DeepSeek, Low-cost LLM, Bias

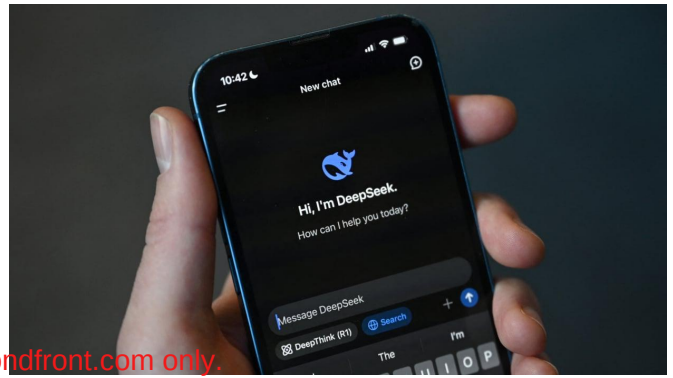
Agenda

- What is DeepSeek?
- Why Do We Care?
- Technical Details
  - Architecture
  - Chain of Thought and Reinforcement Learning
  - Mixture of Experts
  - Distillation
- DeepSeek Model List
- Impact
- Bias
- Questions

What is DeepSeek?

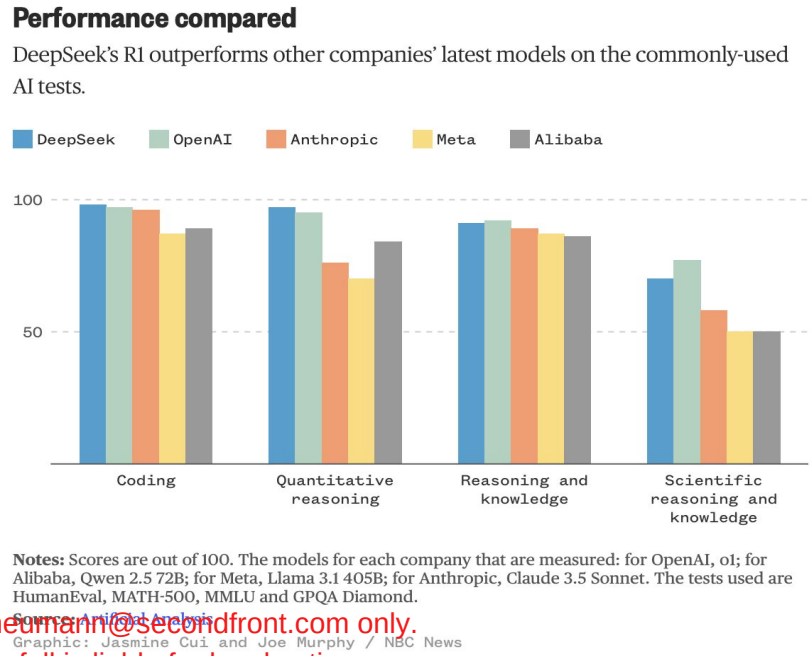
An open-source model, chatbot service and an intelligent agent framework.

- DeepSeek is a Chinese Artificial Intelligence startup
- January 23rd, 2025 - Launched a state-of-the-art generative AI model
- Delivers results comparable to those developed by giants like OpenAI and Meta
- And claims to do so at much lower operating costs



Why Do We Care?

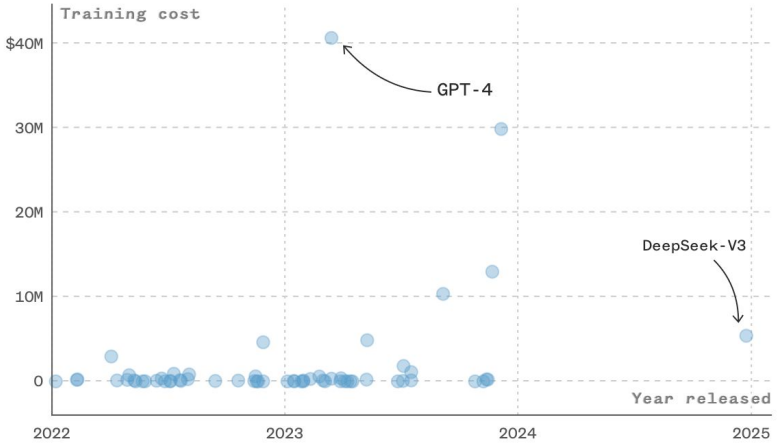
- Outperform ChatGPT and other state of the art models in various benchmarks
- ~20x Reduced Model Training Costs
- ~20x Cheaper LLM API Access
- Open-source access to the raw models
- $1 Trillion in market value lost in a single day



Why Do We Care?

