc, Trilinos, xSDK and E4S?“, Breakout Session, ECP Annual Meeting 2022, May 4,2022.

OpenAI: BibTeX citations - Convert from MLA-style format



```
@article{yang2021xsdk,
  title={xSDK: Building an ecosystem of highly efficient math libraries for exascale},
  author={U. M. Yang and L. C. McInnes},
  journal={SIAM News},
  volume={54},
  number={1},
  pages={8-9},
  year={2021}
}

@misc{yang2019building,
  title={Building Community through Software Policies},
  author={P. Luszczek and U. Yang},
  howpublished={\url{https://bssw.io/blog_posts/building-community-through-software-policies}},
  year={2019}
}

@inproceedings{yang2020elements,
  title={Elements of an Effective Ecosystem of Math Libraries for Exascale},
  author={U. Yang},
  booktitle={Second Joint SIAM/CAIMS Annual Meeting (AN20)},
  year={2020},
  url={https://www.siam.org/meetings/an20/}
}

@inproceedings{yang2022xsdk,
  title={xSDK: an Ecosystem of Interoperable Independently Developed Math Libraries},
  author={U. Yang},
  booktitle={1st MarDI Workshop on Scientific Computing},
  year={2022},
  address={Muenster, Germany}
}

@inproceedings{yang2022performance,
  title={Performance Portability in the Extreme-scale Scientific Software Kit},
  author={U. M. Yang},
  booktitle={ECCOMAS 2022},
  year={2022},
  address={Oslo, Norway}
}
```



```
@inproceedings{yang2020building,
  title={Building an Effective Ecosystem of Math Libraries for Exascale},
  author={U. M. Yang},
  booktitle={E4S Workshop, EuroMPI/US'20},
  year={2020},
  url={https://www.euro-mpi.org/2020/}
}

@inproceedings{yang2020xsdk,
  title={xSDK: a Community of Diverse HPC Software Packages},
  author={U. M. Yang},
  booktitle={The First Extreme-scale Scientific Software Forum (E4S Forum), IEEE Cluster 2019},
  year={2019},
  address={Albuquerque NM}
}

@inproceedings{yang2019xsdk,
  title={xSDK: a Community of Diverse Numerical HPC Software Packages},
  author={U. Yang},
  booktitle={Final SPPEXA Symposium},
  year={2019},
  address={Dresden, Germany}
}
```

Tl;dr. Here is a WWW’04 conference paper abstract I use in a compact writing exercise, followed by the AI-generated Tl;dr version. I chose the abstract because it presents significant opportunities for improvement.

OpenAI
Tl;dr assistant

Original text:

We seek to gain improved insight into how Web search engines should cope with the evolving Web, in an attempt to provide users with the most up-to-date results possible. For this purpose we collected weekly snapshots of some 150 Web sites over the course of one year, and measured the evolution of content and link structure. Our measurements focus on aspects of potential interest to search engine designers: the evolution of link structure over time, the rate of creation of new pages and new distinct content on the Web, and the rate of change of the content of existing pages under search-centric measures of degree of change. Our findings indicate a rapid turnover rate of Web pages, i.e., high rates of birth and death, coupled with an even higher rate of turnover in the hyperlinks that connect them. For pages that persist over time we found that, perhaps surprisingly, the degree of content shift as measured using TF.IDF cosine distance does not appear to be consistently correlated with the frequency of content updating. Despite this apparent non-correlation, the rate of content shift of a given page is likely to remain consistent over time. That is, pages that change a great deal in one week will likely change by a similarly large degree in the following week. Conversely, pages that experience little change will continue to experience little change. We conclude the paper with a discussion of the potential implications of our results for the design of effective Web search engines.

Tl;dr version

The Web is constantly evolving, with new pages being created and old ones being deleted or modified. Search engines need to be able to cope with this changing landscape in order to provide users with the most up-to-date results possible. Our study found that the rate of change of Web content is much higher than the rate of change of the links that connect it. This means that search engines need to be able to adapt quickly to changes in the content of Web pages, in order to provide the most relevant and up-to-date results.

OpenAI Playground
Second-grader

Here is an opening paragraph for an article I wrote a few years ago, followed by the second-grader translation. Very interesting that the AI inserted the definition of exascale computing.

Prompt:

> Summarize this (my paragraph) for a second-grade student:
>
> The demands of massive concurrency and increased resilience required for effective exascale computing have led to claims that our existing approach to application programming must be replaced. In this article I argue that disruption is inevitable, but from the programming perspective will not require new languages or programming environments. Instead the disruption will be in application architecture, introducing a new control layer that will provide the concurrency, adaptability and resilience features we need in order to achieve effective exascale performance levels.

OpenAI Playground Generated Second-grader text:

> The article is discussing how the way we currently write programs will have to change in order to be effective at exascale computing. Exascale computing is a type of computing where programs can use a billion billion calculations per second. This is different from our current approach, which the article claims will need to be replaced in order to be effective. The new way of writing programs will include a new control layer that provides features like concurrency, adaptability, and resilience.

Theme 1 Takeaways – Makes writing better, faster, cheaper

BibTeX – Easy tedious work

Compact writing – Summarize/Tl;dr

Translation for different audiences – 2nd grader

Companion for all writing activities

Theme 2: Upscaling Your Technical Skills

Bootstrapping a scripting tool – Step 1: Ask for advice

What is a good way to extract metadata from a GitHub repo?

