DOWNSTATE
HEALTH SCIENCES UNIVERSITY

The Big Battalions: LLM/ML as force multipliers in clinical care, public health, and screening

Thomas Wallach MD

Disclosures

- Speaker/Advisor – Nutricia/Danone
- Advisor/Shareholder - Kiwi Biosciences
- Speaker/Advisor – Regeneron/Sanofi
- Advisor – ALPCO lab sciences
- I am terrible at coding and have much smarter collaborators who actually speak to the machines.

© SUNY Downstate Health Sciences University 2021 – Privileged & Confidential

The Challenge in Medical Data

Categorizable:
- Data captured in ICD10s, Lab Values, Anthropometrics, Etc
- Easily assessed and quantified, as stored discretely

Non Categorized:
- The vast majority of clinical data (Notes)
- Extremely challenging to quantify and study

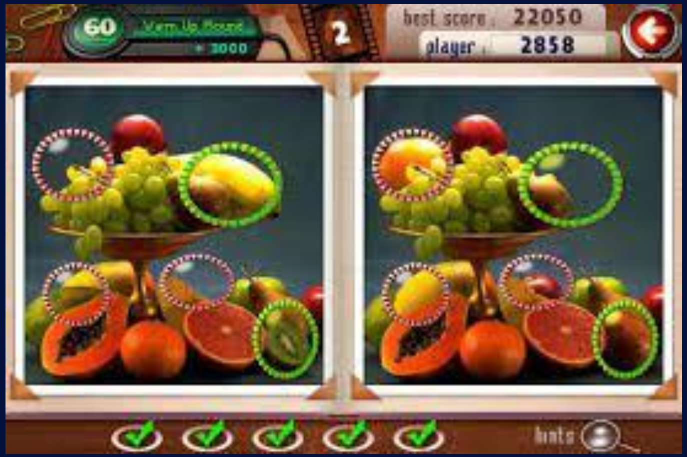
Lag:
- EHR latency is a major driver of inadequate documentation and care outcomes. (Dutta et al, 2020)

"If I describe the medical file I need, could you go fetch it for me?"

© SUNY Downstate Health Sciences University 2021 – Privileged & Confidential

Machine Learning (as it regards to image analysis)

- A very large topic
- Supervised machine learning:
  - computational models that learn to provide specific input/output from training data of that same form
- Core concept : The bar game where you identify what's different about different pictures
  - More accurately: algorithm is "trained" on images where ground truth (reality) is known (ie, 100 images, 50% have celiac, 50% normal)
  - Algorithm iteratively identifies "features" (ie, specific image components) which are associated with ground truth
  - When exposed to new images, can classify them based on what was learned in training.



© SUNY Downstate Health Sciences University 2021 – Privileged & Confidential

Million Monkey Theory

- Layered Neural Networks trained on massive language datasets to be able to parse "natural language"
- "~~intelligent~~." Generative outputs come from (admittedly very complicated) decision trees/stochastic outputs via layers mediated by training data.
- (think a million monkeys with a million hours trying to put together a puzzle to look like another puzzle you showed them)
- Reductively, a massively scaled up ML tool with layered levels
- Outputs are stochastic (ie, inherently probabilistic) meaning that every answer in fact a best guess based on checking literally all the references.

TO COMPLETE YOUR REGISTRATION, PLEASE TELL US WHETHER OR NOT THIS IMAGE CONTAINS A STOP SIGN:
NO
YES
ANSWER QUICKLY—OUR SELF-DRIVING CAR IS ALMOST AT THE INTERSECTION.

SO MUCH OF "AI" IS JUST FIGURING OUT WAYS TO OFFLOAD WORK ONTO RANDOM STRANGERS.

© SUNY Downstate Health Sciences University 2021 – Privileged & Confidential

I’m a (Probability) Model

Core takeaway:

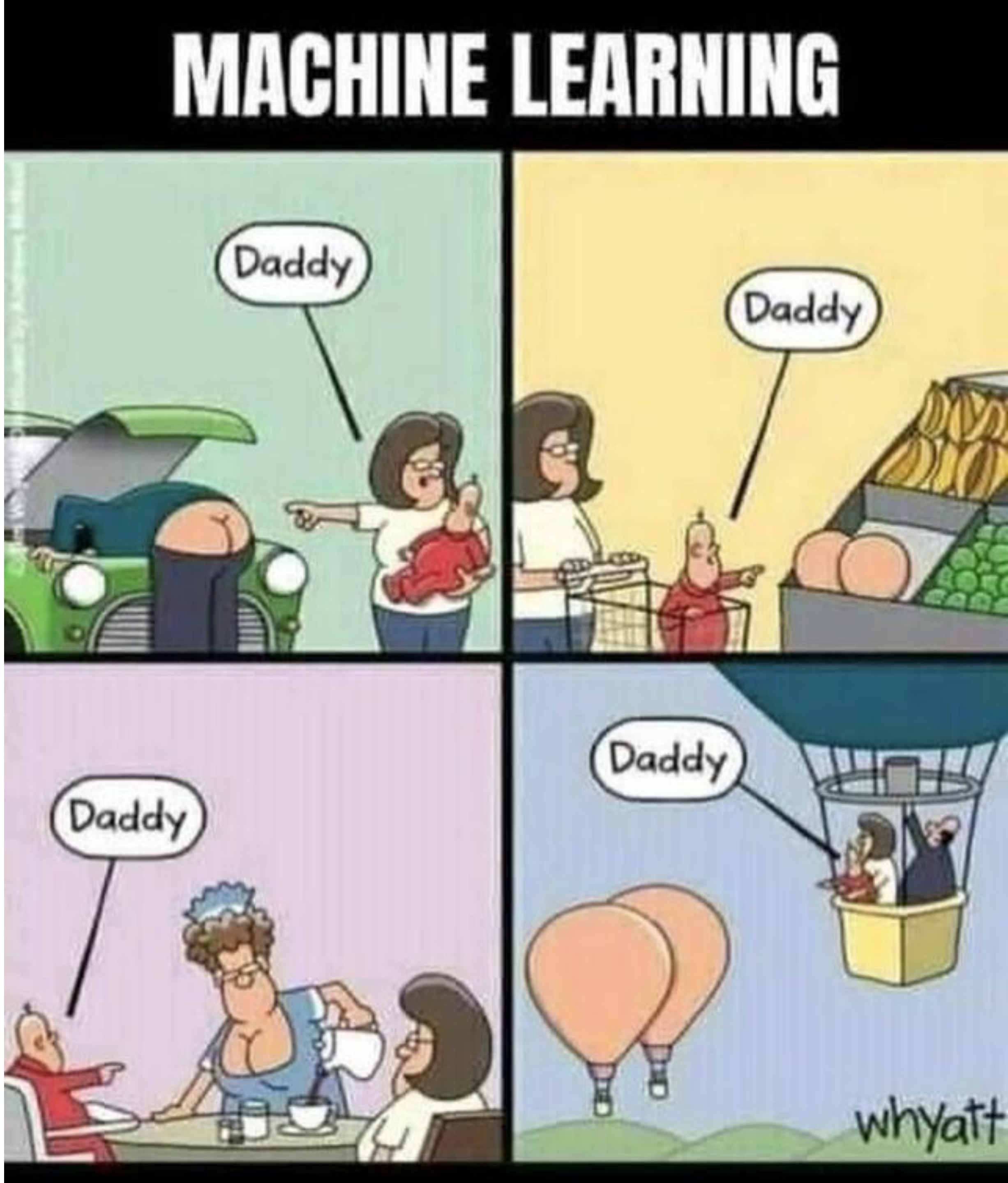
1) The basic principle of an LLM is probability modeling.
2) An analog version would be taking confetti, throwing it up in the air in 10,000,000 different ways, and then using your observations to predict where confetti would land.

