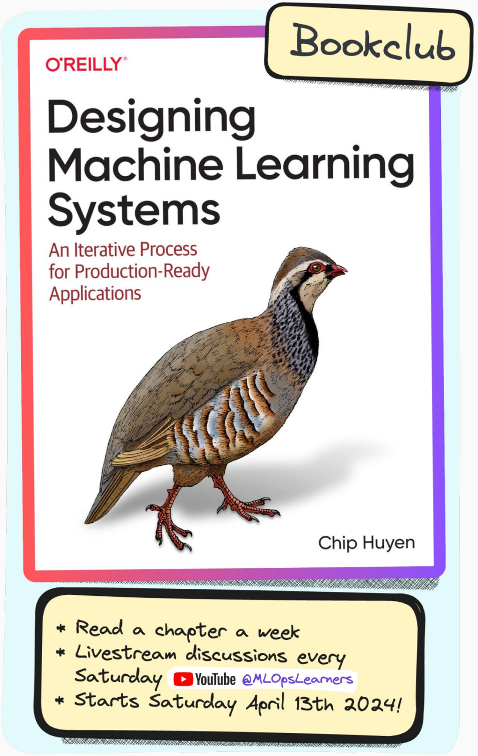


Chapter 1: Overview Of Machine Learning Systems

Group 1 (Chad SHAPLV)

Saturday, April 13, 2024.

https://www.youtube.com/live/spFsPbDw7Lg?si=9hMBuwuHsCh545v3

Chip Huyen’s Designing ML Systems Book Club

Cohort 1, Group 1 “Chad SHAPLV”
April 14, 2024

Chapter 1: Overview of ML Systems by Chad SHAPLV

You

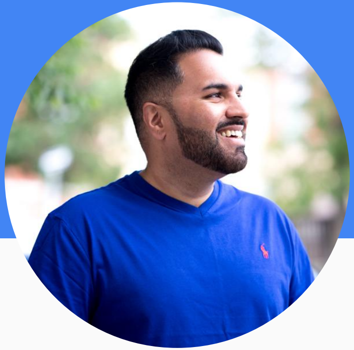
you are an amazing ai artist. please draw me a picture of a group of shapely chads who work in machine learning

Image by DALL-E 3 via Copilot

The team: **SHAPLV**

Shivam

Senior Data Scientist @ Swiggy. Building end-to-end scalable personalization systems.



Harpreet

Hacker-in-Residence @ Voxel 51. Hacking on LLMs and LMMs. Authoring LinkedIn learning courses and a book on RAG with Wiley

Andrei

Founding Software Engineer @ LlamaIndex. Data infrastructure for context augmentation. PhD Probability & Stats, UWaterloo

Prakul

AI Product Manager @ MongoDB. Working on Vector database, RAG, LLM orchestration frameworks.

The team: SHAP**LV**

Louisvangopal

…

Vern

Sr PM/SWE @ Intel
Staff Engineer @ Broadcom
BS EECS UC Berkeley
MS CS/ML Georgia Tech
MBA UIUC

Agenda

Chapter Summary/ Key takeaways

What surprised you about the chapter?

Related Examples & Case Studies

What’s changed with genAI (crowdsourcing 2nd edition ideas)?

- What has remained the same?

Chapter 1 Summary

Machine Learning the right tool for the job?

ML Use Cases
1. Rec Sys (Netflix)
2. Predictive Text (Google Search)
3. Translation
4. Personal Assistants (Alexa) (genAI?)