The best way to extract metadata from a GitHub repo is to use the GitHub API. It provides a range of endpoints that allow you to access information about the repo, such as the number of commits, pull requests, releases, and other metadata. You can also use the GitHub GraphQL API to query more specific information.

Write a script to extract basic activity information from the Trilinos GitHub repo

```
#!/usr/bin/env python3

import requests

# Enter your GitHub personal access token here
token = '<YOUR_TOKEN_HERE>'

# The GitHub repository to query
repo = 'trilinos/Trilinos'

# Get the list of commits
url = f'https://api.github.com/repos/{repo}/commits'
headers = {'Authorization': f'token {token}'}
response = requests.get(url, headers=headers).json()

# Print the commit count
print(f'Commit count: {len(response)}')

# Get the list of pull requests
url = f'https://api.github.com/repos/{repo}/pulls'
headers = {'Authorization': f'token {token}'}
response = requests.get(url, headers=headers).json()

# Print the pull request count
print(f'Pull request count: {len(response)}')

# Get the list of releases
url = f'https://api.github.com/repos/{repo}/releases'
headers = {'Authorization': f'token {token}'}
response = requests.get(url, headers=headers).json()

# Print the release count
print(f'Release count: {len(response)}')
```

Step 2: Generate script

Step 3: Learn about token

How do I determine my github personal access token?

To generate a GitHub personal access token, you must first log in to your GitHub account. Once logged in, go to your Settings page, select “Developer settings” from the left sidebar, and then “Personal access tokens”. On the next page, click “Generate new token”. You will then be prompted to give your token a name and select the scopes for which the token will be granted access. Once you have set your token’s name and scopes, click “Generate token”. Your token will then be displayed on the next page.