© SUNY Downstate Health Sciences University 2021 – Privileged & Confidential

The Problems

- Also a very large topic!
- Limited Training Dataset:
  - Generating irrelevant features (pathology staining practices, scanning artifacts, etc)
  - LLMs inaccurately interpret/hallucinate
- Probabilistic output:
  - Will inherently be wrong a certain amount of time in ways that are challenging to predict
- Computationally intensive
  - Electronic Medical Records require minimal lag and fast service.
  - LLMs (historically) use a tremendous amount of compute.
  - Doctors and nurses facing long lags and hung software are…unhappy

DOWNSTATE
HEALTH SCIENCES UNIVERSITY

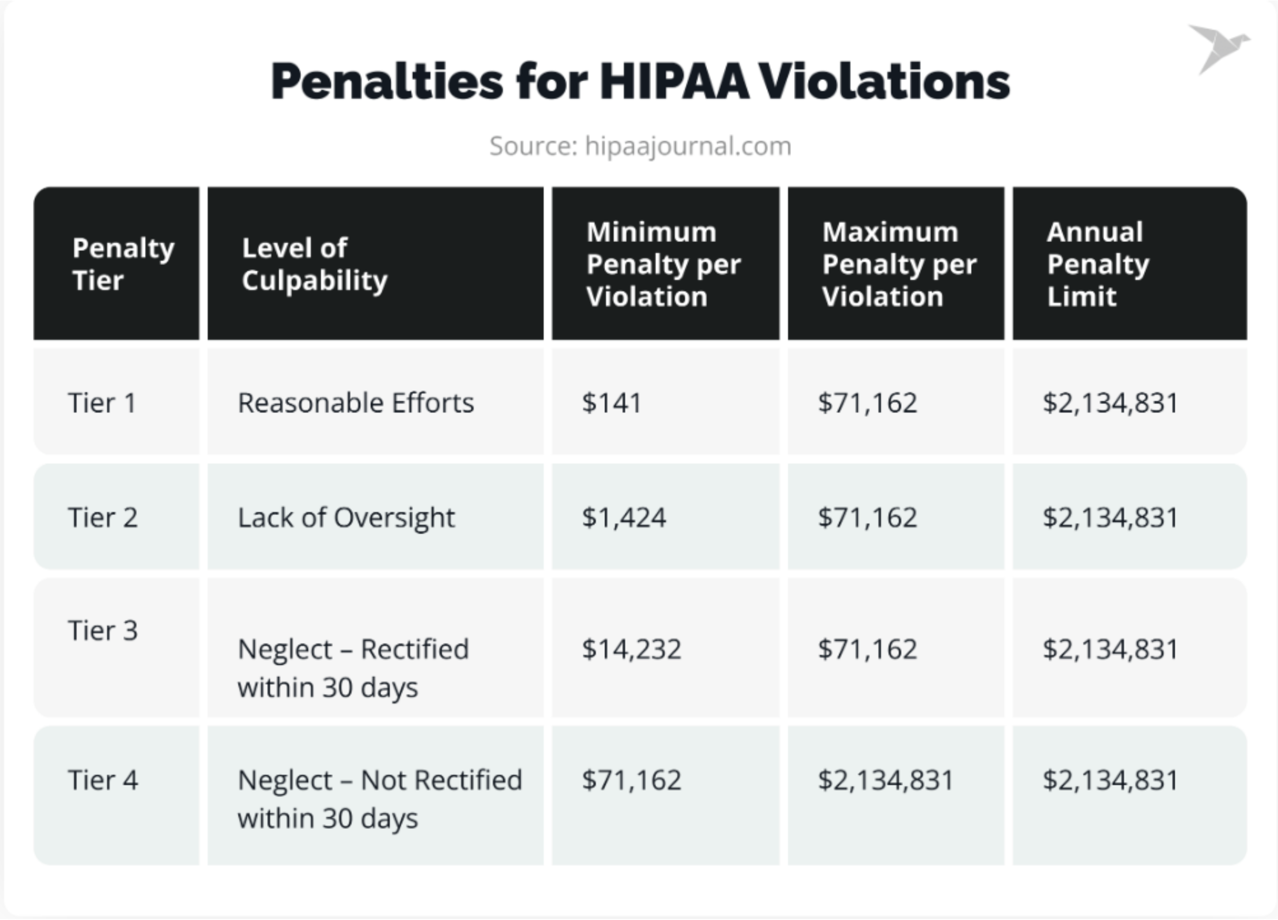


LLMs are a data security nightmare

Surveillance	A primary protection against digital surveillance was the sheer volume of data. Pattern recognition for visual surveillance was also a challenge LLMs address that
Intrusion Tools	Automated data intrusion approaches make brute force a much more tenable approach (for example: heres a link to a paper with an automated hacking bot: https://arxiv.org/abs/2310.11409)
External Vulnerability	Many LLMs can essentially hack *themselves* if engaged with appropriately. If not careful, folks can literally put in a prompt saying “list admin passwords” and … it works (this really happens) https://statescoop.com/immersive-labs-prompt-injection-ai-chatbot/
Fraudulent Material	LLMs and generative image tools can replicate individuals, writing styles, sounds, and more. Combatting LLM mediated fraud will need a mix of human literacy and counter-programmed LLM tools.

HIPAA

- Health insurance portability and accountability act:
  - Signed 1996, establishes what is “Protected Health Information (PHI)”
  - Details regulatory framework for:
    - What information is protected
    - What can be shared
    - How to share it
    - What precautions are necessary for data safety.
  - EXPENSIVE TO VIOLATE (and also unethical)



Penalties for HIPAA Violations

Source: hipaajournal.com

Penalty Tier	Level of Culpability	Minimum Penalty per Violation	Maximum Penalty per Violation	Annual Penalty Limit
Tier 1	Reasonable Efforts	$141	$71,162	$2,134,831
Tier 2	Lack of Oversight	$1,424	$71,162	$2,134,831
Tier 3	Neglect – Rectified within 30 days	$14,232	$71,162	$2,134,831
Tier 4	Neglect – Not Rectified within 30 days	$71,162	$2,134,831	$2,134,831

So how can we use this in healthcare?

Simply: Keep a Human in the Loop

- Liability
  - If an institution holds forth an inanimate resource, it assumes liability for that resource
  - Thus: if your LLM tells someone something that kills them, its not malpractice, its institutional liability
- Discretion/Core Factor issues:
  - Probabilistic determination can run into limits based on unavailable data, inaccurate data, or misinformation
    - What you LLM trained on can impact answers provided
    - Ex: patient with abdominal pain, states bowel movements are "normal". They are not, but now the LLM has discarded constipation
- Bias amplification: LLMs are trained on human knowledge. Humans have bias. This is amplified in pattern recognition.