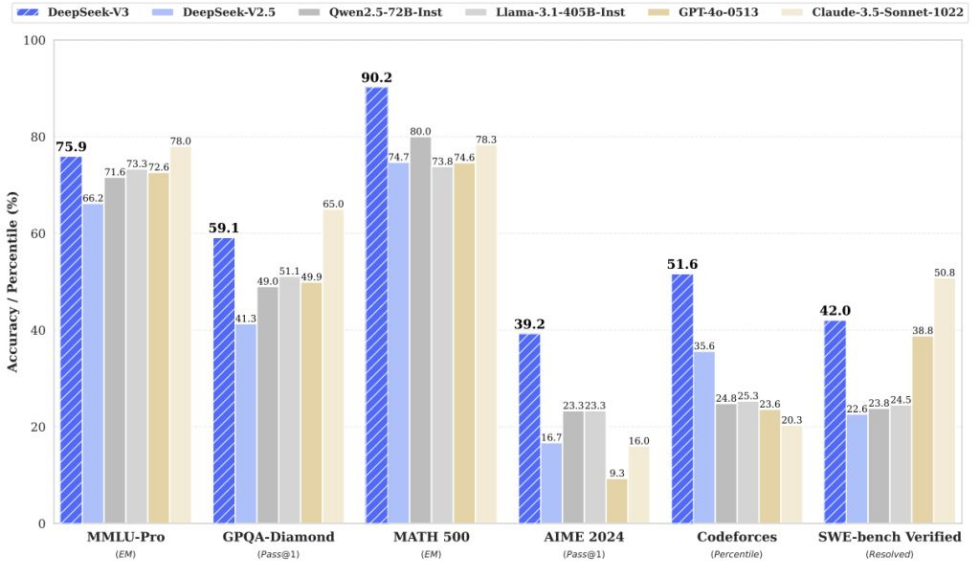
Spectrum of costs

Training costs have been rising throughout the industry, but the price for DeepSeek-V3 – the precursor to R1 – was significantly lower than other popular AI models.



Source: Epoch AI

Graphic: Jasmine Cui and Ash Reynolds / NBC News

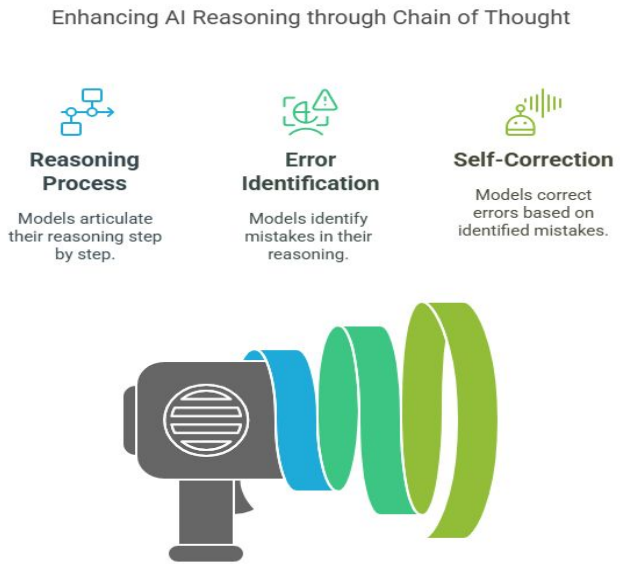


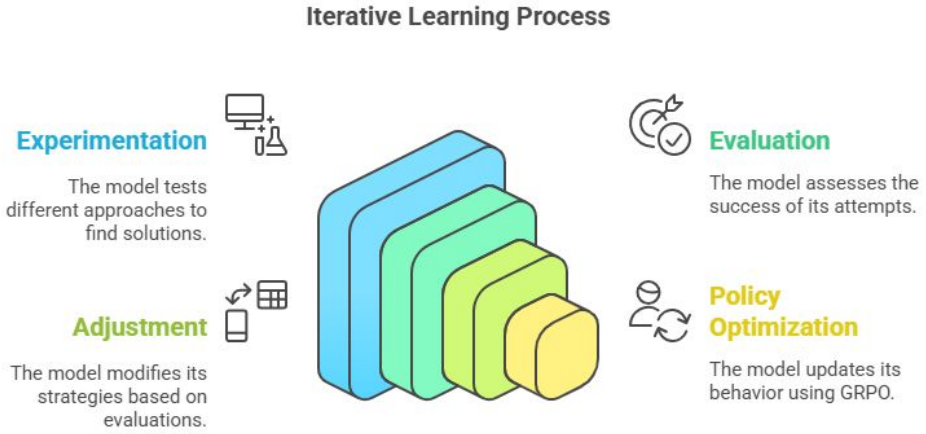
DeepSeek AI: Architecture

- **DeepSeek R1** uses a **Mixture-of-Experts (MoE) architecture** with 671 billion parameters but only activates 37 billion per query, optimizing computational efficiency. It also uses **reinforcement learning (RL) post-training** to enhance reasoning. DeepSeek was trained in 55 days on 2,048 Nvidia H800 GPUs at a cost of $5.5 million, significantly less than ChatGPT's training expenses.