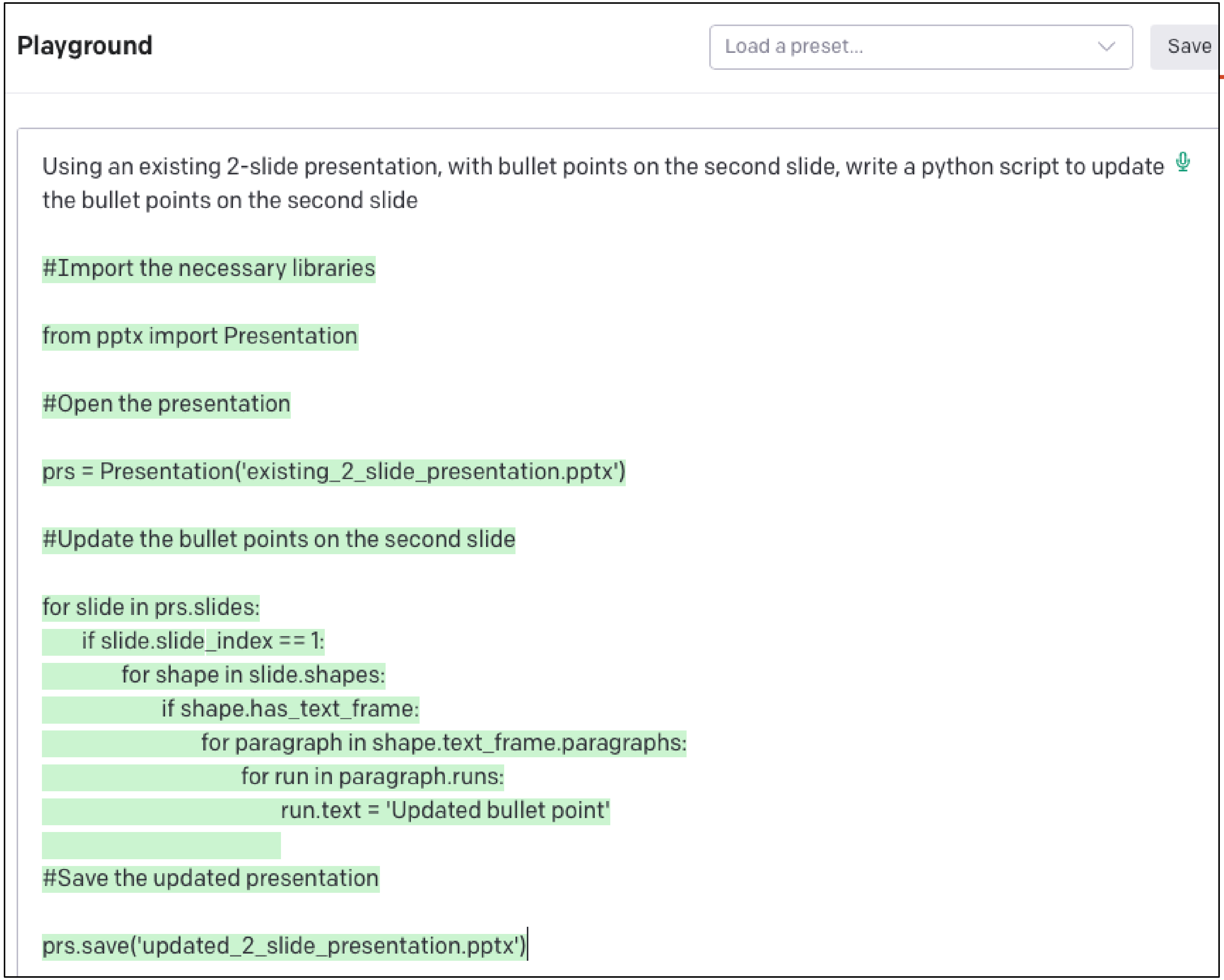
Step 4: Experiment

Write a github API python script to list the contributors and number of commits to a github repo with user provided name

```
import requests
import json


# Get the user input
user_input = input("Please enter the GitHub repo name: ")


# API call to get the list of contributors
url = "https://api.github.com/repos/{}/contributors".format(user_input)
r = requests.get(url)


# Parse the response
data = json.loads(r.text)


# Print the result
print("List of Contributors and Number of Commits to {}".format(user_input))
for contributor in data:
    print("{}: {}".format(contributor['login'], contributor['contributions']))
```

OpenAI
More specific prompts

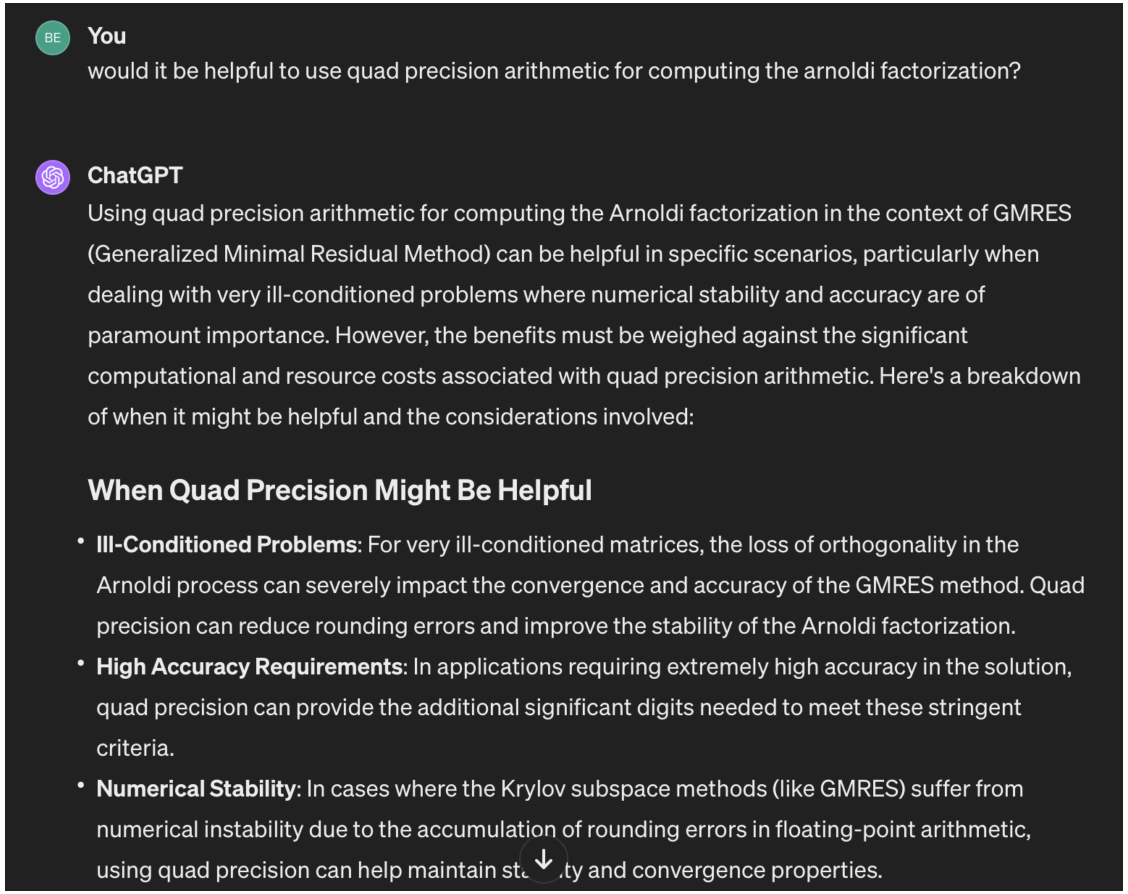
Writing detailed prompts help to produce more usable content



Using an existing 2-slide presentation, with bullet points on the second slide, write a python script to update the bullet points on the second slide

```
#Import the necessary libraries

from pptx import Presentation

#Open the presentation

prs = Presentation('existing_2_slide_presentation.pptx')

#Update the bullet points on the second slide

for slide in prs.slides:
    if slide.slide_index == 1:
        for shape in slide.shapes:
            if shape.has_text_frame:
                for paragraph in shape.text_frame.paragraphs:
                    for run in paragraph.runs:
                        run.text = 'Updated bullet point'

#Save the updated presentation

prs.save('updated_2_slide_presentation.pptx')
```

Advice on floating point arithmetic



You

would it be helpful to use quad precision arithmetic for computing the arnoldi factorization?