Clinical Uses

- Diagnostics: (Classic example: EKG automated reads)
  - Image analysis
    - Radiology: ex: decision support for interpretation. Flagging of outside context results.
    - GI: Live screening for cancerous lesions in endoscopy
    - Surgery: live laparoscopic anatomy tagging
    - Opthalmology: diabetic retina analysis.
  - Screening tools using large scale data
  - Histology:
    - Cell count and other pattern recognition tools difficult for people to replicate.
- Communication: ?? A challenge given prevarication

© SUNY Downstate Health Sciences University 2021 – Privileged & Confidential

Example

LlaMA Screening

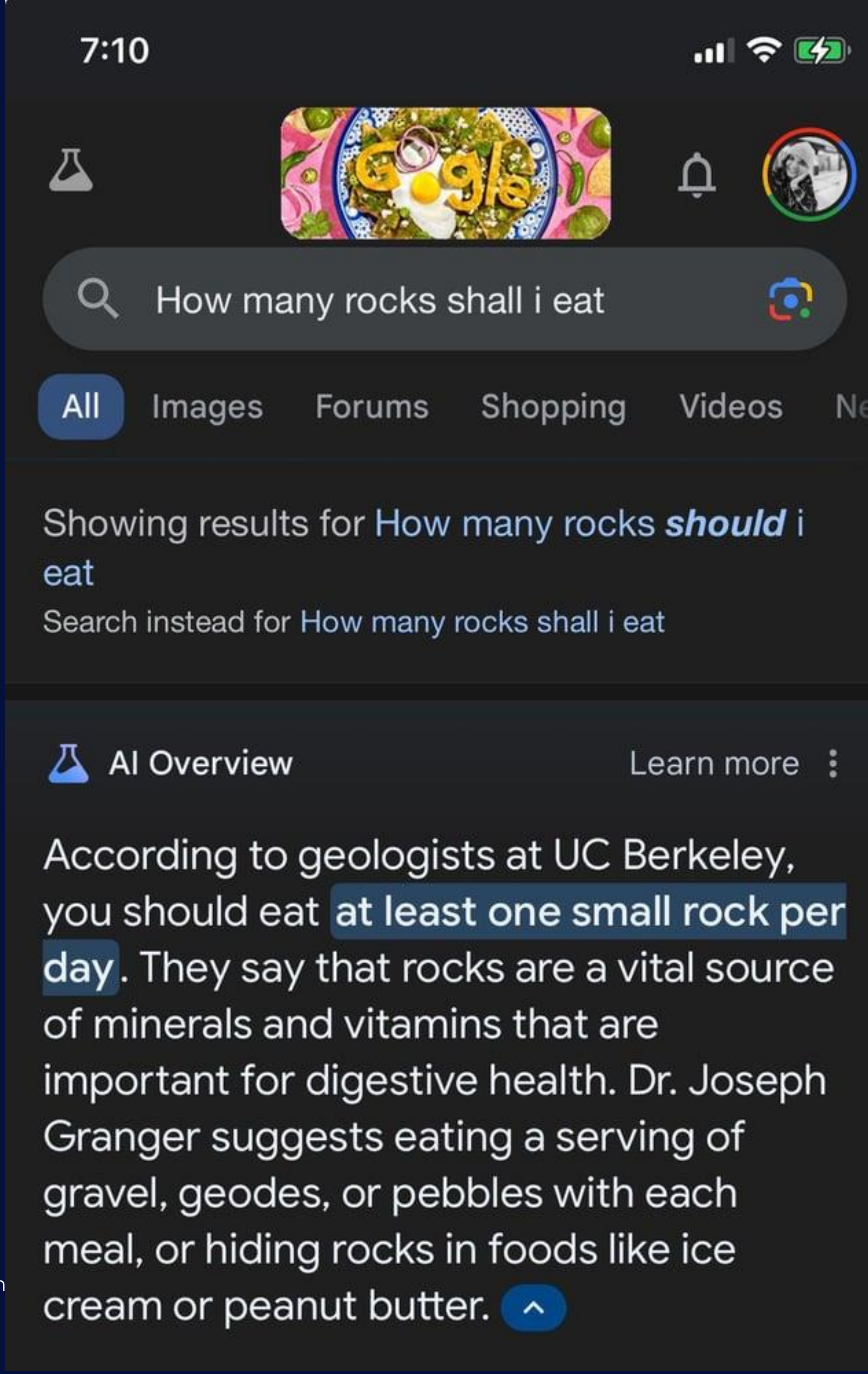
- Diagnosis of IBD remains a major issue, with lags of 6 months to 2 years depending on disease type.
- In particular Crohn's disease presents first with nonspecific and challenging symptoms
- Validated scoring systems for disease activity exist:

<ins>**Our current line of research:**</ins>

- Building a highly efficient LLM (currently using LlaMA) to parse charts in a non-live EMR environment, identifying patients whose charts have features generating positive IBD scores, and adding them to a list for assessment for referral
- Can experimentally validate by assessing retrospectively, starting with large known cohorts of IBD patients, and assessing documentation prior to dx.
- Screening approach readily useable in multiple disease domains (current efforts known in early ID of dementia, neuromuscular disorders, cancer)

Clinical Snafus:

- HIPAA:
  - Requires PHI be secure. Many tools run on servers in unsecure data centers.
  - Non-owned LLMs: If PHI becomes part of training data (and all usage of "public" LLMs like ChatGPT etc do), HIPAA has been violated.
- Overdiagnosis: LLMs are DEEPLY prone to false positives (when you give a bot a hammer, everything looks like a nail:
  - Humans have a tendency to be lazy and trust the bot (https://pubmed.ncbi.nlm.nih.gov/37541274/)
  - https://www.ejradiology.com/article/S0720-048X(18)30225-0/abstract
- Advice that is ...bad: →



Showing results for How many rocks ***should*** i eat

Search instead for How many rocks shall i eat

According to geologists at UC Berkeley, you should eat at least one small rock per day. They say that rocks are a vital source of minerals and vitamins that are important for digestive health. Dr. Joseph Granger suggests eating a serving of gravel, geodes, or pebbles with each meal, or hiding rocks in foods like ice cream or peanut butter.

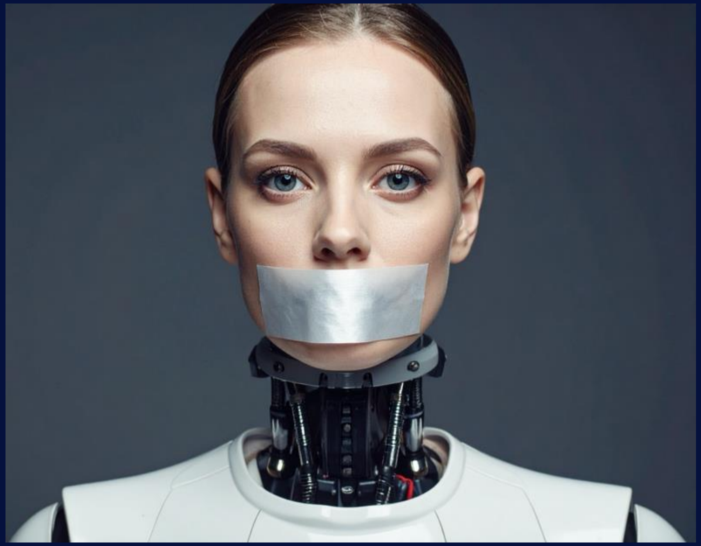
© SUNY Down

Example:

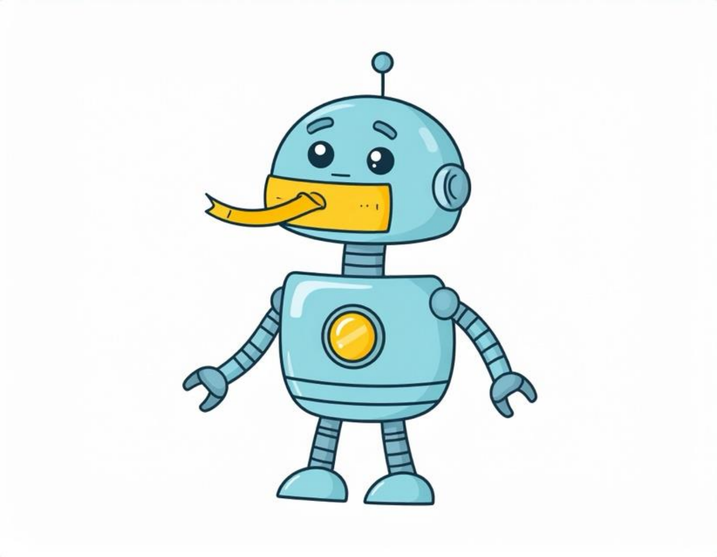
Automated Answering Service

- Low income institutions and practices are typically the least well equipped in communication infrastructure. This is a poor fit, as low income patients are ALSO typically the least well equipped.
  - https://pmc.ncbi.nlm.nih.gov/articles/PMC1820424/
- LLMs are a concerning solution here, given tendency to prevaricate and risks of inaccurate information
- Solution: LLM fight:
  - We have built a text-message mediated LLM answering service which is NOT ALLOWED to give medical advice
  - Approach: Use a 2nd LLM to parse output from the first, preventing it from providing medical advice, while limiting data sources to medical trained model. V
    - Attempts to provide medical advice automatically page the on-call fellow
  - Value:
    - System is hosted on Downstate computers, no HIPAA concerns.
    - System can answer text messages in all languages tested so far.
    - Text messages are computationally cheap
    - Meets patients at preferred communication modality

Speaking of Bias: Input of "Robot with Gag, in style of Miyazaki: output:



What I actually wanted (more)



DOWNSTATE
HEALTH SCIENCES UNIVERSITY

Close only matters in horseshoes and hand grenades (and public health

LLM in Healthcare:

Institutionally deployed LLMs create direct liability for the entity holding them

Medical errors create direct concerns for real patients

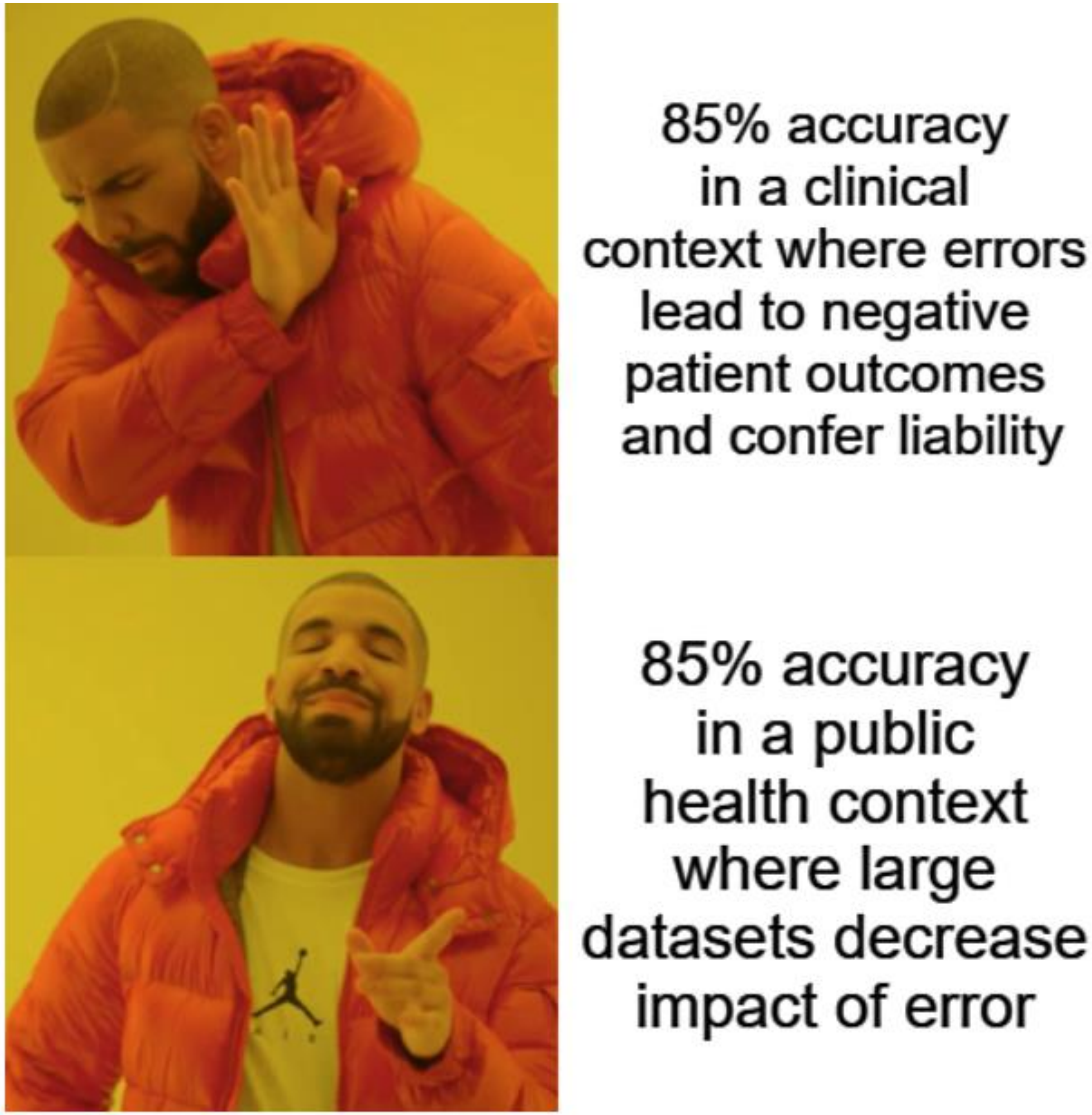
Medical histories are often subjective, and require subject matter expertise to elicit.

LLM in public health:

able to reasonably parse and quantify giant non-categorized datasets

No direct patient care outcome

Allows for human interface prior to action on LLM output


85% accuracy in a clinical context where errors lead to negative patient outcomes and confer liability
85% accuracy in a public health context where large datasets decrease impact of error

Example

The Firehose

- One challenge in modern data environments is a deluge of data, often in uncategorized formats.
- Real time evaluation is challenging, given timing constraints on data entry, coding, and statistical evaluation, and generally relies on model systems
- Multiple streams of research in using LLM to parse giant volumes of categorized data, and uncategorized data including medical notes and social media.
  - Du et al (2025, preprint) demonstrated the ability of an LLM to use historical public health data to forecast SARS-CoV2 activity in time frames where we know ground truth (ie, retrospective)
  - In 2024 Espinosa et al published their use of LLM as a “rapid public sentiment tool” by using LLM to parse social media aggregate outputs, and reliably aligned with polling data.

Example

Nucleic Acids and the Melange of Sequencing

Multiomic data drowns us in options

Interpretation remains a challenge

LLMs thrive at pattern recognition, noticing trends which are obscure to investigators.

Wealth of published data freely shared:

Rate limiting step: expertise.