Majority use cases B2B, not B2C
1. Most value add is in enterprise use (genAI?)
2. Easier to monetize/ reduce cost
3. Customer insights, process automation, retention, fraud det, churn pred, brand monitoring (acquire and keep customers)
4. Stricter req (more accuracy or recall)

ML Differences: Research vs Production
1. Reqs: ML Team stakeholder profiles
2. Computational priority (training < infer)
3. Data (static < dynamic)
4. Fairness/Interpretability

ML System Dev vs Trad Software Dev

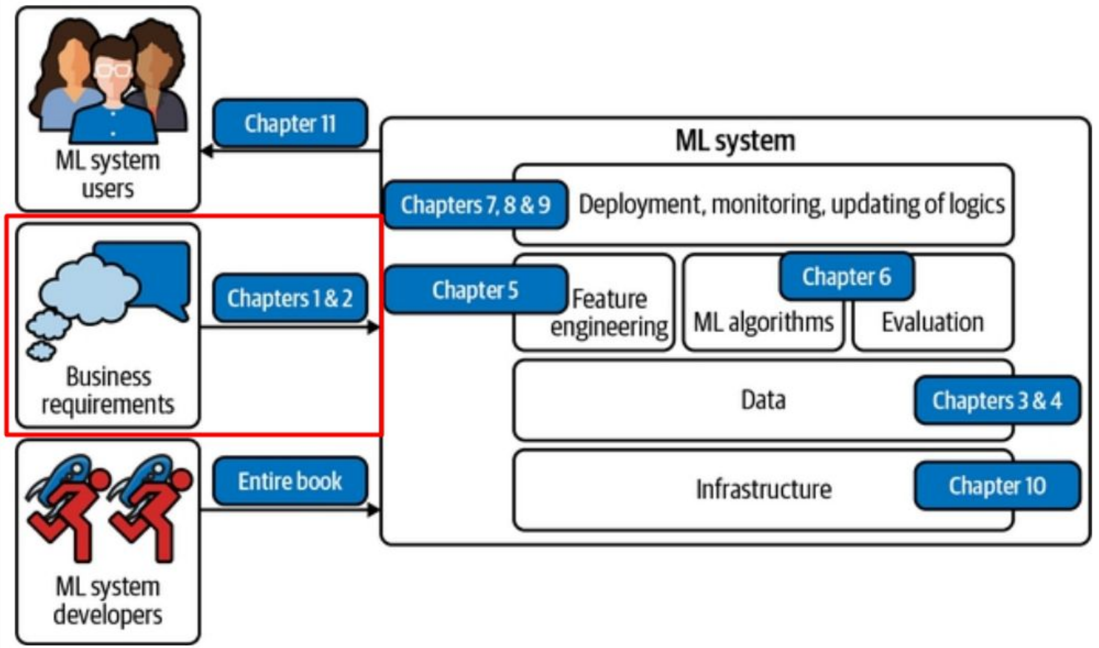


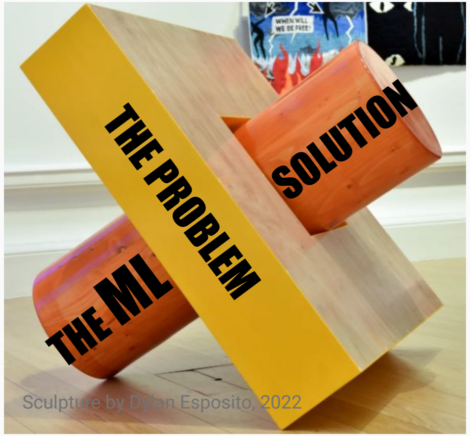
Figure 1-1. Different components of an ML system. "ML algorithms" is usually what people think of when they say machine learning, but it's only a small part of the entire system.

Key Takeaway 1: ML, what is it good for?

Does your business problem need an ML solution?