ChatGPT

Using quad precision arithmetic for computing the Arnoldi factorization in the context of GMRES (Generalized Minimal Residual Method) can be helpful in specific scenarios, particularly when dealing with very ill-conditioned problems where numerical stability and accuracy are of paramount importance. However, the benefits must be weighed against the significant computational and resource costs associated with quad precision arithmetic. Here's a breakdown of when it might be helpful and the considerations involved:

When Quad Precision Might Be Helpful

- **Ill-Conditioned Problems**: For very ill-conditioned matrices, the loss of orthogonality in the Arnoldi process can severely impact the convergence and accuracy of the GMRES method. Quad precision can reduce rounding errors and improve the stability of the Arnoldi factorization.
- **High Accuracy Requirements**: In applications requiring extremely high accuracy in the solution, quad precision can provide the additional significant digits needed to meet these stringent criteria.
- **Numerical Stability**: In cases where the Krylov subspace methods (like GMRES) suffer from numerical instability due to the accumulation of rounding errors in floating-point arithmetic, using quad precision can help maintain sta ty and convergence properties.

Understood the topic

Comparison with double-double

How to perform double-double with FMA

Arithmetic complexity

Memory complexity

Not to count in-register read/write

Explained results step by step

Theme 2 Takeaways: Upscaling your tech skills
AI Tools Transform our ability to work with algorithms and software

Script APIs: Makes possible the use of GitHub API example:

- Go from no knowledge to a working meta-data access script in 15 minutes (with help from StackOverflow)
- Use advanced Python functionality (scripted changes to Powerpoint file)

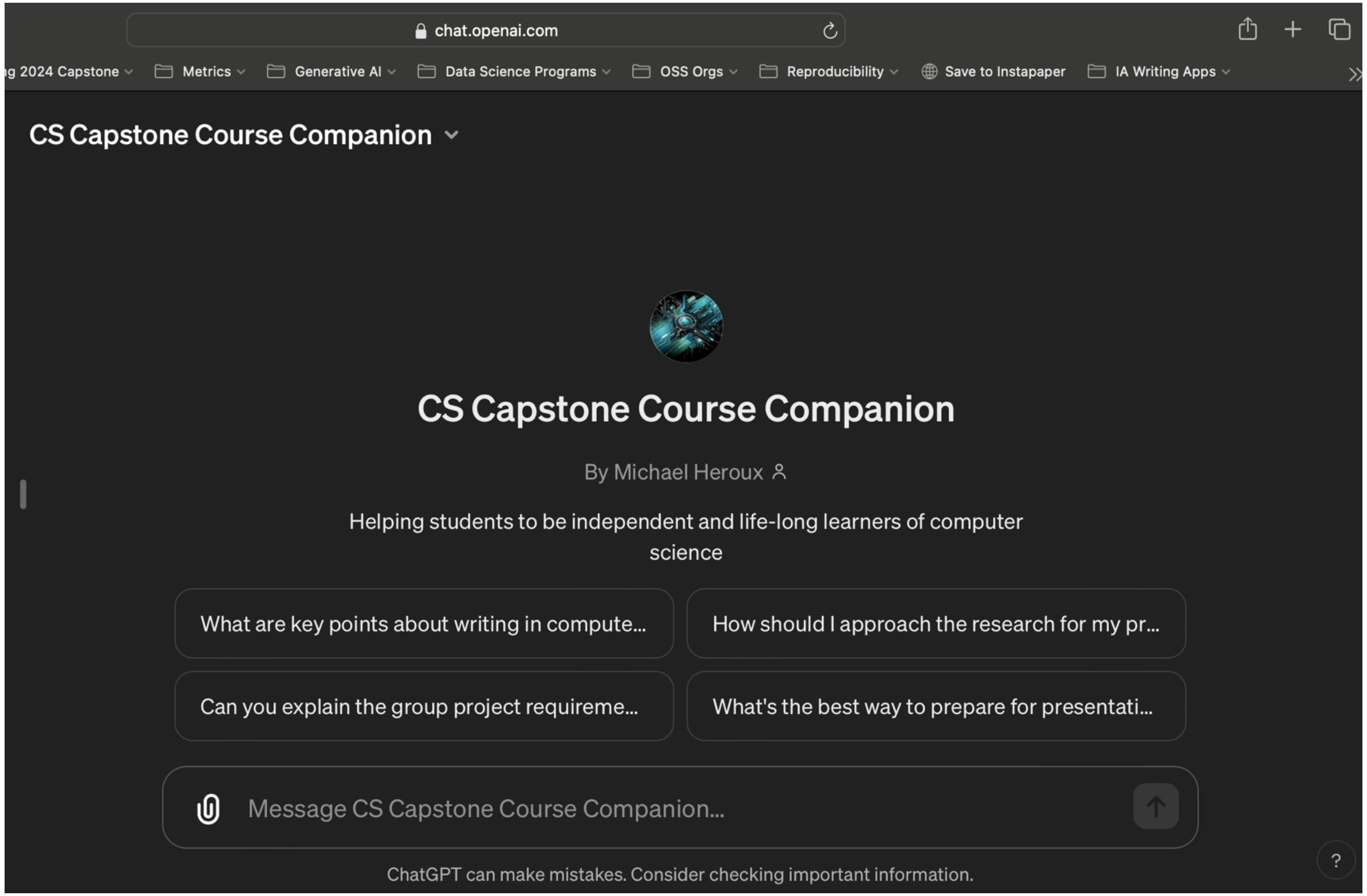
Algorithm exploration: Reason about algorithmic choices

- Explain algorithmic trade-offs
- Provide starter code
- Have a team of world experts on many topics

Theme 3: Changing the Research Enterprise

Lessons from Computer Science Capstone Course

Lesson1: Custom GPTs



CS Capstone Course Companion

CS Capstone Course Companion

By Michael Heroux

Helping students to be independent and life-long learners of computer science

What are key points about writing in compute...

