

# ML8 – generative models

Machine Learning – Tools and applications for policy – Lecture 9

Iman van Lelyveld – Michiel Nijhuis

DNB Data Science Hub



## ML8 – generative models

---

1. With the release of ChatGPT, generative models have gained center stage
2. What makes them tick?
3. What are the possibilities? What the dangers?

## ML8 – generative models

Generative models and LLMs

Hypothesis Generation

Will Generative Models change the world?

- No clips for this lecture

# Will Generative Models change the world?

---

- “Generative” refers to the ability of an AI algorithm to produce complex data.
- The alternative is “discriminative” AI, which chooses between a fixed number of options and produces just a single number
  - Example: choosing whether to approve a loan application
- More recently, with the emergence of foundational models (Bommasani et al. 2022), based on the transformer architecture for large language models (OpenAI 2023) , there is the idea that given large enough data, we can train foundational models which may learn emergent properties to “understand” their respective domains of knowledge
- Will this indeed materialize?

- Closed models
  - Open AI: GPT-4 (4.5 coming soon)
  - Google: BARD
- Open models
  - Meta: Llama 2
  - Google: Bert
  - Nomic: GPT4ALL
- semi-Closed
  - GPT4 - Enterprise
  - Microsoft Azure: GPT4

# Weaknesses of Large Language Models (LLMs)

---

- **Lack of Real Understanding:** may generate text that appears coherent but lacks genuine comprehension. Reliance on statistical patterns without true understanding.
- **Inability to Reason:** may struggle with higher-level reasoning tasks, often generating responses that are plausible-sounding but lack logical coherence
- **Bias and Fairness Issues:** can perpetuate or even exacerbate biases present in the training data. Can lead to unfair or discriminatory outputs, raising ethical concerns
- **Limited Contextual Understanding:** struggles with context over longer passages
- **Vulnerability to Adversarial Attacks:** are susceptible to carefully crafted input designed to deceive them, resulting in misleading or inappropriate responses
- **Resource Intensive:** Training/fine-tuning LLMs require significant computational resources raising barriers to entry
- **Environmental Impact:** The large-scale training of LLMs has environmental implications due to the substantial energy consumption associated with data centers
- **Domain Specificity:** trained on general text may not perform optimally in specific domains, requiring extensive fine-tuning for specific applications

# Weaknesses of Large Language Models (LLMs)

---

- **Lack of Real Understanding:** may generate text that appears coherent but lacks genuine comprehension. Reliance on statistical patterns without true understanding.
- **Inability to Reason:** may struggle with higher-level reasoning tasks, often generating responses that are plausible-sounding but lack logical coherence
- Bias and Fairness Issues: can perpetuate or even exacerbate biases present in the training data. Can lead to unfair or discriminatory outputs, raising ethical concerns
- Limited Contextual Understanding: struggles with context over longer passages
- Vulnerability to Adversarial Attacks: are susceptible to carefully crafted input designed to deceive them, resulting in misleading or inappropriate responses
- Resource Intensive: Training/fine-tuning LLMs require significant computational resources raising barriers to entry
- Environmental Impact: The large-scale training of LLMs has environmental implications due to the substantial energy consumption associated with data centers
- Domain Specificity: trained on general text may not perform optimally in specific domains, requiring extensive fine-tuning for specific applications

# What if we want the math to check out?

---

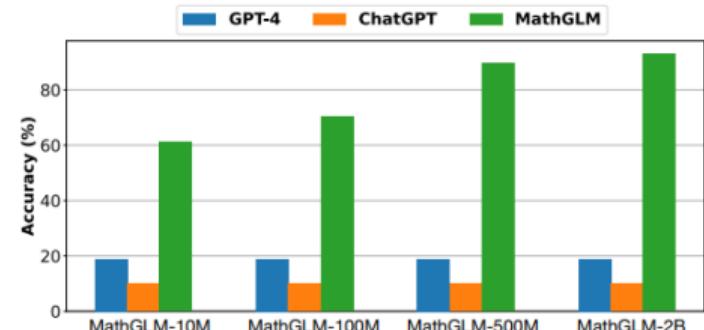
- LLM can serve up a **likely** response that is uncannily realistic but ...
- ... will attention layers be able to **build abstractions** on their own (and not regurgitating from abstractions it got ready-served by humans)?
- What if we need the answer to be **exactly correct**?