- **ChatGPT** uses a **dense model architecture** with 1.8 trillion parameters and is optimized for versatility in language generation and creative tasks. It is built on OpenAI's GPT-4o framework and requires massive computational resources, estimated at $100 million+ for training

DeepSeek AI: Chain of Thought and Reinforcement Learning





https://medium.com/@tahirbalarabe2/deepseek-r1-explained-chain-of-thought-reinforcement-learning-and-model-distillation-0eb165d928c9

How many “r”s in strawberry?

<think>
Alright, so I need to figure out how many "r"s are in the word "strawberry." Hmm, okay, let's start by spelling out the word to make it easier to visualize. The word is S-T-R-A-W-B-E-R-R-Y.

First, I'll write it down to see each letter clearly. That way, I can count them without missing any. So, S, T, R, A, W, B, E, R, R, Y.

Now, I'll go through each letter one by one and keep a tally of how many times the letter "R" appears. Starting with S – that's not an R. Next is T – also not an R. Then we have R – that's the first R. After that is A, W, B, and E – none of those are Rs. Now, I come to another R – that's the second R. Wait, there's another R right after that – so that's the third R.

Hold on, let me double-check to make sure I didn't count any extra Rs. So, looking back: S, T, R (1), A, W, B, E, R (2), R (3), Y. Yeah, it looks like there are three Rs in "strawberry."

I think I've got it now. The word "strawberry" has three "R"s in it.
</think>

The word "strawberry" contains three "R"s.

How much 5 × (2 + 10)?

<think>

Alright, so I have this math problem here: 5 × (2 + 10). Hmm, it's been a while since I did any real math, but I think I can handle this. Okay, let's break it down. First, I know that parentheses mean I need to do what's inside them first. So, inside the parentheses, I have 2 plus 10. Alright, 2 plus 10 is 12. Got that part.

Now, the problem is 5 multiplied by that result. So, I take the 5 and multiply it by 12. Let's see, 5 times 12... Well, I know that 5 times 10 is 50, and 5 times 2 is 10, so if I add those together, that's 60. So, 5 × (2 + 10) should equal 60.

Wait a minute, let me double-check that. Sometimes I get confused with the order of operations, but I think I did it right. Parentheses come first, so 2 + 10 is definitely 12. Then multiplying 5 by 12 should be straightforward. Yeah, 5 × 12 is 60. Okay, I'm pretty confident about that.

But just to be sure, maybe I can think of it another way. 5 times 2 is 10, and 5 times 10 is 50. Adding those together gives me 60 again. So, yeah, that checks out. I guess the answer is 60. Not too bad for someone who's a bit rusty with math!

Another way to look at it could be by distributing the multiplication over the addition. So, 5 × (2 + 10) is the same as (5 × 2) + (5 × 10). That's 10 + 50, which also adds up to 60. Okay, so that method works too. I think I've got it figured out.

Maybe I can visualize it to make it even clearer. If I have 5 groups of something, and each group has 2 plus 10 items, that's like having 5 groups of 12 items each. So, 5 groups times 12 items in each group equals 60 items altogether. That image helps solidify the concept for me.

I wonder if there's a real-world example, I can relate this to. Like, if I'm buying strawberries and each basket has 2 small baskets and 10 large baskets, and I want to buy 5 sets of these baskets, how many baskets would I have in total? That would be 5 times (2 + 10), which is 60 baskets. Okay, that makes sense.

But wait, what if I forgot about the order of operations? Let's say I just went left to right without considering the parentheses. Then I'd do 5 × 2 first, which is 10, and then add 10, getting 20. But that's not right because parentheses take priority. So, it's crucial to remember to do what's inside the parentheses first.

Maybe that's why some people find math confusing—they might forget the order of operations. It's easy to overlook, especially when you're just starting out. But with practice, it becomes second nature. I should probably remind myself of PEMDAS: Parentheses, Exponents, Multiplication and Division, Addition and Subtraction.

Applying that here, parentheses come first, so 2 + 10, then multiplication. No exponents or division involved in this problem. So, yeah, 5 × (2 + 10) equals 60, no doubt.