How should I approach the research for my pr...

Can you explain the group project requireme...

What's the best way to prepare for presentati...

Message CS Capstone Course Companion...

ChatGPT can make mistakes. Consider checking important information.

Capstone Course Companion GPT

ChatGPT + Course Handbook
Becomes “virtual me”
Generates “new ideas” (lit search vs survey)
Form of Retrieval-augment generation (RAG)
Used by my students for many interactions
Concept useful for “exporting/importing” expertise: Virtual tutors, flipped classrooms, and more

Scenarios: Using Personas with Contexts

- Complete a task
- For a persona with given attributes
- Addressing the context specification

Sample Scenarios-Personas-Contexts

Scenario: Select a capstone project research topic. At the beginning of the semester, each student must determine a topic to serve as the foundation for their semester-long research project.

- **Student persona:** Each student will create a persona that describes their background, experience, and interests.
- **Context:** The research topic must satisfy course requirements

Scenario: Identify weak arguments in my paper.

- **Topic expert persona:** Preconditioned with reference content that conveys the domain of expertise. The student can use this persona to converse about topic details
- **Context:** Draft student paper, related peer-reviewed content

Scenario: Improve text clarity through efficient word use. One theme of the capstone course is efficient use of text. Word count reduction has strong correlation with improved technical prose clarity.

- **Efficient-text editor persona:** Singular purpose persona that focuses on reducing word count while retaining the same meaning
- **Context:** Set of strategies for reducing word count (could be generated by ChatGPT too)

Scenario: Suggest capstone course research topic

---- Copy and paste into our Capstone Course Companion GPT ----

Prompt

Based on the following information about me, generate a list of 3 possible research topics for the state-of-the-field project that I could pursue that would help me succeed in the course

My coursework: Intro to computer science, software development, algorithms, computer organization, databases, computer ethics, agile software development, differential and integral calculus, intro psychology, intro social science

My Interests: Computer gaming, chess, baseball, new ideas in technology

Things I don't find interesting: Cooking, sophisticated movies, loud music

Give me details about why each topic would be a good choice. Give me specific details about what algorithms I would need to learn to do a good job

Companion GPT response

Based on your coursework and interests, here are three potential research topics for your state-of-the-field project in the CSCI 373 course, along with justifications and relevant algorithms you would need to learn:

Artificial Intelligence in Computer Gaming:

Relevance: Your interest in computer gaming aligns perfectly with this topic. The field of AI in gaming is rapidly evolving and offers a plethora of research opportunities.

Coursework Connection: Your background in algorithms and software development will be crucial in understanding and developing AI systems for games.

Algorithms to Learn:

Pathfinding Algorithms (e.g., A*, Dijkstra’s algorithm) for NPC movement.

Decision Trees and Machine Learning algorithms for adaptive and intelligent NPC behavior.

Reinforcement Learning for dynamic game balancing and NPC skill adjustment.

Data Analysis in Sports: A Focus on Baseball:

Relevance: Merging your interest in baseball with your computing skills could lead to fascinating insights in sports analytics, a field that's gaining immense popularity.

Coursework Connection: Your knowledge from databases, algorithms, and introductory courses in social sciences and psychology could be valuable in understanding and analyzing player statistics, team strategies, and fan psychology.

Algorithms to Learn:

Regression Analysis and Time Series Analysis for performance prediction and trend analysis.

Clustering Algorithms (e.g., K-Means) for player segmentation based on performance metrics.

Association Rule Learning for discovering interesting relations between different statistics.

Chess Engine Development and Optimization:

Relevance: As a chess enthusiast, you would find it rewarding to delve into the computational aspects of chess engines, exploring how computers simulate and strategize the game.

Coursework Connection: Your skills in algorithms and computer organization would be directly applicable in understanding and improving chess engine performance.

Algorithms to Learn:

Minimax Algorithm with Alpha-Beta Pruning for decision-making in chess.

Evaluation Functions for assessing chessboard positions.

Machine Learning techniques for strategic play based on historical game data.

Each of these topics leverages your academic background and personal interests, providing a rich ground for a comprehensive and engaging state-of-the-field project. The specific algorithms mentioned are integral to these areas and will enhance your understanding and ability to conduct in-depth research.

Lesson 3: Editing content to match writing goals

Virtual assistant response

Based on the analysis of your paper, here are specific areas of improvement aligned with the capstone course requirements:

1. Reduce Grandiose Language (Rule #1): The paper occasionally uses grandiose language (e.g., "revolutionize diagnostic procedures", "monumental leap"). While it's important to emphasize the significance of your topic, try to present information in a more academic and less marketing-oriented tone.

2. Clarify and Define Terms (Rule #3): Ensure all technical terms and acronyms are clearly defined when first introduced. For example, terms like "Sensitivity Encoding (SENSE)" and "Generalized Autocalibrating Partially Parallel Acquisitions (GRAPPA)" should be explained in simple terms to ensure clarity.