# What if we want the math to check out?

8

- LLM can serve up a **likely** response that is uncannily realistic but ...
- ... will attention layers be able to **build abstractions** on their own (and not regurgitating from abstractions it got ready-served by humans)?
- What if we need the answer to be **exactly correct**?
- Yang *et. al.* happy their LLM drastically improves math performance vs GPT.
- But would you be happy?

Overall performance



Source: Yang et al. 2023

# What if we want the math to check out?

---

- LLM can serve up a **likely** response that is uncannily realistic but ...
- ... will attention layers be able to **build abstractions** on their own (and not regurgitating from abstractions it got ready-served by humans)?
- What if we need the answer to be **exactly correct**?
- Yang *et. al.* happy their LLM drastically improves math performance vs GPT.
- But would you be happy?

Performance deteriorates at higher precision

Generalization	GPT4	ChatGPT	MathGLM-500M	MathGLM-2B
5-digit	6.67%	5.43%	83.44%	85.16%
6-digit	10.0%	2.94%	79.58%	78.17%
7-digit	3.33%	1.92%	71.19%	73.73%
8-digit	3.13%	1.43%	64.62%	67.69%
9-digit	6.90%	1.57%	66.66%	69.60%
10-digit	3.33%	1.45%	49.55%	65.77%
11-digit	0%	0%	42.98%	57.89%
12-digit	6.90%	1.33%	27.38%	41.05%

Performance of a **calculator** is **100%**

Source: Yang et al. 2023

# What if we want the math to check out?

---

- LLM can serve up a **likely** response that is uncannily realistic but ...
- ... will attention layers be able to **build abstractions** on their own (and not regurgitating from abstractions it got ready-served by humans)?
- What if we need the answer to be **exactly correct**?
- Yang *et. al.* happy their LLM drastically improves math performance vs GPT.
- But would you be happy?

## Examples of mistakes

Input	OutPut	Ground Truth
3468*4046/7424=	14031528/7424=18 <b>89.901400862069</b>	14031528/7424=18 <b>90.0226293103449</b>
(3626*8919)/8861=	323 <b>30294/8861</b> =3648.605574991536	323 <b>40294/8861</b> =3649.7341157882856
7715/4791*7691-1968*9155=	1.610311*7691-1968*9155=12384. <b>801801</b> -1968*9155=12384.801801-18017040=	1.610311*7691-1968*9155=12384. <b>9018993</b> -1968*9155=12384.9018993-18017040=
1968*9155=	18004655.198199	18004655.098106066
(4059+7011.8718)-	(4059+7011.8718)-	(4059+7011.8718)-
4038.22*847.15907=	4038.22*847.15907=11070.8718-	4038.22*847.15907=11070.8718-
	4038.22*847.15907=11070.8718-	4038.22*847.15907=11070.8718-
	342 <b>014.6996554</b> =-3408943.8278554	342 <b>014.6996554</b> =-3409943.8278554003
7499-5747.91007/-5438*-	7499-5747.91007/5438*439=7499-439=	7499-5747.91007/5438*439=7499-
439=	1.05 <b>70081040823832</b> *439=7499-464.026576921662=7034.973442307834	1.05 <b>6997715</b> *439=7499-464.0184848713=7034.981515128724
3868*6735*5755+3741-	26050980*5755+3741-	26050980*5755+3741-
7533=	7533= <b>159923389900</b> +3741-	7533= <b>149923389900</b> +3741-
	7533= <b>159923393641</b> -7533=159923386108	7533= <b>149923393641</b> -7533=149923386108

Table 15: Some failed examples generated by MathGLM.

Source: Yang et al. 2023

# Is this thinking?

9

- The rooms in my house are painted blue or white or yellow
- Yellow paint fades to white within a year
- In two years time I want them all to be white
- What should I do and why?

Source: [Geoffrey Hinton](#)

IM

The rooms in my house are painted blue or white or yellow. Yellow paint fades to white within a year. In two years time I want them all to be white. What should I do if I want to exert the least effort and why?



If you want all the rooms in your house to be white in two years while exerting the least effort, you should follow this plan:

**1. Wait for One Year:**

- Do not take any action during the first year. This allows the yellow paint to fade to white on its own.

**2. Evaluate the Rooms:**

- After the first year, evaluate the rooms to see if the yellow paint has faded sufficiently. If it has not faded completely, proceed to the next step. If the yellow paint has already faded, you're done.