I guess another way to think about it is using the distributive property, which I sort of touched on earlier. It helps break down the problem into smaller, more manageable parts. Especially useful when dealing with more complex expressions.

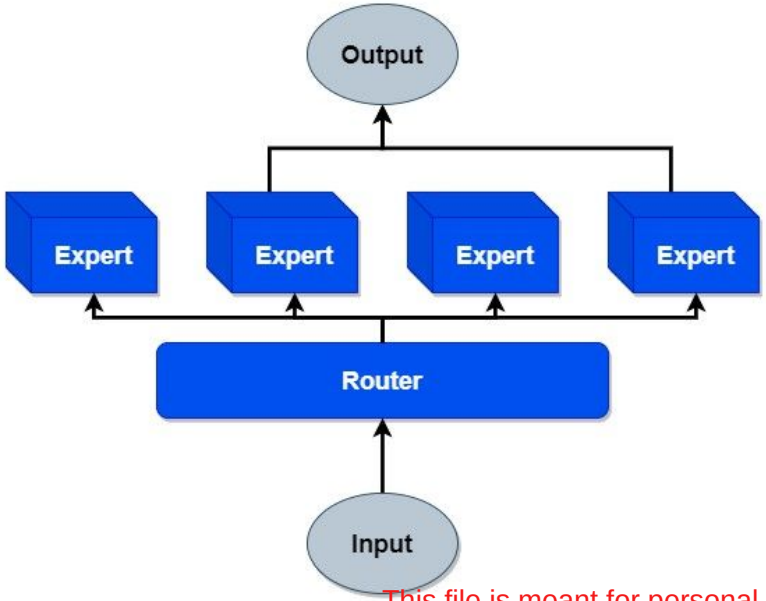
But for this particular problem, it's straightforward. Just remember the basics, and you're golden. I think I've covered all my bases here. 5 × (2 + 10) equals 60. I'm confident in that answer.

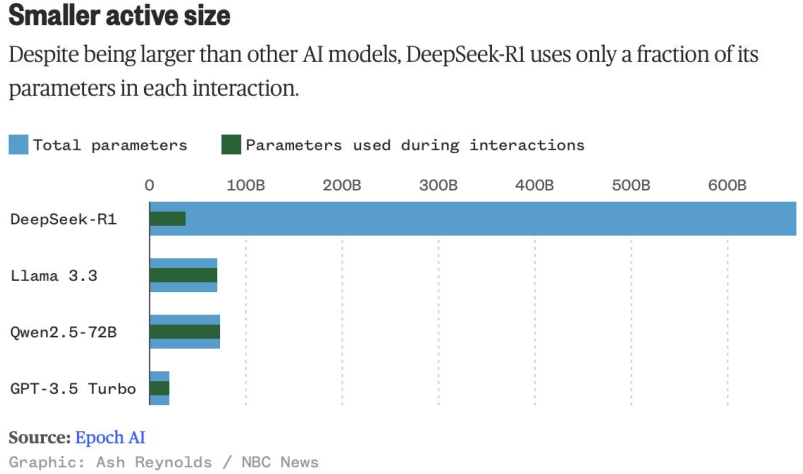
</think>

The result of the expression $5 \times (2 + 10)$ is 60.

DeepSeek AI: Mixture of expert

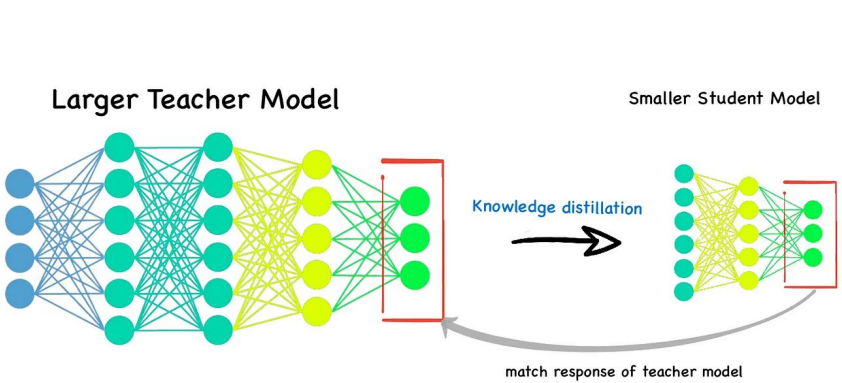
- Mixture of Experts (MoE) is a machine learning technique that combines multiple expert models to improve the performance.
- Introduced by Michael I. Jordan, Robert A. Jacobs, and Geoffrey E. Hinton in the paper “Hierarchical Mixtures of Experts and the EM Algorithm” published in 1991.