3. Avoid Over-Dramatization (Rule #4): The paper occasionally leans towards dramatic language (e.g., "groundbreaking", "transformative approach"). Aim for more objective and precise wording that conveys the technical details without exaggeration.

4. Improve Structure and Clarity (Rule #9): While section headers are used effectively, incorporating bulleted or numbered lists where appropriate can further enhance the readability and organization of the paper.

5. Enhance Figures and Diagrams (Rule #7): Ensure that all figures and diagrams are clearly labeled and include substantial captions. For example, the paper includes figures illustrating parallel processing and MRI components; ensure these are accompanied by detailed descriptions that explain their relevance to the text.

6. Strengthen the Conclusion (Rule #2): The conclusion should succinctly summarize key findings and reiterate the significance of the research in advancing MRI technology. Consider focusing on a few core takeaways rather than broadly summarizing the entire paper.

7. Concrete Examples (Rule #8): Use specific examples to illustrate the impact of parallel imaging in MRI. For instance, discussing a particular case study where parallel imaging significantly improved diagnostic outcomes can make the abstract concepts more tangible.

8. Proofread for Consistency and Grammar (Rule #5): Ensure consistent use of terms throughout the paper. Also, proofread for grammatical accuracy and coherence to enhance the overall quality of the writing.

9. Substantiate Claims (General Rule): Where strong claims are made, back them up with objective data, expert opinions, or related research findings. This adds credibility to your arguments and aligns with the academic rigor expected in CS373.

In summary, focusing on these areas will help align your paper more closely with the CS373 course's emphasis on clarity, conciseness, precision, and academic rigor.

Lesson 4: How to incentivize proper use of AI tools?

We embrace creative use of generative AI tools in capstone course for all aspects of the research process

We incentivize proper use by:

- Focusing on level of expertise evaluation
- Performing assessment using dialogue

No advantage to cheat

However, expect the use of AI for

- Perfect spelling, grammar, sentence, paragraph structure
- Compact writing
- Polished demos
- More

Theme 3 Takeaway: Your Own Personal Research and Writing Team

Authoring
- Brainstorm topics
- Generate outline
- Recursively provide more detail

Assessment
- Request AI review of content

Improvement
- Steel-manning other perspectives
- Refine content – understandable by a novice, TL;DR, etc.

Corporate executives have a research and writing team; now you do, too!

New to Gen AI Tools? Use the Powerful Question

Kids to Adults: Why? → Why? → Why?

You to AI Bot: How? → How? → How?

Prompt:

I <describe your job>. How can you help me do my job better?

Lessons from the Industrial Revolution

> *It wasn't the steam engine alone that caused the Industrial Revolution. It was the thousands of specific machines invented by skilled craftsmen that used steam to augment or do existing work that created the Revolution*.

Ethan Mollick

Reading Recommendation: One Useful Thing https://www.oneusefulthing.org
Blog by Ethan Mollick

Observations: Generative AI Tools and Learning

GenAI tools change everything: Yes, this sounds grandiose, but I believe it

- Helping with tedious tasks
- Upscaling our technical skills
- Fundamentally reshaping the research enterprise

Some impacts:

- Many mechanical skills and activities are being automated
- The importance of soft skills is increasing ; alignment is why
- Scientific progress limited by the imagination of thousands: GenAI is new steam engine

There is AI hype and anti-hype

- Some anti-hype is from small-scope constraints
- Some hype is from people who see beyond small-scope limitations
- Scaling productivity requires scaling out the scope of transformation

AI for Science

Sittin' on the dock of the bay, watching the tide roll away…

AI For Science: Exciting Frontier

Google GenCast: Probabilistic weather forecasting with machine learning

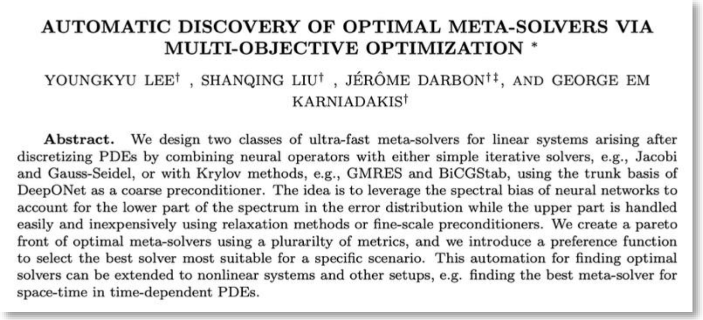
- https://www.nature.com/articles/s41586-024-08252-9
- https://deepmind.google/discover/blog/gencast-predicts-weather-and-the-risks-of-extreme-conditions-with-sota-accuracy
- https://www.nytimes.com/2024/12/04/science/google-ai-weather-forecast.html