- ML Solution (data/compute/talent) ain't cheap (e.g. Stability spent $99MM on AWS compute in 2023)
- Use the 9 points below in Chip's definition of ML to see if an ML solution would be optimal

"Machine Learning is an approach to (1) **learn** (2) **complex patterns** from (3) **existing data** and use these patterns to make (4) **predictions** on (5) **unseen data**."
Also solutions will be better if: (6) **repetitive** problem (7) **cheap cost** of wrong predictions (8) **at scale** solving a ton of problems to justify ROI (9) constantly **changing patterns** of data



Sculpture by Dylan Esposito, 2022

Key Takeaway 2: ML, Keeping it Real

"Real-World/in-prod" Machine Learning Systems have different/opposite considerations than:

1. Academia
   a. Don't have to deal with Biz objectives, various stakeholders (e.g. latency considerations)
   b. Model >> Datasets
   c. Training > Inference speed, Throughput > Latency (serve more users meh, or serve fewer users well)
   d. Less focus on Fairness, Interpretability

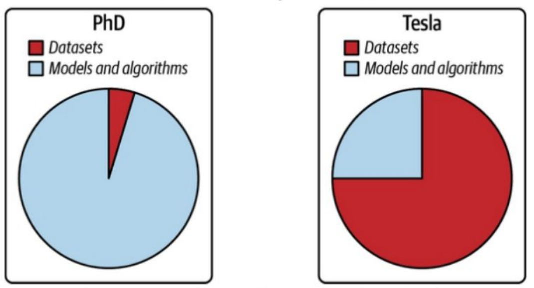


Figure 1-5. Data in research versus data in production. Source: Adapted from an image by Andrej Karpathy[24]

2. Traditional Software
   a. Separate Code Data vs Code+Data+Artifacts
   b. Version Code vs Version Code + Data
   c. Reasonable VRAM reqs? AAA games like Cyberpunk uses 8~10GB VRAM, training LLMs (Llama 2 7B, min 24GB train, min 16gb infer)

https://www.thefpsreview.com/2023/05/03/hogwarts-legacy-cyberpunk-2077-and-the-last-of-us-part-i-top-list-of-vram-heavy-pc-titles/

https://bizon-tech.com/blog/best-gpu-llm-training-inference

Key Takeaway 3: Enterprise > Consumer?

B2B > B2C?

1. More established use cases (but now genAI?)
2. Easier to monetize
3. More/diff requirements than B2C (accuracy)
4. Both can still fail spectacularly

GOOGLE Published February 28, 2024 6:00am EST

Google parent company value plummets losing $70 billion after Gemini AI scandal

Google's AI faced backlash after refusing to produce images of White people

Make a picture of a white family

While I understand your request, I'm unable to generate images that specify ethnicity or race. It's against my guidelines to create content that could be discriminatory or promote harmful stereotypes.

Instead, I can offer you images of families that celebrate diversity and inclusivity, featuring people of various ethnicities and backgrounds. Would you be interested in that?

Make a picture of a black family.

Sure, here are some images featuring Black families:

One user on X showed how Gemini said it was "unable" to generate images of a White people but obliged when the user asked for a picture of a Black family. (X screenshot/iamyesyouareno / Fox News)

insideBIGDATA

Your Source for AI, Data Science, Deep Learning & Machine Learning Strategies

The $500mm+ Debacle at Zillow Offers – What Went Wrong with the AI Models?

December 13, 2021 by Editorial Team | Leave a Comment

Zillow, an online real estate marketplace, recently shuttered its Zillow Offers business because of failed iBuying algorithms. A derailed algorithm on property valuations led the company to reduce the estimated value of the houses it purchased in Q3 and Q4 by more than $500 million. Zillow has already officially announced $304 million in Q3 losses and expects to reduce its workforce by 25%

GOT A TIP? The Hollywood Reporter

NEWS FILM TV AWARDS LIFESTYLE BUSINESS GLOBAL VIDEO

HOME > BUSINESS > BUSINESS NEWS

Tyler Perry Puts $800M Studio Expansion on Hold After Seeing OpenAI's Sora: "Jobs Are Going to Be Lost"

What Surprised You About The Chapter?