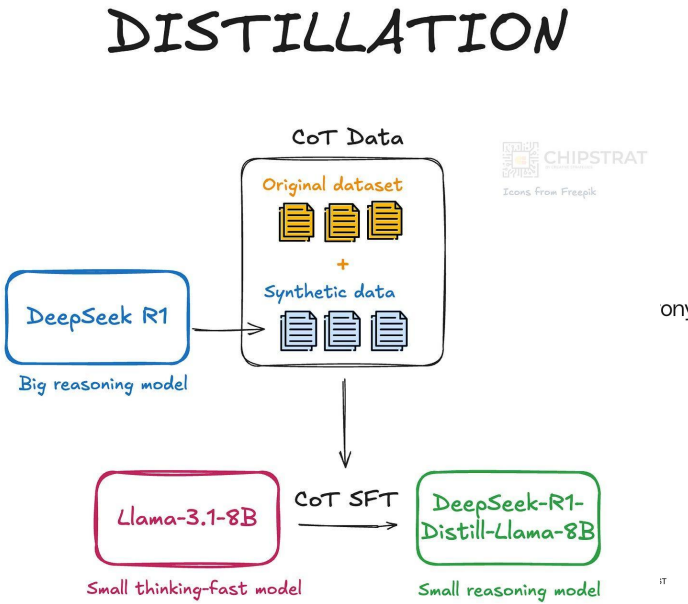
DeepSeek AI: Distillation

- DeepSeek researchers distilled their model into Llama 3 and Qwen. The surprising part? The smaller models sometimes performed better than the original. This makes AI far more accessible. Instead of needing a supercomputer, you can run a powerful model on a single GPU.



Model A
Accuracy: 98%
Runtime: 1.5 sec
Size: 200 mb

Model B
Accuracy: 96%
Runtime: 0.2 sec
Size: 20 mb .



DeepSeek AI: Distillation

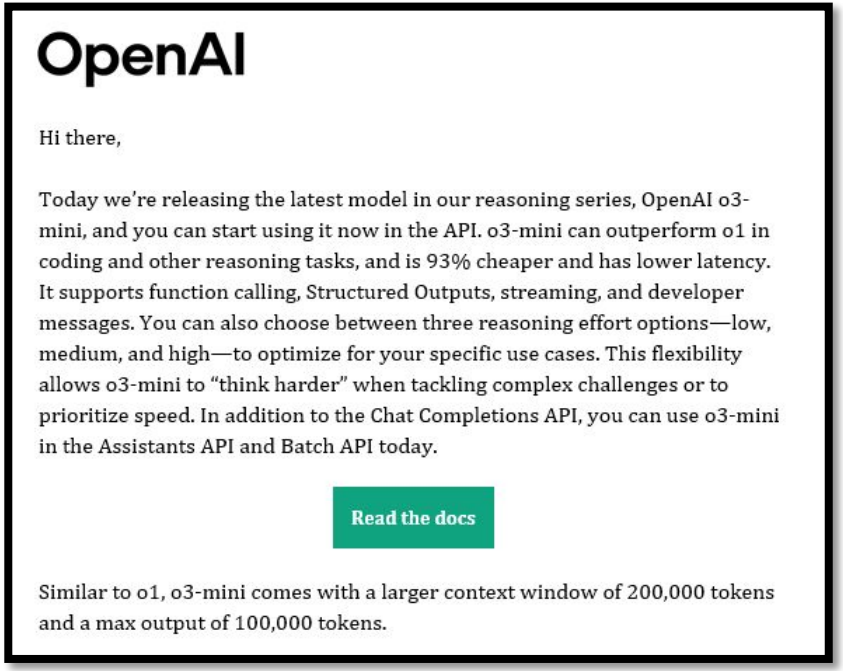
- DeepSeek researchers distilled their model into Llama 3 and Qwen. The surprising part? The smaller models sometimes performed better than the original. This makes AI far more accessible. Instead of needing a supercomputer, you can run a powerful model on a single GPU.