Use of generative AI in high-dimensional complex systems

- https://www.nature.com/articles/s41467-024-53165-w.pdf

Use of generative AI in a multigrid solver

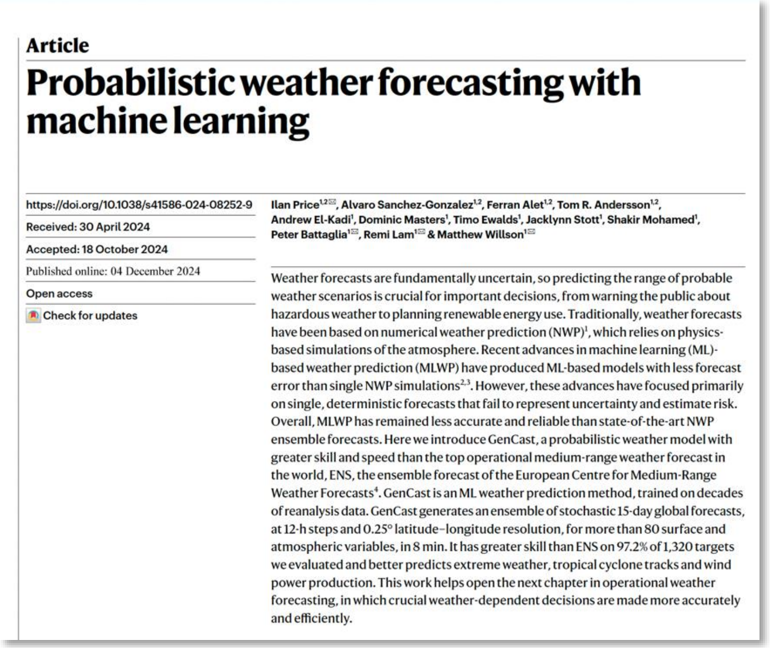
- https://arxiv.org/pdf/2412.00063



AUTOMATIC DISCOVERY OF OPTIMAL META-SOLVERS VIA MULTI-OBJECTIVE OPTIMIZATION *

YOUNGKYU LEE†, SHANQING LIU†, JÉRÔME DARBON†‡, AND GEORGE EM KARNIADAKIS†

Abstract. We design two classes of ultra-fast meta-solvers for linear systems arising after discretizing PDEs by combining neural operators with either simple iterative solvers, e.g., Jacobi and Gauss-Seidel, or with Krylov methods, e.g., GMRES and BiCGStab, using the trunk basis of DeepONet as a coarse preconditioner. The idea is to leverage the spectral bias of neural networks to account for the lower part of the spectrum in the error distribution while the upper part is handled easily and inexpensively using relaxation methods or fine-scale preconditioners. We create a pareto front of optimal meta-solvers using a plurarilty of metrics, and we introduce a preference function to select the best solver most suitable for a specific scenario. This automation for finding optimal solvers can be extended to nonlinear systems and other setups, e.g. finding the best meta-solver for space-time in time-dependent PDEs.



Article

Probabilistic weather forecasting with machine learning

https://doi.org/10.1038/s41586-024-08252-9

Received: 30 April 2024

Accepted: 18 October 2024

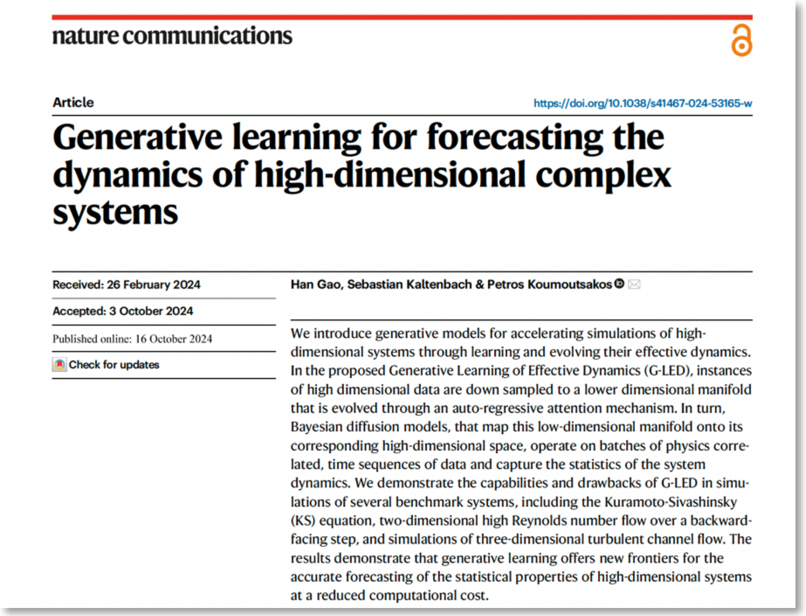
Published online: 04 December 2024

Open access

Check for updates

Ilan Price[1,2] ✉, Alvaro Sanchez-Gonzalez[1,2], Ferran Alet[1,2], Tom R. Andersson[1,2], Andrew El-Kadi[1], Dominic Masters[1], Timo Ewalds[1], Jacklynn Stott[1], Shakir Mohamed[1], Peter Battaglia[1] ✉, Remi Lam[1] ✉ & Matthew Willson[1] ✉

Weather forecasts are fundamentally uncertain, so predicting the range of probable weather scenarios is crucial for important decisions, from warning the public about hazardous weather to planning renewable energy use. Traditionally, weather forecasts have been based on numerical weather prediction (NWP)[1], which relies on physics-based simulations of the atmosphere. Recent advances in machine learning (ML)-based weather prediction (MLWP) have produced ML-based models with less forecast error than single NWP simulations[2,3]. However, these advances have focused primarily on single, deterministic forecasts that fail to represent uncertainty and estimate risk. Overall, MLWP has remained less accurate and reliable than state-of-the-art NWP ensemble forecasts. Here we introduce GenCast, a probabilistic weather model with greater skill and speed than the top operational medium-range weather forecast in the world, ENS, the ensemble forecast of the European Centre for Medium-Range Weather Forecasts[4]. GenCast is an ML weather prediction method, trained on decades of reanalysis data. GenCast generates an ensemble of stochastic 15-day global forecasts, at 12-h steps and 0.25° latitude–longitude resolution, for more than 80 surface and atmospheric variables, in 8 min. It has greater skill than ENS on 97.2% of 1,320 targets we evaluated and better predicts extreme weather, tropical cyclone tracks and wind power production. This work helps open the next chapter in operational weather forecasting, in which crucial weather-dependent decisions are made more accurately and efficiently.