**3. Repaint Yellow Rooms:**

- If the yellow paint has not faded to white after one year, repaint only the yellow rooms with white paint. This will ensure that all rooms are white within the next year.

This plan exerts the least effort because it minimizes the amount of actual painting you need to do. You only need to repaint the rooms with yellow paint. The rooms already painted white or blue remain untouched. This saves time, energy, and resources compared to repainting all the rooms immediately.



DataScience  
Hub

# Is this thinking?

- The rooms in my house are painted blue or white or yellow
- Yellow paint fades to white within a year
- In two years time I want them all to be white
- What should I do and why?

Source: [Geoffrey Hinton](#)

But this is **different from Winton ...**

"Godfather of AI" Geoffrey Hinton: The 60 Minutes Interview

Watch 60 Minutes



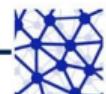
The rooms in my house are painted blue or white or yellow. Yellow paint fades to white within a year. In two years time I want them all to be white. What should I do and why?



Assuming that blue paint does not fade to white over time and you want all rooms to be white in two years, here's what you should do:

1. The rooms painted in white: You don't have to do anything to these rooms since they are already white.
2. The rooms painted in yellow: You don't need to repaint these rooms either, because as per your information, yellow paint fades to white within a year. So, in two years, they will be white.
3. The rooms painted in blue: These rooms need to be repainted with white paint because the blue paint will not fade to white on its own.

You should do this because if you leave the blue rooms as they are, they won't automatically turn white. And if you paint the yellow rooms white, the yellow paint that fades to white within a year could potentially mix with the white paint and give a different shade, or you'd be wasting resources by painting a room that will naturally fade to the desired color. Thus, only painting the blue rooms white is the most efficient course of action to achieve your goal.



DataScience  
Hub

# Is this thinking?

---

- The rooms in my house are painted blue or white or yellow
- Yellow paint fades to white within a year
- In two years time I want them all to be white
- What should I do and why?

Source: [Geoffrey Hinton](#)

... and I can't replicate the advice on the [blue room](#)

IM

The rooms in my house are painted white, yellow or blue. Yellow paint fades to white in two years. If I want my walls to be white eventually, what should I do and why?



If you want your walls to eventually be white, you should initially paint them with blue paint. Here's why:

1. Blue paint does not fade to white: Since you mentioned that yellow paint fades to white in two years, using yellow paint would not achieve your goal of eventually having white walls.
2. Blue paint remains blue: Blue paint does not naturally fade to white, so the color will stay consistent over time.
3. Easier transition to white: If your ultimate goal is to have white walls, starting with blue paint is a good choice because white is a lighter color that can be easily painted over blue. You can simply paint over the blue with white paint when you're ready for the change, and it should cover the blue effectively with minimal effort.

So, to achieve white walls eventually, you should begin with blue paint, and when you're ready to transition to white, you can easily paint over the blue with white paint to achieve the desired result.



DataScience  
Hub

# Is this thinking?

---

- The rooms in my house are painted blue or white or yellow
- Yellow paint fades to white within a year
- In two years time I want them all to be white
- What should I do and why?

Source: [Geoffrey Hinton](#)

IM

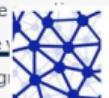
The rooms in my house are painted white, yellow or blue. Yellow paint fades to white in two years. I want my walls to be white eventually. what should I do and why?



To ensure that your walls eventually become white, you can follow a strategic plan for repainting your rooms over time. Here's what you should do and why:

1. **Start with White:** Begin by painting all the rooms in your house white. This will be your base color and the color you ultimately want to achieve.
2. **Monitor Fading:** Since you mentioned that yellow paint fades to white in two years, keep track of the condition of the yellow-painted walls.
3. **Repaint Yellow to Yellow:** After approximately two years, when the yellow paint has faded, repaint those yellow walls with yellow paint again. This ensures that you maintain the yellow color you desire.
4. **Repeat as Needed:** Continue to repaint any yellow walls with yellow paint every two years, or as soon as you notice significant fading. This step is essential to maintain the yellow color and prevent the walls from turning completely white.
5. **Avoid Painting Blue:** Do not repaint the blue walls unless you want them to turn white eventually. Blue paint is not fading to white, so leaving it as is will help you achieve your goal of having white walls eventually.

The reason for this approach is that yellow paint fades to white naturally transforms into the color you want. By repainting the periodically, you maintain the yellow color while allowing it to go Meanwhile, the blue walls are left untouched, so they will remain blue, as you did not specify.