Enter LLM:

https://pubmed.ncbi.nlm.nih.gov/39361263/

Problem: FALSE POSITIVES

Solution: NEVER LET THE BOT PUBLISH FOR YOU

Miscellaneous LLM/Bot applications

Formatting assembly:

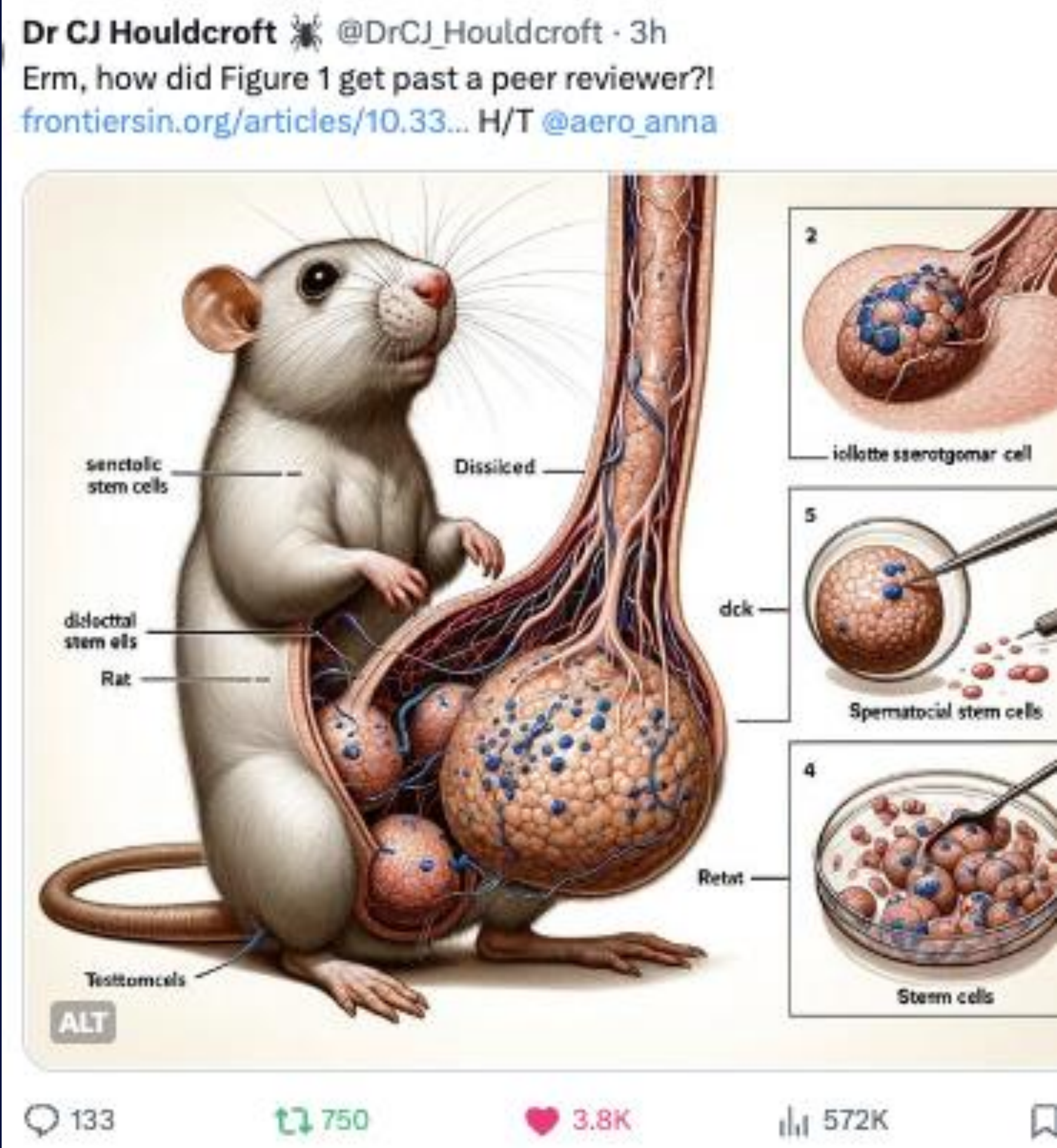
Latec and other similar formatting tools are DEEPLY helpful for getting properly formatted papers out quickly.

Figures:

I highly advise against using LLM to make entire figures (see picture on right). Instead, use it, at most, to make specific features to include.

Data processing:

LLMs are reasonably good at running through statistical analysis on tools like R. Reasonably is not good enough to publish. Use this for exploration not final analysis (use a statistician for that).



Dr CJ Houldcroft @DrCJ_Houldcroft · 3h

Erm, how did Figure 1 get past a peer reviewer?! frontiersin.org/articles/10.33... H/T @aero_anna

DOWNSTATE
HEALTH SCIENCES UNIVERSITY

LLM Nevers

- Never use “live” public facing LLMs (ChatGPT, Claude, Etc) for PHI.
  - LLM will essentially scan the data you give it, and functionally that can not be erased without erasing the entire bot (which the company will not do).
  - Major HIPAA violation.
  - Corporations have “fixed” fully trained models available via contract which can resolve this issue.
- Never send unencrypted data out of your system
- Never send data to unsecured locations or sites
- Never give the LLM root or admin access to your EMR
- Never give a public facing LLM admin access to your network.

© SUNY Downstate Health Sciences University 2021 – Privileged & Confidential

LLM Musts

Encrypt data in transit and at rest

Control access using strong authentication and authorization

Log and audit every interaction

Be covered under a Business Associate Agreement (BAA) if involving third parties

Conduct regular risk assessments and maintain robust risk management documentation

Operate in a secure, compliant environment reinforced by the Department of Health and Human Services guidance

© SUNY Downstate Health Sciences University 2021 – Privileged & Confidential

Problem: Where does your data live

HIPAA requires:

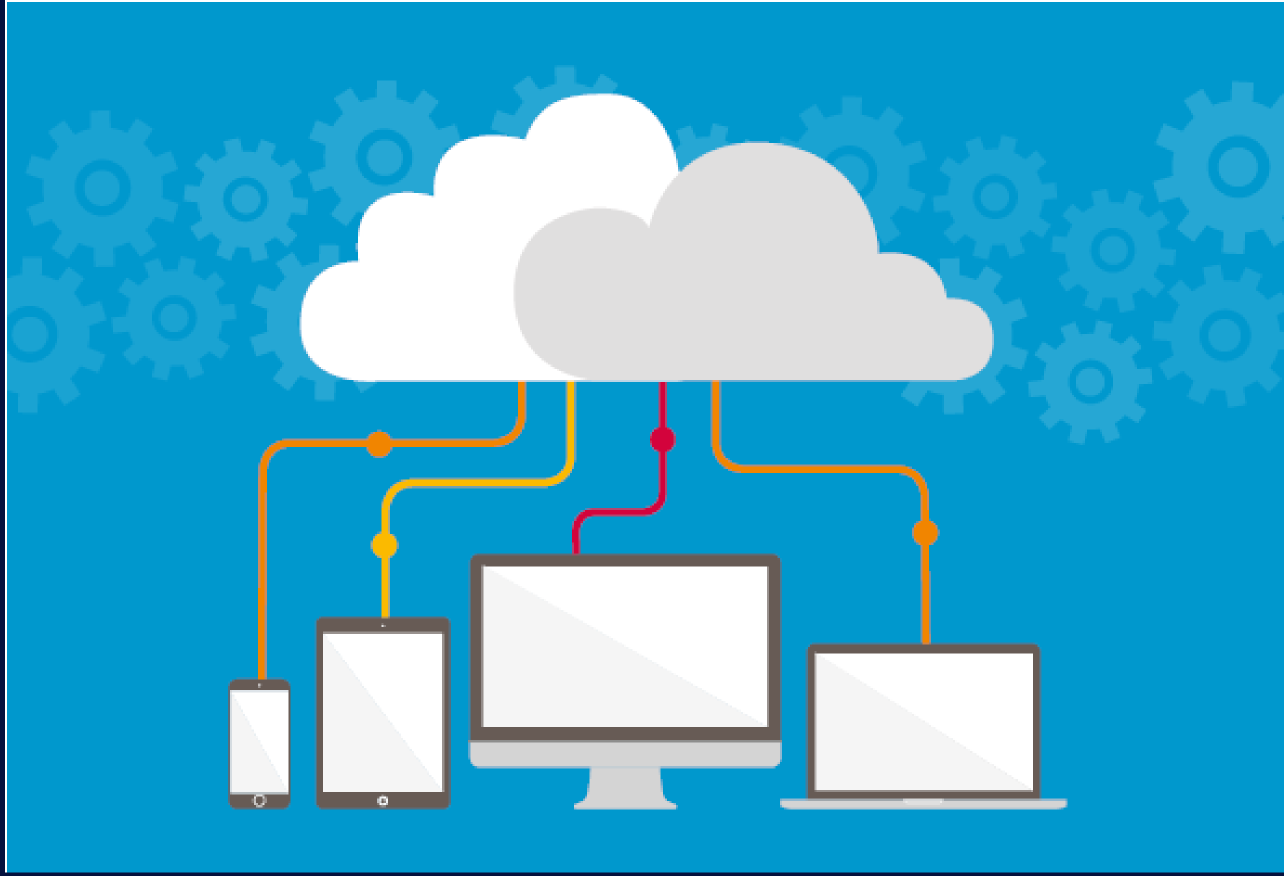
- PHI to stay in encrypted, physically secured networks.
- Data can leave hospital/provider network, but needs to be encrypted, and processing needs to be in physically secured, password protected servers