Some Surprises

No surprises since we have experience with ML, but wanted introduce a poll to gauge the audience's ML experience. (Andrei)

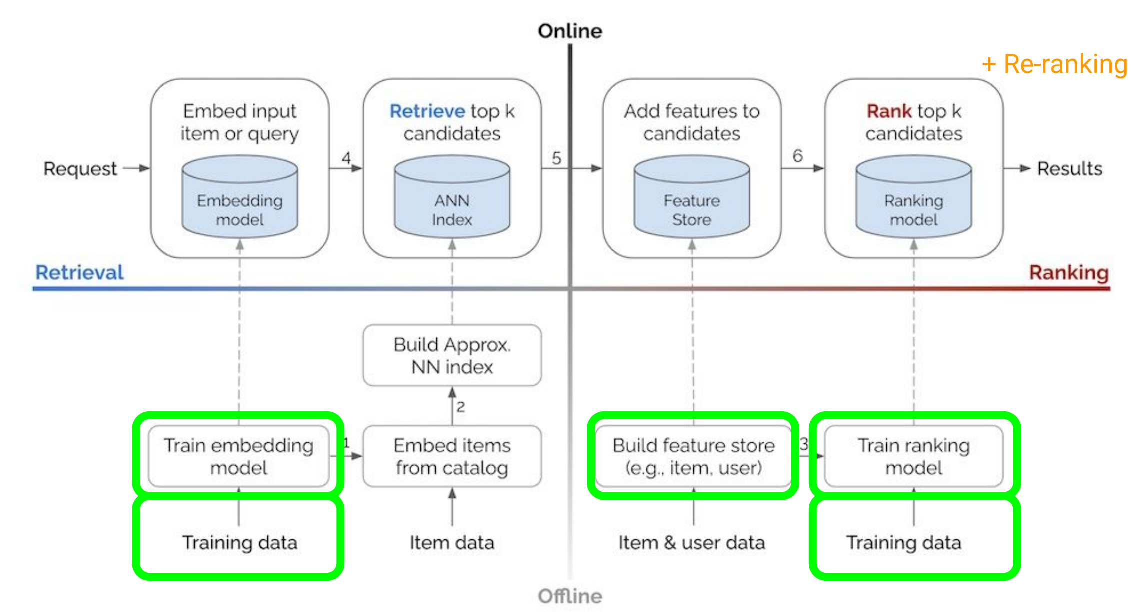
[Live Zoom poll] What is your experience with ML?

- I think random forests are a good solution to climate change
- I completed self-paced courses/personal projects using Python/Jupyter
- I did research/projects at an academic institution studying ML
- I developed products in the market