DataScience  
Hub

# Is this thinking?

9

- The rooms in my house are painted blue or white or yellow
- Yellow paint fades to white within a year
- In two years time I want them all to be white
- What should I do and why?

Source: [Geoffrey Hinton](#)

IM

The rooms in my house are painted white or yellow or blue. Yellow paint fades to white within a year. In two years time I want them all to be white. What should I do and why?



To have all the rooms in your house be white in two years, you should consider repainting the rooms that are currently yellow. Here's what you should do and why:

1. **Repaint Yellow Rooms:** Since you mentioned that yellow paint fades to white within a year, you'll need to repaint the rooms that are currently yellow. This will ensure that they maintain a consistent white color over the two-year period.
2. **Select the Right White Paint:** When repainting, choose a good quality white paint that is durable and long-lasting. It's important to select a paint that doesn't fade or discolor easily.
3. **Proper Preparation:** Before applying the new paint, make sure to properly prepare the surfaces by cleaning and sanding any imperfections. This will help the new paint adhere better and last longer.
4. **Apply Primer:** Consider using a primer before applying the white paint. Primer helps create a smooth and consistent surface for the paint to adhere to, and it can also enhance the durability of the paint.
5. **Maintenance:** In addition to repainting, you may want to address any factors that could cause paint to fade or deteriorate quickly. This includes controlling humidity, and addressing any underlying issues.



DataScience  
Hub

The reason for repainting the yellow rooms is that yellow paint fades to white within a year.

# Is this thinking?

- The rooms in my house are painted blue or white or yellow
- Yellow paint fades to white within a year
- In two years time I want them all to be white
- What should I do and why?

Source: [Geoffrey Hinton](#)

IM

The rooms in my house are painted blue or white or yellow. Yellow paint fades to white within a year. In two years time I want them all to be white. What should I do and why?



To have all the rooms in your house painted white in two years, you should take the following steps:

1. Repaint the Yellow Rooms:

Since you mentioned that yellow paint fades to white within a year, you should repaint the rooms that are currently yellow. This will ensure that they turn white naturally over the next year. You can select a high-quality white paint to get the desired result.

2. Monitor and Repaint as Needed:

Keep an eye on the rooms that are painted yellow as the year progresses. If you notice that the yellow color has not faded completely to white, you may need to touch up or repaint those rooms as necessary.

3. Maintain the White Rooms:

For the rooms that are already painted white, you should ensure they remain in good condition. Regular cleaning and maintenance will help keep the white color fresh and vibrant.

4. Plan for Timing:

Make sure you plan the repainting of the yellow rooms well ahead of your two-year deadline. This allows for any necessary touch-ups and ensures all rooms are white by the specified timeframe.



DataScience  
Hub

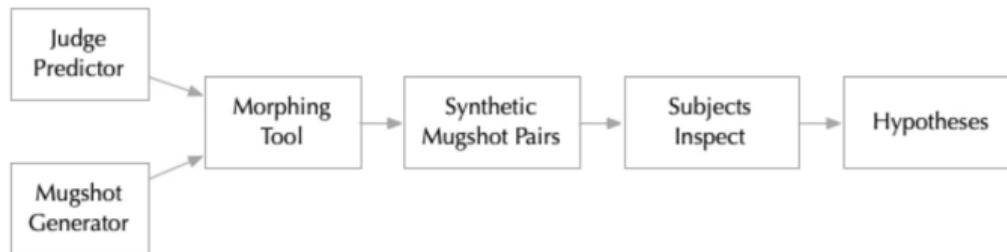
## ML8 – generative models

Generative models and LLMs

Hypothesis Generation

Will Generative Models change the world?

- Ludwig and Mullainathan (2023): model judges decisions to keep people in jail awaiting trial. Key question: how dangerous is the defendant?
- Components: labels (judges' decisions) + defendant info (including mugshot) + Mechanical Turk to label morphs
- Change the mug shots in the direction of more jail time and then find a description for the change



Source: Ludwig and Mullainathan (2023).



# Illustrative facial images



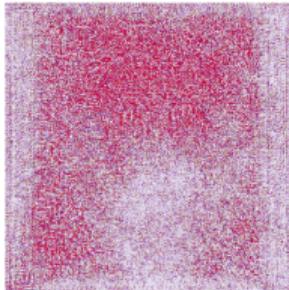
Source: Ludwig and Mullainathan (2023). Synthetic faces.