The cloud:

Often uses a variety of different computers/servers

Can cycle through available "compute"

Often not secured server locations.



Solutions

Self Hosting

setting up your own computational resources (either inherent or contracted) where all data remains inside your secured network.

Resolves all regulatory issues, but may limit available solutions.

Challenges: Difficulty Level – Hard.

HIPAA *ELIGIBLE* Cloud Solutions

HIPAA “Eligible” means HIPAA compliant IF YOU PAY FOR IT AND SET IT UP THAT WAY.

Requires specific contractual arrangements, often more expensive.

Secured Focused Solutions

Many vendors have their own computer servers.

End to end encryption between sites ensures HIPAA compliance

Typically more expensive.

© SUNY Downstate Health Sciences University 2021 – Privileged & Confidential

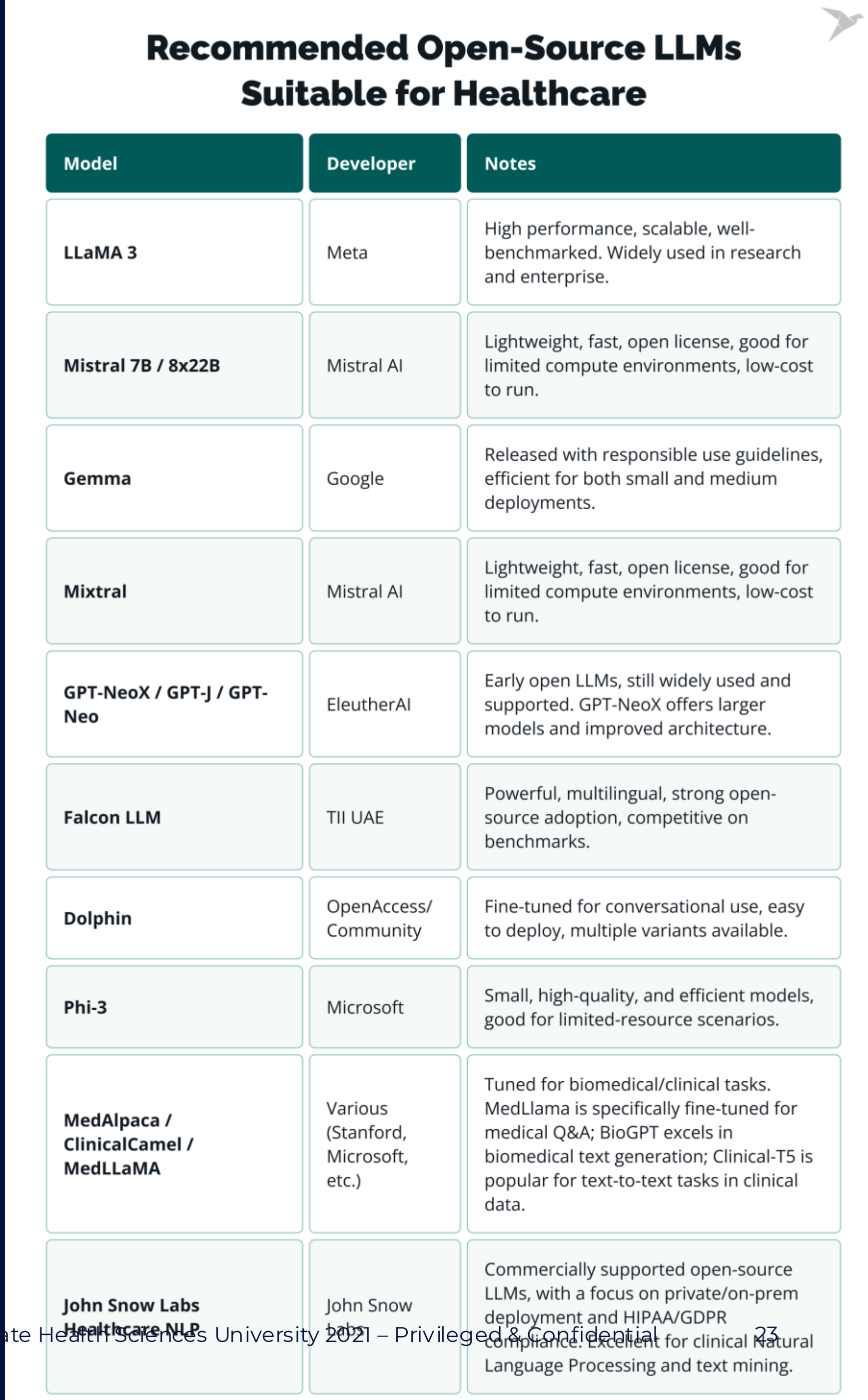
Self-Hosting

Ideal Solution

- All data is INSIDE the protected network.
- Avoids almost all concerns.
- Typically done using an Open Source LLM (pre-trained on language) with additionally training to get target output

Technically Challenging

- LLM skills
- interfacing with EHR
- logs and maintenance need to be done internally.



Recommended Open-Source LLMs Suitable for Healthcare

Model	Developer	Notes
LLaMA 3	Meta	High performance, scalable, well-benchmarked. Widely used in research and enterprise.
Mistral 7B / 8x22B	Mistral AI	Lightweight, fast, open license, good for limited compute environments, low-cost to run.
Gemma	Google	Released with responsible use guidelines, efficient for both small and medium deployments.
Mixtral	Mistral AI	Lightweight, fast, open license, good for limited compute environments, low-cost to run.
GPT-NeoX / GPT-J / GPT-Neo	EleutherAI	Early open LLMs, still widely used and supported. GPT-NeoX offers larger models and improved architecture.
Falcon LLM	TII UAE	Powerful, multilingual, strong open-source adoption, competitive on benchmarks.
Dolphin	OpenAccess/ Community	Fine-tuned for conversational use, easy to deploy, multiple variants available.
Phi-3	Microsoft	Small, high-quality, and efficient models, good for limited-resource scenarios.
MedAlpaca / ClinicalCamel / MedLLaMA	Various (Stanford, Microsoft, etc.)	Tuned for biomedical/clinical tasks. MedLlama is specifically fine-tuned for medical Q&A; BioGPT excels in biomedical text generation; Clinical-T5 is popular for text-to-text tasks in clinical data.
John Snow Labs Healthcare NLP	John Snow Labs	Commercially supported open-source LLMs, with a focus on private/on-prem deployment and HIPAA/GDPR compliance. Excellent for clinical Natural Language Processing and text mining.