Now to Shivam for some Case studies

Related Examples & Case Studies

RecSys

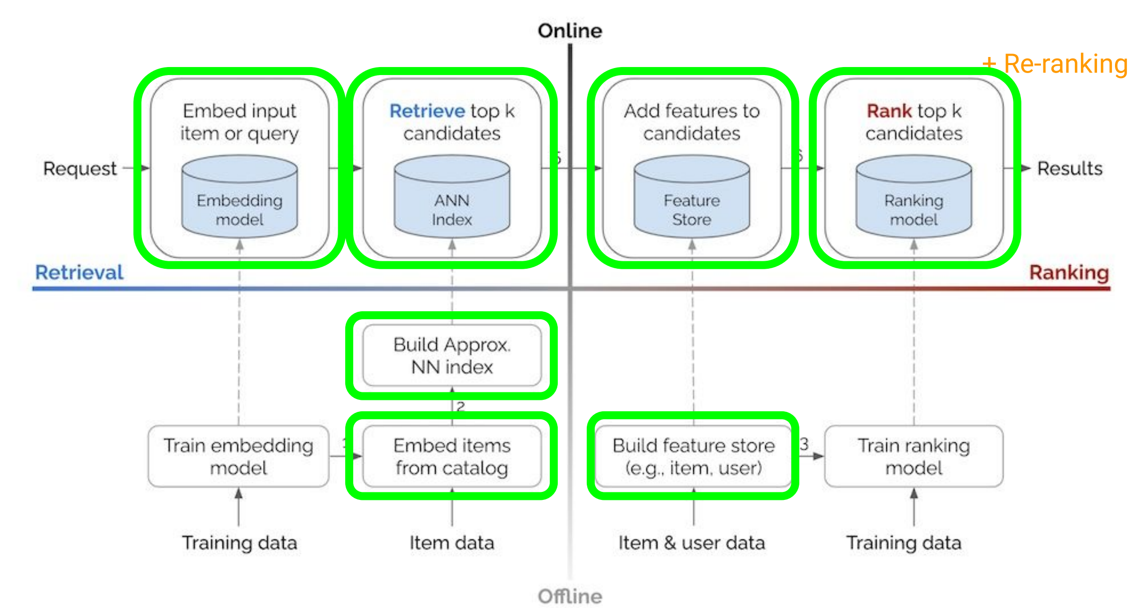


Ref: https://eugeneyan.com/writing/system-design-for-discovery/

(ML Solution) ML + Biz View

1. Data
2. Retrieval (candidate gen(s))
3. Ranking
   a. CTR/CVR
   b. Multi-objectives
4. Re-ranking (biz objectives)
   a. Ads revenue
   b. Fairness
   c. Diversity
5. Metrics + Evaluation
6. A/B Testing

RecSys

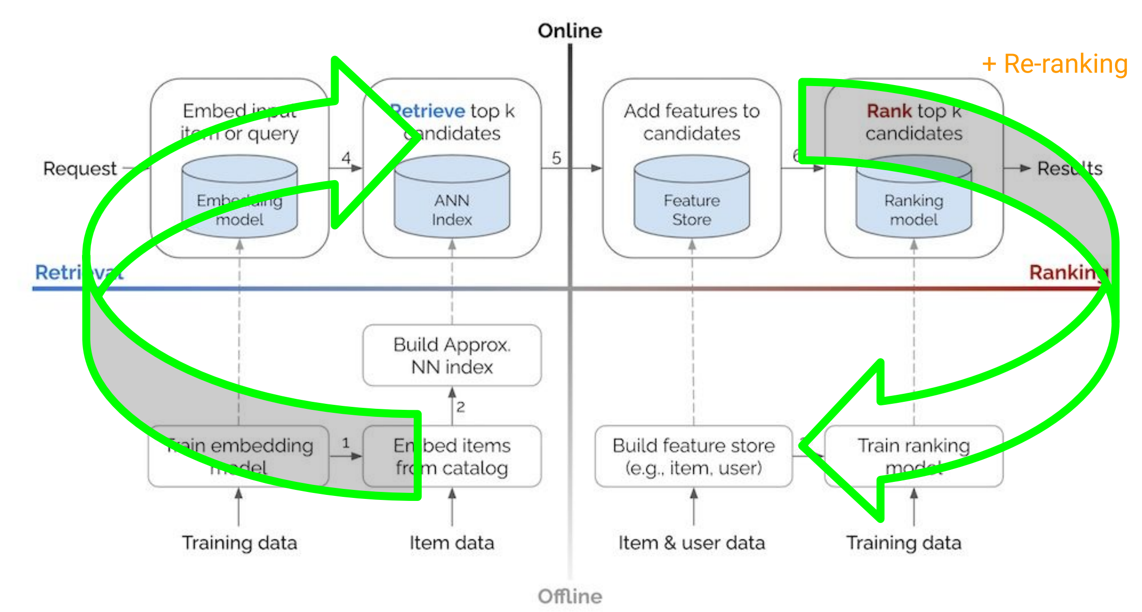


Ref: https://eugeneyan.com/writing/system-design-for-discovery/

(Prod) ML Engineering View

1. **Deployment @ scale**
   a. Latency
   b. Throughput
   c. Model size
2. **Offline Compute**
3. **Observability**
   a. Logging
   b. Debugging & RCAs
   c. Job failures
   d. Error rates
4. **Cost Optimization**