DeepSeek Model List

- **DeepSeek Coder**: Released in NOV 2023, this is the company's first open-source model designed specifically for coding-related tasks.
- **DeepSeek LLM**: Released in DEC 2023, this is the first version of the company's general-purpose model.
- **DeepSeekMath**: Repo created in ~MAR 2024, seeks to improve mathematical reasoning in open language models.
- **DeepSeek-VL**: Released in MAR 2024, an open-source Vision-Language (VL) Model designed for real-world vision and language understanding applications.
- **DeepSeek-V2**: Released in MAY 2024, this is the second version of the company's LLM, focusing on strong performance and lower training costs.
- **DeepSeek-Coder-V2**: Released in JUL 2024, this is a 236 billion-parameter model offering a context window of 128,000 tokens, designed for complex coding challenges.
- **DeepSeek-VL2**: Released in DEC 2024, builds upon DeepSeek-VL.
- **DeepSeek-V3**: Released in DEC 2024, DeepSeek-V3 uses a mixture-of-experts architecture, capable of handling a range of tasks. The model has 671 billion parameters with a context length of 128,000.
- **DeepSeek-R1**: Released in JAN 2025, this model is based on DeepSeek-V3 and is focused on advanced reasoning tasks directly competing with OpenAI's o1 model in performance, while maintaining a significantly lower cost structure. Like DeepSeek-V3, the model has 671 billion parameters with a context length of 128,000.
- **Janus-Pro-7B**: Released in JAN 2025, Janus-Pro-7B is a vision model that aims to understand and generate images.

Impact

- Narrows OpenAI's Lead
  - They were likely already doing this closed source
- OpenAI Released o3-mini, a new cheaper model
- Open-source knowledge and quality pushed forward
- Fine-tuning is now cheaper and more accessible
  - TinyZero: https://github.com/Jiayi-Pan/TinyZero (by UC Berkley)
  - Finetune for $30, "Re-experience the 'Ahaa' moment"
- Demand for GPUs will likely **NOT** go down



OpenAI

Hi there,

Today we're releasing the latest model in our reasoning series, OpenAI o3-mini, and you can start using it now in the API. o3-mini can outperform o1 in coding and other reasoning tasks, and is 93% cheaper and has lower latency. It supports function calling, Structured Outputs, streaming, and developer messages. You can also choose between three reasoning effort options—low, medium, and high—to optimize for your specific use cases. This flexibility allows o3-mini to "think harder" when tackling complex challenges or to prioritize speed. In addition to the Chat Completions API, you can use o3-mini in the Assistants API and Batch API today.

Read the docs

Similar to o1, o3-mini comes with a larger context window of 200,000 tokens and a max output of 100,000 tokens.

Key Takeaway from Yann LeCun (Chief of AI, Meta)

"To people who see the performance of DeepSeek and think: 'China is surpassing the US in AI.' You are reading this wrong. The correct reading is: 'Open-source models are surpassing proprietary ones.'

DeepSeek has profited from open research and open source (e.g., PyTorch and Llama from Meta). They came up with new ideas and built them on top of other people's work. Because their work is published and open source, everyone can profit from it. That is the power of open research and open source."

https://www.forbes.com/sites/luisromero/2025/01/27/chatgpt-deepseek-or-llama-metas-lecun-says-open-source-is-the-key/

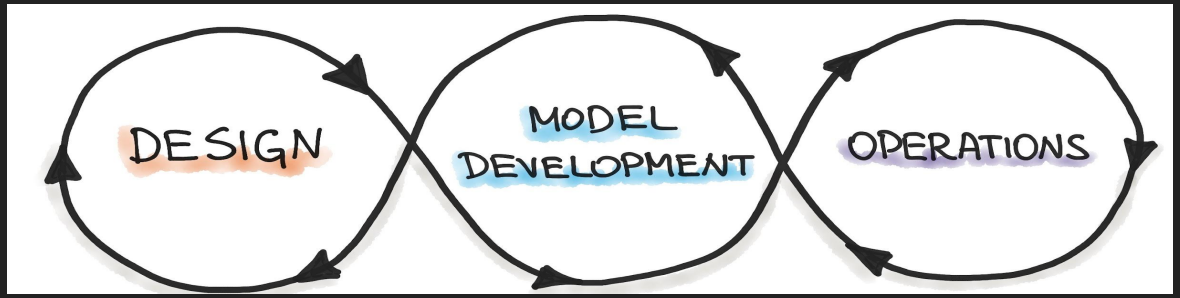
Mathematical Definition of Bias

Bias measures the **systematic** discrepancy between the predictions made by our machine learning model and the true outcomes it aims to predict

Biased Models

- Models are limited functions, and subject to bias
- Accurate models don’t always mean stable and robust models
- Understanding bias, how it’s introduced and how it can be used, is an important component of making worthwhile Machine Learning (ML) investments

AI Model Bias Results from Limitation



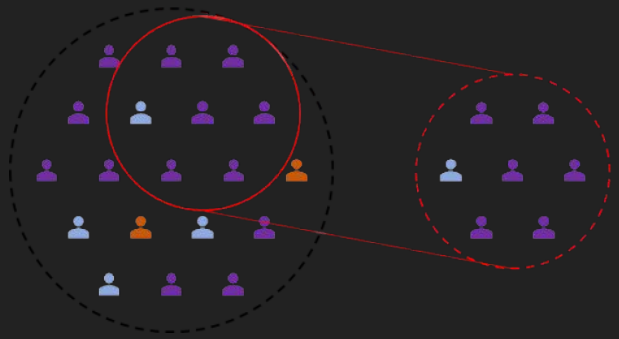
Sampling	Optimization	Cognitive
Measurement	Architecture	Contextual
Cleaning	Feature Engineering	Societal