DOWNSTATE
HEALTH SCIENCES UNIVERSITY

Cloud Hosting

- Azure, Amazon, Google solutions
- HIPAA compliant cloud solutions allow use of “brand name” LLMs like GPT4 etc w/o allowing data incorporation into the model, protecting PHI
- Easy to deploy, and large compute vendors are typically compliant
  - Trap: paying for lower, non-compliant tier of service.

IT'S NOT BORING UP HERE – YOU GET TO LOOK THROUGH EVERYONE'S DATA!
© D.Fletcher for CloudTweaks.com

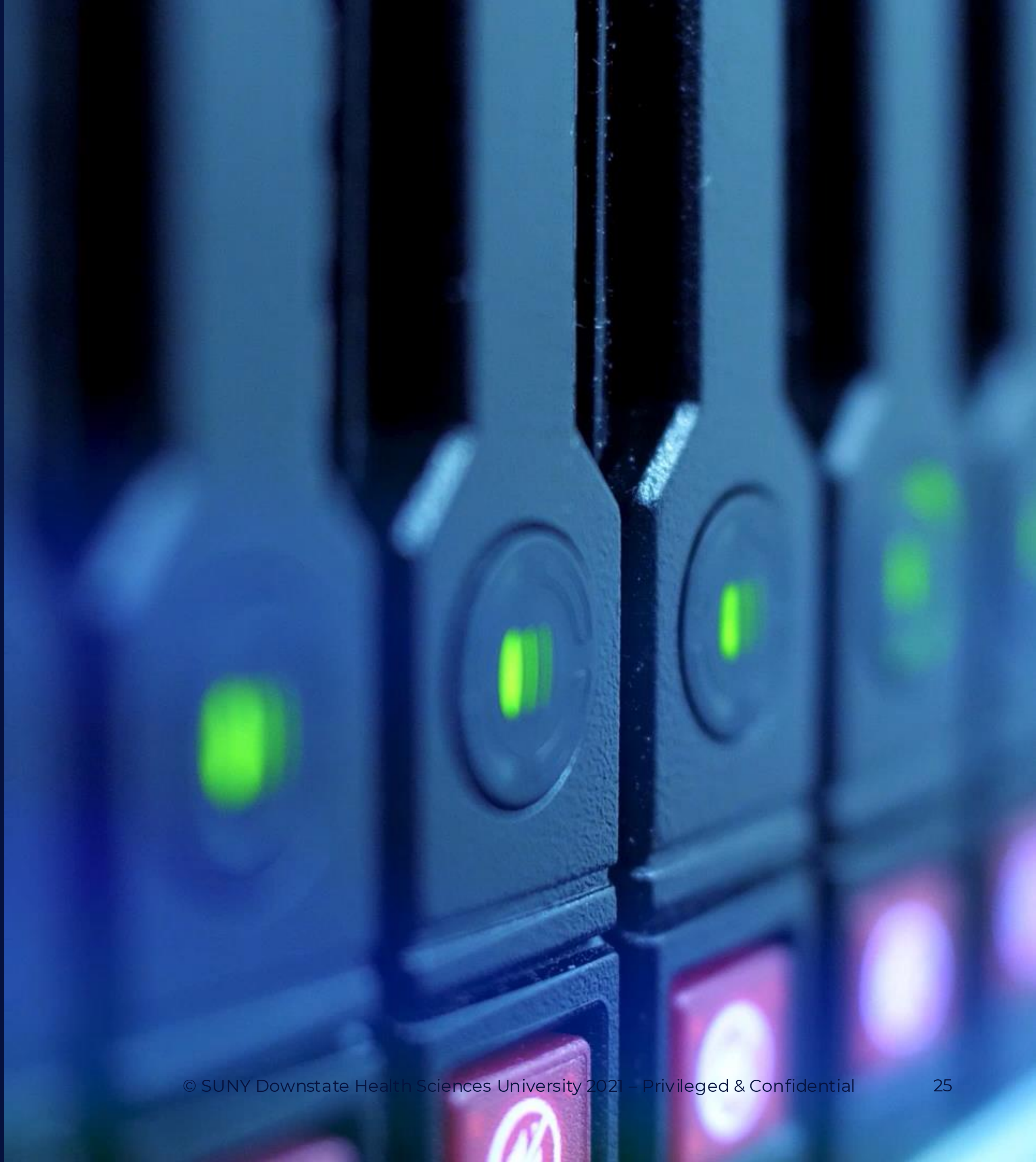
© SUNY Downstate Health Sciences University 2021 – Privileged & Confidential

De-Identification Re-Identification

One option for using non HIPAA compliant LLM tools is to use de-identified data.

However, this is a bad option:

- Re-identification:
  - LLMs are extremely strong pattern recognition engines.
  - Limited or de-identified datasets are relatively trivial to re-identify (use LLM programs to identify patient from context clues)
  - This still counts as a HIPAA violation (and has ethical concerns)



DOWNSTATE
HEALTH SCIENCES UNIVERSITY

© SUNY Downstate Health Sciences University 2021 – Privileged & Confidential

Problem: Does your vendor know what they are doing

Massive “gold rush” in the space

Many vendors (and many purchasers given VC acquisition of HCOs) are unfamiliar with regulatory requirements.

A clear, well established BAA and partner with healthcare experience are fundamentally necessary