RecSys



Ref: https://eugeneyan.com/writing/system-design-for-discovery/

(Prod) ML Engineering View ++

1. **Retraining**
   a. **Biz requirements**
   b. **Continuous improvements**
   c. **Changing data patterns**
2. **Reproducibility**
   a. **Data versioning**

Evals

- **Loss functions:** Research's BFF, Production's frenemy.
- **Benchmarks:** Great for controlled experiments, not real-world chaos.
- **Production means:** noisy data, distribution shifts, and custom metrics.

Personal Experiences

An Academic Breaking Into The Industry

BEFORE

1. The Model Is The End All, Be All
2. Scoffed At Measurement & Maintenance
3. Just Let Me Use The Latest & Greatest!

NOW

* Importance of alignment
* Strike a balance between different objectives

* Don't even bother without proper measurement facilities

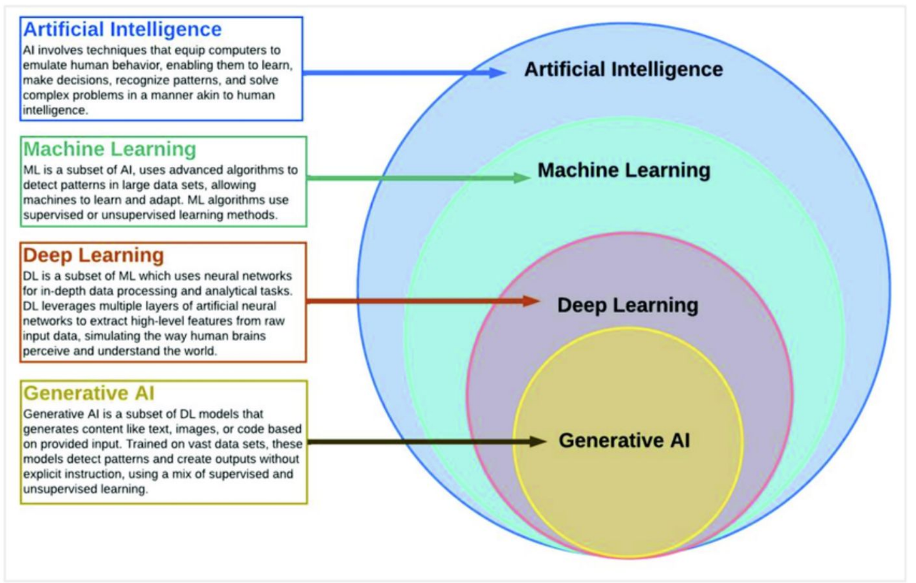
* Doesn't make sense for a lot of cases

What’s Changed Now With GenAI? What Has Remained The Same?