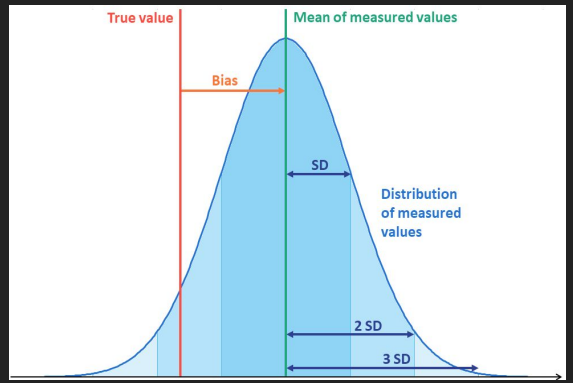
Design / Data Biases

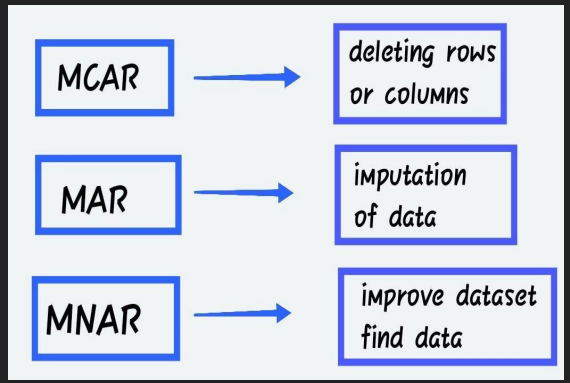
Sampling
- Convenience
- Historical
- Split/Test