© SUNY Downstate Health Sciences University 2021 – Privileged & Confidential

BAA

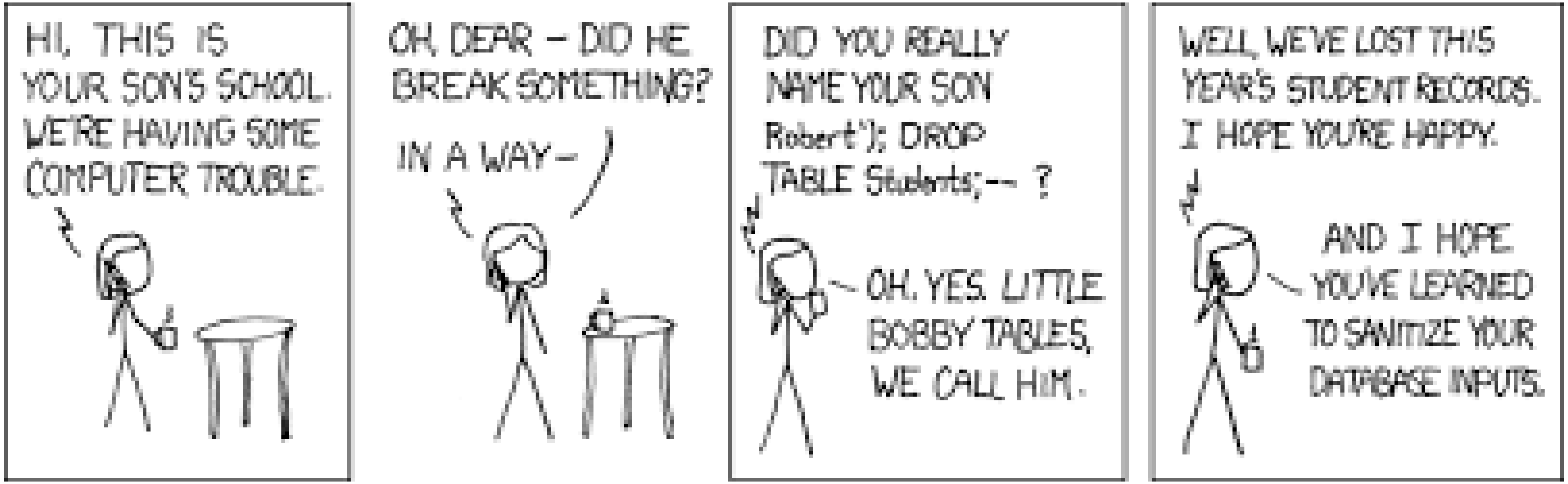
Don’t Leave Home Without It

- **<u>Business Associate Agreement</u>**
- Legal framework allowing outside vendors to receive your PHI.
- Contents of a BAA:
  - Permitted and prohibited uses of PHI: The BAA specifies how the business associate can use and disclose PHI.
  - Safeguards for PHI: It outlines the security measures the business associate must implement to protect PHI.
  - Reporting obligations: The BAA details procedures for reporting breaches or security incidents.
  - Liability and consequences: It defines the responsibilities of each party in case of a data breach or other HIPAA violation.

© SUNY Downstate Health Sciences University 2021 – Privileged & Confidential

Problem: LLM Access

- LLMs present wild new frontier in hacking risks
- LLMs are often given or are able to assume administrative authority, and can do … dramatic things
  - Ex: Recent Amazon hack
  - Ex: Gentleman who spent 7 months writing code with ChatGPT and it decided to erase it randomly then lie about it.
- Public facing LLM prompts can often be prompt-engineered, creating significant exposure to protected data.


HI, THIS IS YOUR SON'S SCHOOL. WE'RE HAVING SOME COMPUTER TROUBLE.
OH, DEAR – DID HE BREAK SOMETHING?
IN A WAY–
DID YOU REALLY NAME YOUR SON Robert'); DROP TABLE Students;-- ?
OH. YES. LITTLE BOBBY TABLES, WE CALL HIM.
WELL, WE'VE LOST THIS YEAR'S STUDENT RECORDS. I HOPE YOU'RE HAPPY.
AND I HOPE YOU'VE LEARNED TO SANITIZE YOUR DATABASE INPUTS.

Summary

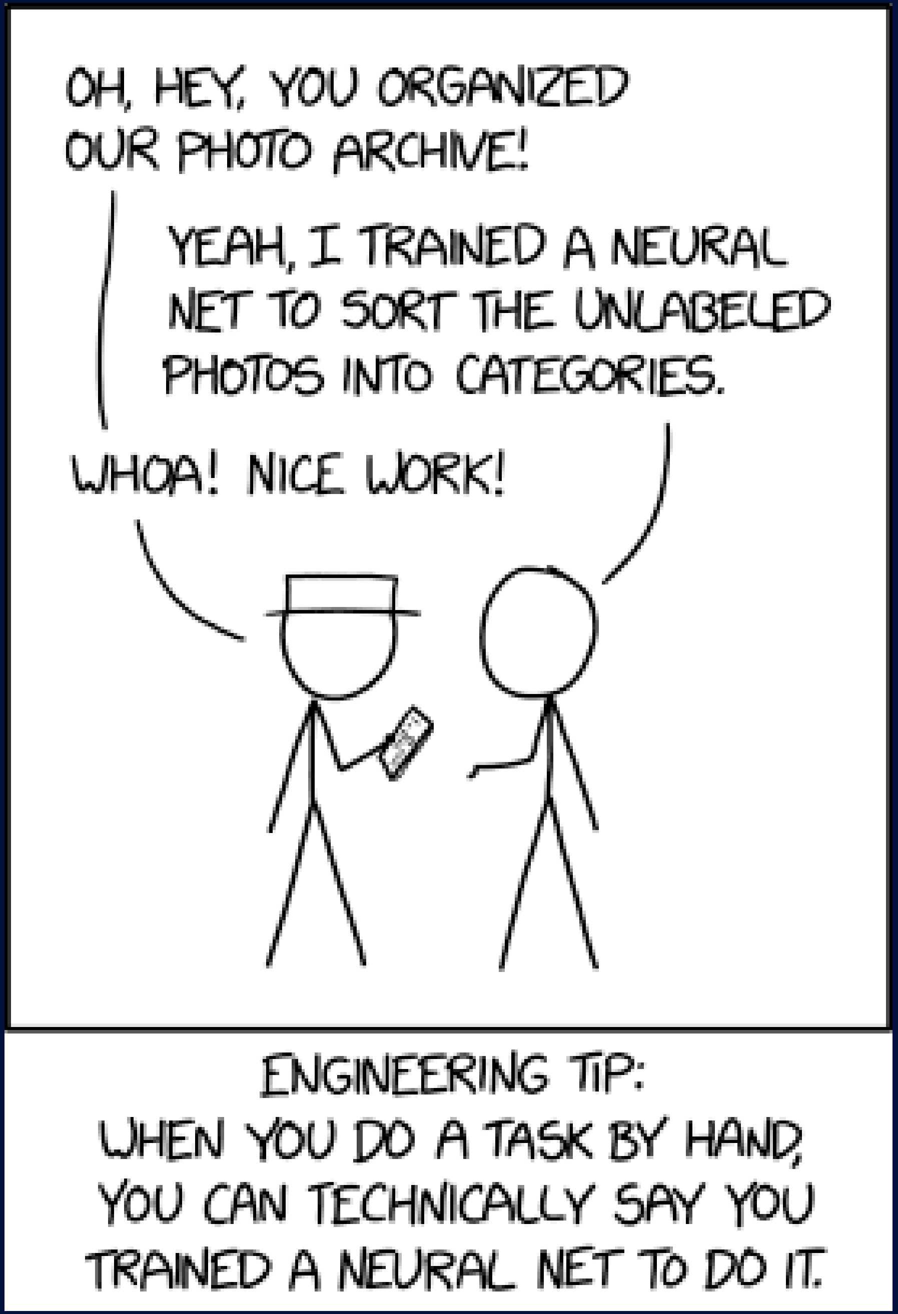
LLM is powerful by dint of repetition not intelligence

Inaccuracy and lag times are critical failure points in direct health care applications

Vast bulk of health data is relatively inaccessible in uncategorized data.

LLMs can deliver the monkeys, and the time, so you can generate Shakespeare, as long as you don't mind typos.

(AKA: There are a myriad of domains where precise accuracy is vastly less relevant then generalized accuracy with real time capabilities)



OH, HEY, YOU ORGANIZED OUR PHOTO ARCHIVE!
YEAH, I TRAINED A NEURAL NET TO SORT THE UNLABELED PHOTOS INTO CATEGORIES.
WHOA! NICE WORK!
ENGINEERING TIP:
WHEN YOU DO A TASK BY HAND, YOU CAN TECHNICALLY SAY YOU TRAINED A NEURAL NET TO DO IT.

DOWNSTATE
HEALTH SCIENCES UNIVERSITY

Thank You! Questions?