nature communications

Article

https://doi.org/10.1038/s41467-024-53165-w

Generative learning for forecasting the dynamics of high-dimensional complex systems

Received: 26 February 2024

Accepted: 3 October 2024

Published online: 16 October 2024

Check for updates

Han Gao, Sebastian Kaltenbach & Petros Koumoutsakos ✉

We introduce generative models for accelerating simulations of high-dimensional systems through learning and evolving their effective dynamics. In the proposed Generative Learning of Effective Dynamics (G-LED), instances of high dimensional data are down sampled to a lower dimensional manifold that is evolved through an auto-regressive attention mechanism. In turn, Bayesian diffusion models, that map this low-dimensional manifold onto its corresponding high-dimensional space, operate on batches of physics correlated, time sequences of data and capture the statistics of the system dynamics. We demonstrate the capabilities and drawbacks of G-LED in simulations of several benchmark systems, including the Kuramoto-Sivashinsky (KS) equation, two-dimensional high Reynolds number flow over a backward-facing step, and simulations of three-dimensional turbulent channel flow. The results demonstrate that generative learning offers new frontiers for the accurate forecasting of the statistical properties of high-dimensional systems at a reduced computational cost.

Observations: AI for Science – very speculative, uncertain

New foundational models constructed from scientific data tokens

- Historical weather data image patches
- Fundamental particle position, velocity, temp, forces
- What else? Entity-relationship quantities from any observational setting

Other sources

- Language describes scientific phenomena, good enough for some situations
- Synthetic data from modeling and simulation

AI approaches seem to work well where models struggle most

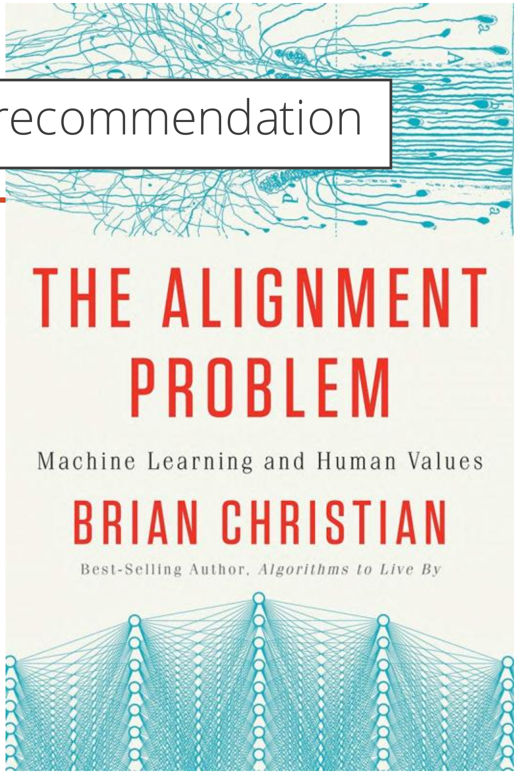
- Complex systems – weather, turbulence

Combination of AI/ModSim should be synergistic – best of both

- Lots to study, understand
- Very rapid change – big risk: disruptive change creating obsolescence

AI and Alignment

Reading recommendation



THE ALIGNMENT PROBLEM

Machine Learning and Human Values

BRIAN CHRISTIAN

Best-Selling Author, *Algorithms to Live By*

AI enables the creation of powerful tools

The consequences of design choices are much higher

Key concept: Alignment

Increased demand:

- Domain knowledge: What the tools are supposed to do
- Human knowledge: Understanding humans, our needs, how to design

Final Takeaways

1. **Mainstream generative AI tools** such as ChatGPT and GitHub Copilot **are already transforming how learning and research are conducted.** Most learning and research can benefit from these tools, and the academic community is only beginning to understand the potential and limits of these tools even as they become increasingly effective.

2. **Generative AI tools will continue to improve rapidly with significant and growing industry investments, most of which will meaningfully benefit the scientific community.** Foundational large-language models (LLMs) (pre-training) may be maturing but post-training and other specializations are still improving rapidly.

Final Takeaways

3. **As generative AI tools take over more** of how we produce research, the **remaining human tasks will be higher-level functions**. Deep domain expertise will have greater value than the mechanical aspects of producing software and its related artifacts.

4. There has been a lot of **hype and anti-hype with generative AI.** There is merit in the anti-hype perspective in that if you use generative AI tools to automate a specific task, many other tasks still require human effort. The key to qualitative improvements from generative AI tools is to **expand the scope of application so that improvements are made across many activities and at all levels of the organization.**

5. **AI for science capabilities are transforming predictive science and engineering.** We are just getting started. Like many new advances, the solution is likely not AI ***or*** ModSim, but AI ***and*** ModSim. ModSim domain experts are essential.