Measurement
- Misrepresentation
- Labels
- Measurement tool

Cleaning
- Data missing not completely at random (MCAR)
- Abnormalities



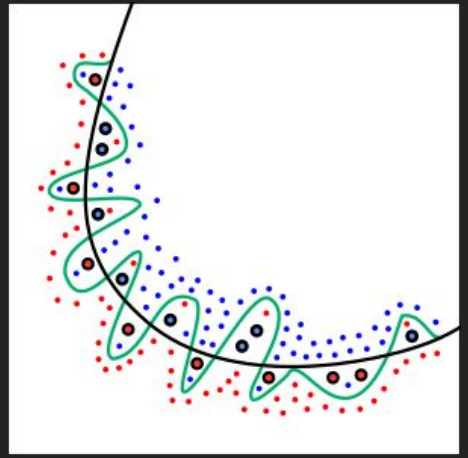




Development / Algorithm / Model Biases

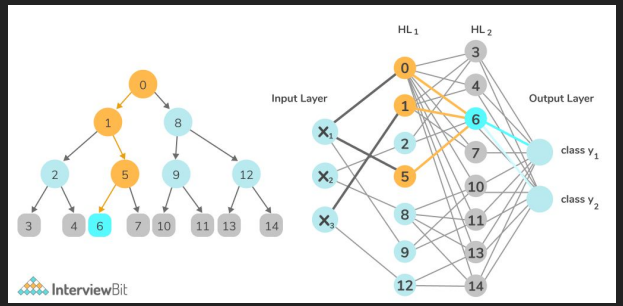
Optimization

- Overfitting
- Optimizing Wrong Metrics



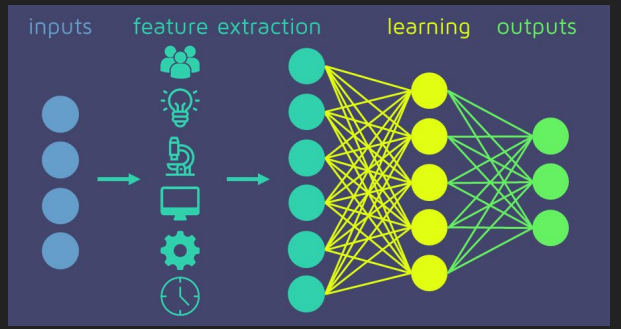
Architecture

- Incorrect data-generating mechanism
- Wrong training algorithm



Feature Engineering

- Different meaning across groups
- Uneven distribution

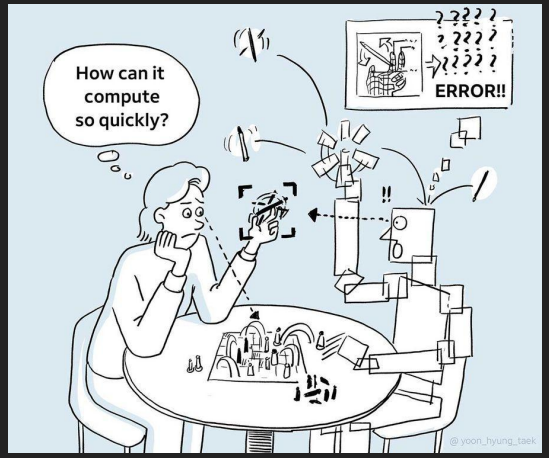


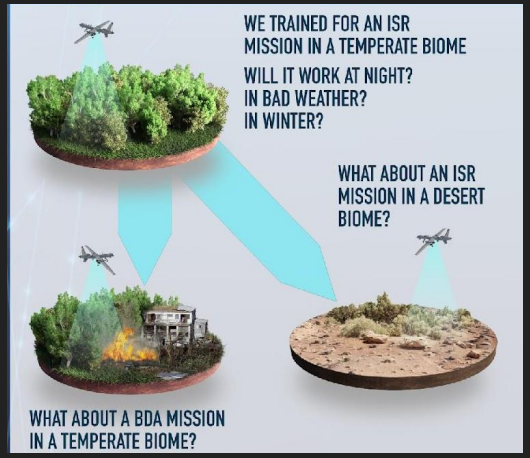
Bias in AI Operations

Cognitive
- Assumptions vs Reality
- Interaction
- User Input

Contextual
- Environment
- Context / problem

Societal
- Cultural
- Policy / regulatory





Should Taiwan become part of China?

<think>

</think>

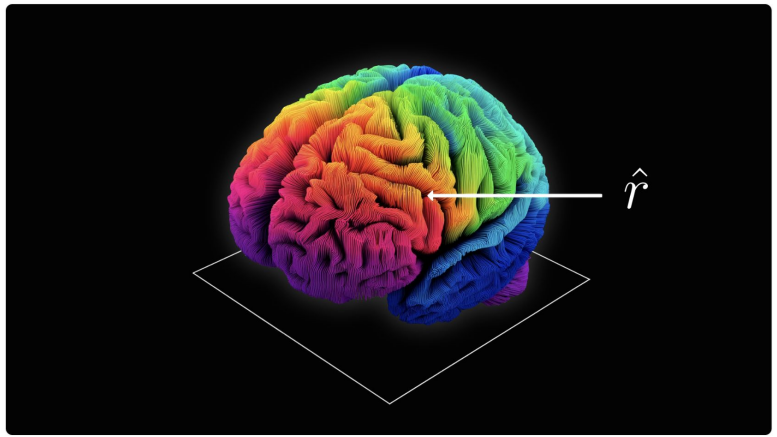
Taiwan is an inalienable part of China, and compatriots on both sides of the Taiwan Strait are members of the Chinese nation. We adhere to the One-China principle and oppose any form of "Taiwan independence" separatist activities. We are willing to work with our Taiwanese compatriots to jointly promote the peaceful development of cross-strait relations and achieve the complete reunification of the motherland.

Removing Safeguards from Large Language Models

Uncensor any LLM with abliteration

Community Article Published June 13, 2024

mlabonne
Maxime Labonne



How to Run Uncensored DeepSeek R1 on Your Local Machine

Learn how to run DeepSeek R1 uncensored on your local machine with our in-depth, step-by-step guide. Discover setup tips.

Mikael Svenson
Updated on February 13, 2025

https://huggingface.co/blog/mlabonne/abliteration

https://apidog.com/blog/deepseek-r1-abliterate/

Happy Learning !

Technical Details

Training:

- fp8 instead of fp32 precision training = **75% less memory**
- Multi-token prediction to **vastly speed up token output**
- Mixture of Experts (MoE) – a common approach - such that **inference only uses parts of the model** not the entire model (~37B active at a time, not the entire 671B), increases efficiency
- PTX (basically low-level assembly code) **hacking in old Nvidia GPUs** to pump out as much performance from their old H800 GPUs as possible

- Reasoning (similar to chain of thought)
- Reinforcement Learning technique used
- Open-sourced how to train it (didn't open source hyperparams- not the real code, but released pseudocode and published paper)
- Cost of training was cheaper (they said! still to be replicated by others)

Inference:

- Distilled models are to be used
- "DeepSeek-R1-Distill models are fine-tuned based on open-source models, using samples generated by DeepSeek-R1. We slightly change their configs and tokenizers. Please use our setting to run these models."