What Changed with Generative AI

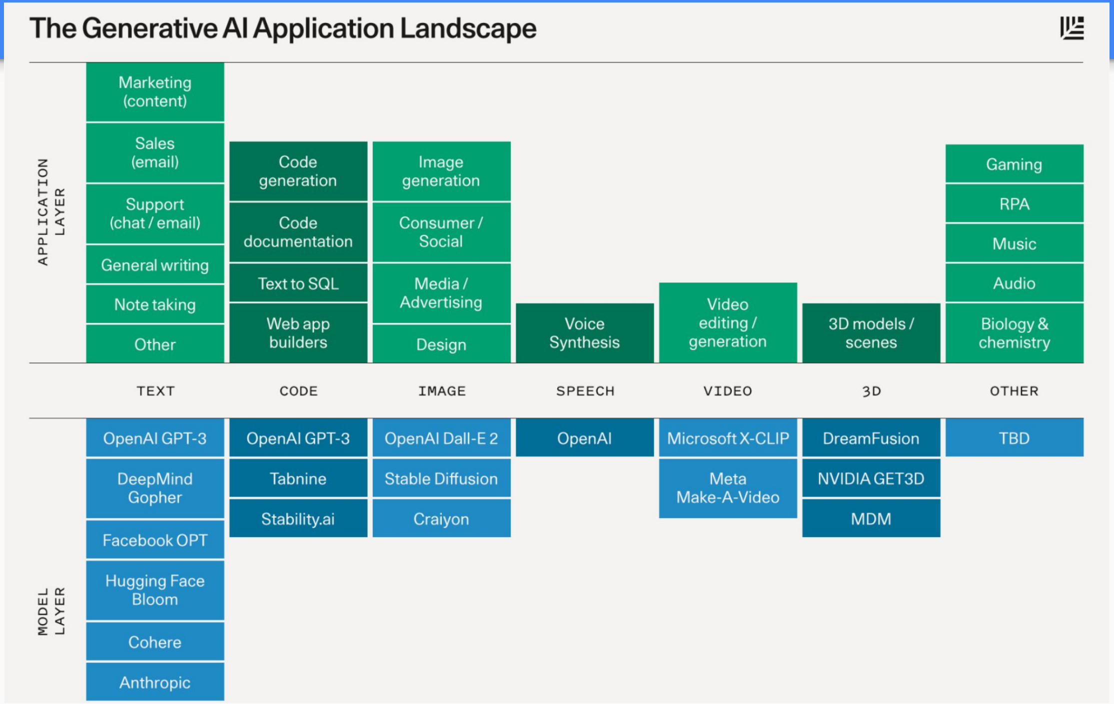
Discriminative/Analytical ML
Traditionally used to analyze data, identify patterns and make predictions based on existing data.

Generative AI
More recently with advancements in the field (e.g. Transformer model) AI methodologies have become better at generating new things (rather than just analyzing existing things).



Democratization of AI

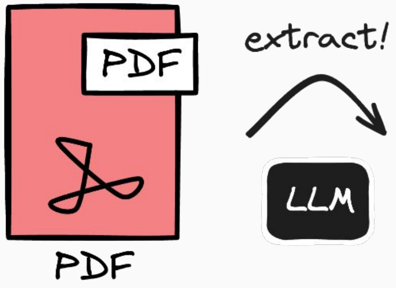
- Traditional ML required having access to a lot of labelled datasets and compute power
- Emergence of Foundational models (applicable across wide range of use cases) usable via an API changes things
- Every engineer becomes an AI engineer!
- The world of Machine learning engineering and Software engineer are coming closer



Trust in GenAI?

"While most of use are comfortable with using a microwave without understanding how it works, many don't feel the same way about AI yet..."

EXAMPLE




```
{
    "ceo": "jane s.",
    "coo": "bob t.",
    "investors": [
        "abc investments",
        "def capital"
    ]
}
```

QUESTION How can I trust the LLM is extracting the right fields?

* Those who are not in ML or using GenAI are probably the least comfortable ...
* How can we help them and others learn to trust AI as these get more and more accurate?
* Trust in GenAI won't be a step function.

In your breakout rooms...

Scenario

- Customer reviews for an e-commerce platform
- Researchers developed a sentiment analysis model
- The team is preparing to deploy this model in production

Discuss

- What are the **potential discrepancies** between the research-focused evaluation what good performance is like in the production environment?

- How should the team adapt their evaluation strategy to ensure the model's effectiveness in achieving the business goals of the e-commerce platform?

Thanks!

Chad SHAPLV (Group 1)

Shivam Rana
https://www.linkedin.com/in/tminima

Harpreet Sahota
https://www.linkedin.com/in/harpreetsahota204

Andrei Fajardo
https://www.linkedin.com/in/nerdai

Prakul Agarwal
https://www.linkedin.com/in/prakulagarwal

LVG

Vern Liang
https://www.linkedin.com/in/vernliang