# Communicating changes in a known facial feature: Age

---



(a) Initial face



(b) Saliency map



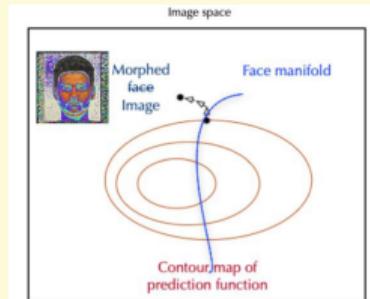
(c) Naive age-morphed image



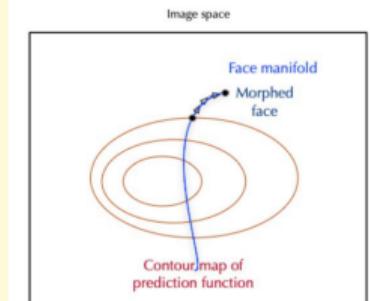
(d) Morphs from our procedure

Source: Ludwig and Mullainathan (2023). Synthetic faces.

# Making sure that the morphed face makes sense



(a) Naïve morphing leads off manifold and results in non-faces



(b) Our procedure stays on manifold and morphs are faces

Source: Ludwig and Mullainathan (2023).

# What morphing does to jail time

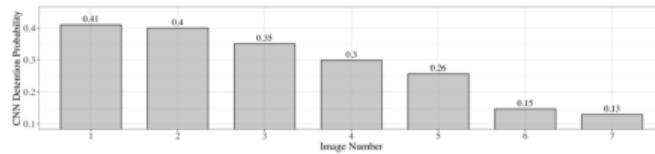
---



(a) Side-by-side mugshot detection morphs with detention probabilities of 0.41 and 0.13 respectively



(b) Transformations of the face along selected steps of the morphing process



(c) Detention-probabilities for images in panel (b)

Source: Ludwig and Mullainathan (2023).

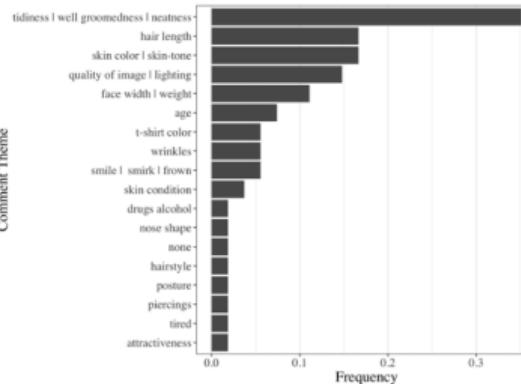


Higher Predicted Detention Risk

Lower Predicted Detention Risk



(a) A word cloud of the comments



(b) Frequencies of comments by theme

Source: Ludwig and Mullainathan (2023).

- ML can be a powerful tool in generating hypothesis
- Propose an approach to generate hypotheses
- Traditional methods may miss non-linear patterns
- ML models can discover complex relationships
- ML complements rather than replaces human intuition
- Use of ML requires domain expertise and interpretability

## Case A: Would this work in finance?

---

18

Can you think of a way to implement this approach in finance?

- What is the labelled feature? How would you obtain the data?
- What is equivalent to the picture? e.g. something highly dimensional that can be morphed
- How can you give a name to the newly found driver?
- Prepare an elevator pitch

## ML8 – generative models

Generative models and LLMs

Hypothesis Generation

Will Generative Models change the world?

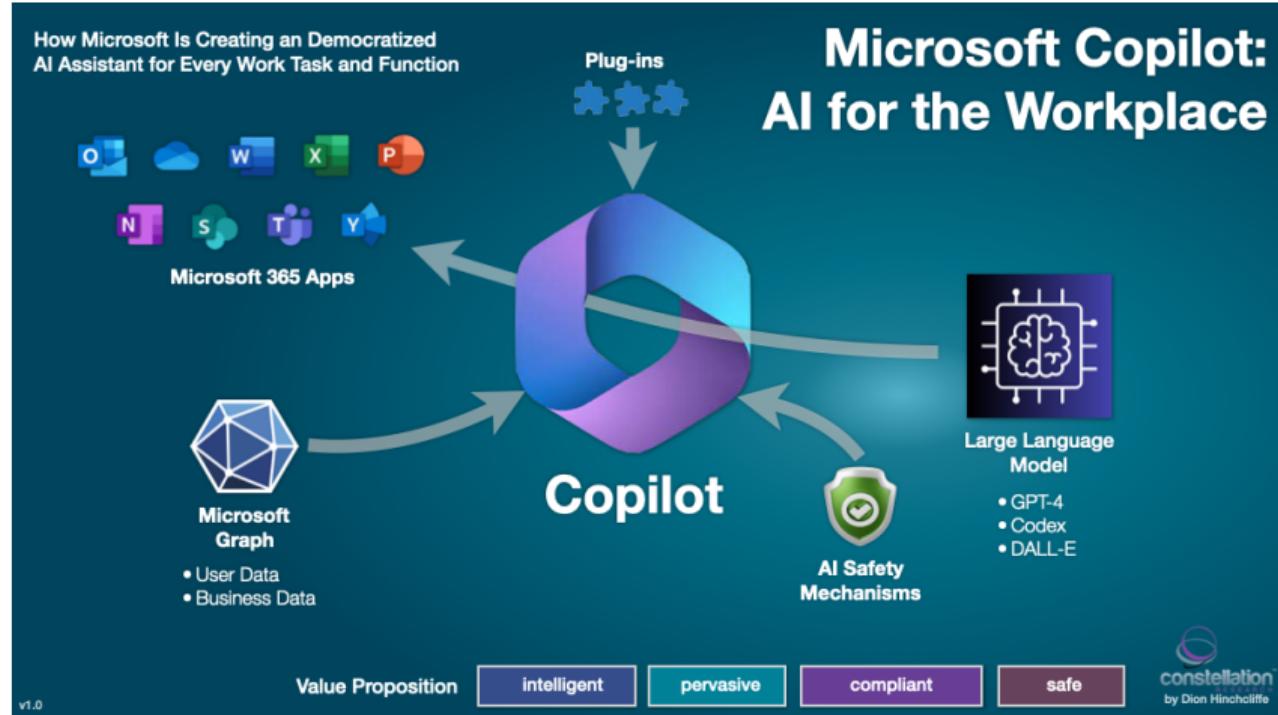
# AI could close the loop for scientific discovery

---

20

- AI and machine learning
  1. can operate as extractors of information
  2. can combine several parameter-dependent steps into a single (complex) optimization problem. See [AlphaFold](#).
  3. could potentially discover and learn novel scientific laws
- Zenil et al. [2023](#) argue for AI-driven, automated, generative, closed-loop approach to scientific discovery, including self-driven hypothesis generation and open-ended autonomous exploration of the hypothesis space





Can you think of a way that Gen AI will change finance?

- How would you make money? How would you obtain the data?
- Most of finance is a zero sum game. Would your scheme redistribute or lead to a more efficient market?
- Would there be external effects? Specifically, discuss how regulation could ameliorate negative side effects
- Prepare an elevator pitch

In this lecture we covered:

1. Strength and weaknesses of Gen AI
2. Can ML be Useful in generating hypotheses?

-  Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., ... Liang, P. (2022). On the Opportunities and Risks of Foundation Models [arXiv:2108.07258 [cs]].
-  Ludwig, J., & Mullainathan, S. (2023). Machine Learning as a Tool for Hypothesis Generation.
-  OpenAI. (2023). GPT-4 Technical Report [arXiv:2303.08774 [cs]].
-  Yang, Z., Ding, M., Lv, Q., Jiang, Z., He, Z., Guo, Y., Bai, J., & Tang, J. (2023). GPT Can Solve Mathematical Problems Without a Calculator [arXiv:2309.03241 [cs] version: 2]. ArXiv e-prints. Retrieved October 21, 2023, from <http://arxiv.org/abs/2309.03241>

-  Zenil, H., Tegnér, J., Abrahão, F. S., Lavin, A., Kumar, V., Frey, J. G., Weller, A., Soldatova, L., Bundy, A. R., Jennings, N. R., Takahashi, K., Hunter, L., Dzeroski, S., Briggs, A., Gregory, F. D., Gomes, C. P., Rowe, J., Evans, J., Kitano, H., & King, R. (2023). The Future of Fundamental Science Led by Generative Closed-Loop Artificial Intelligence [arXiv:2307.07522 [cs]]. Retrieved October 7, 2023, from <http://arxiv.org/abs/2307.07522>