

From Netflix Recommendations to  
Conversational Multi-agents:

# The (R)Evolution of AI-Driven Product Innovation

**Xavi Amatriain**  
VP of Product, Google Cloud



Create a few friendly-looking robots. Each one is holding either a movie DVD or a music CD and is recommending to watch or listen to it.  
DALL-E 3

# About me

Researcher in Recommender Systems

 Started and led ML Algorithms at Netflix

 VP of Engineering at Quora

 Co-founder/CTO at Curai

 VP of AI Product Strategy at LinkedIn

 Google Cloud

VP of Product

Core ML/AI at Google

# Outline

01

The past:

Data & Algorithms-driven  
product innovation

02

Yesterday:

ML -> Deep Learning

03

Today:

Product innovation  
in the Age of (Gen)AI

01

# The Past: Data & Algorithms-driven product innovation

## Netflix's New 'My List' Feature Knows You Better Than You Know Yourself (Because Algorithms)

The Huffington Post | By Dino Grandoni  

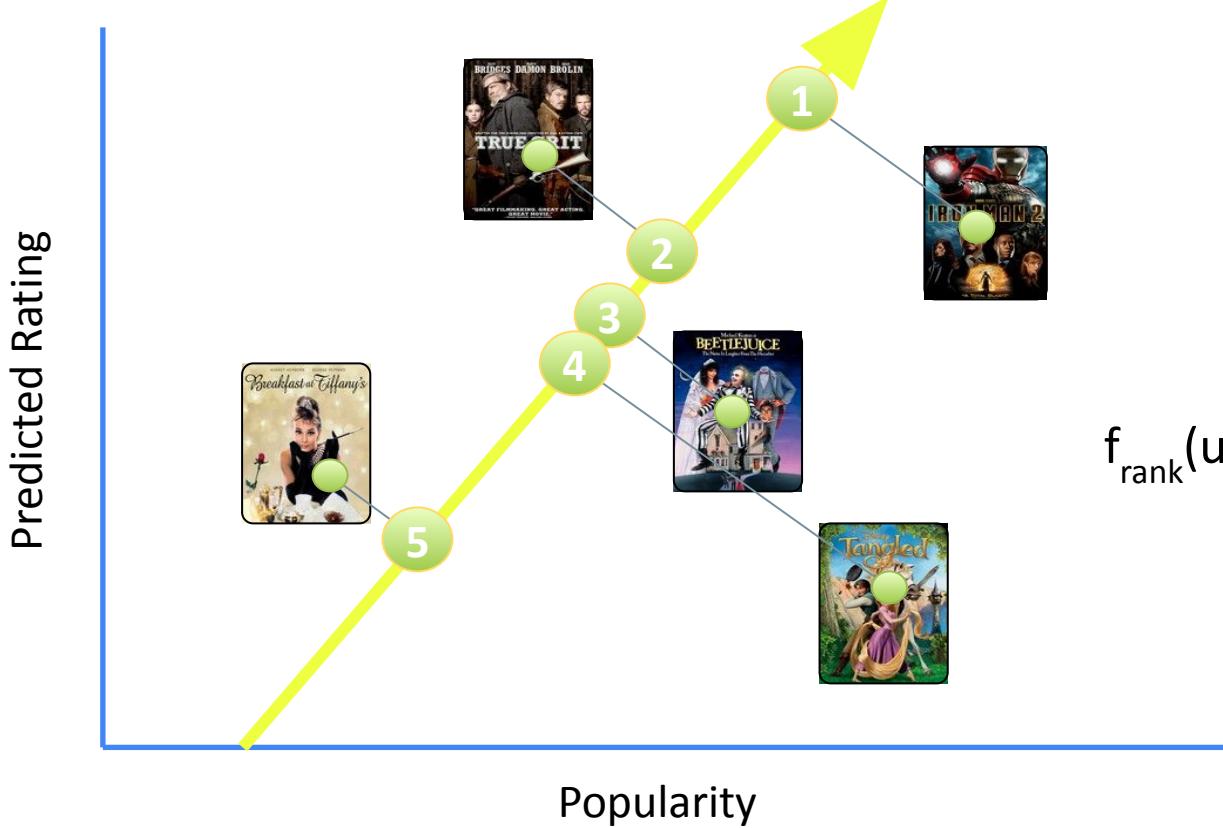
Posted: 08/21/2013 1:44 pm EDT | Updated: 08/22/2013 8:31 am EDT



55 people like this. Be the first of your friends.

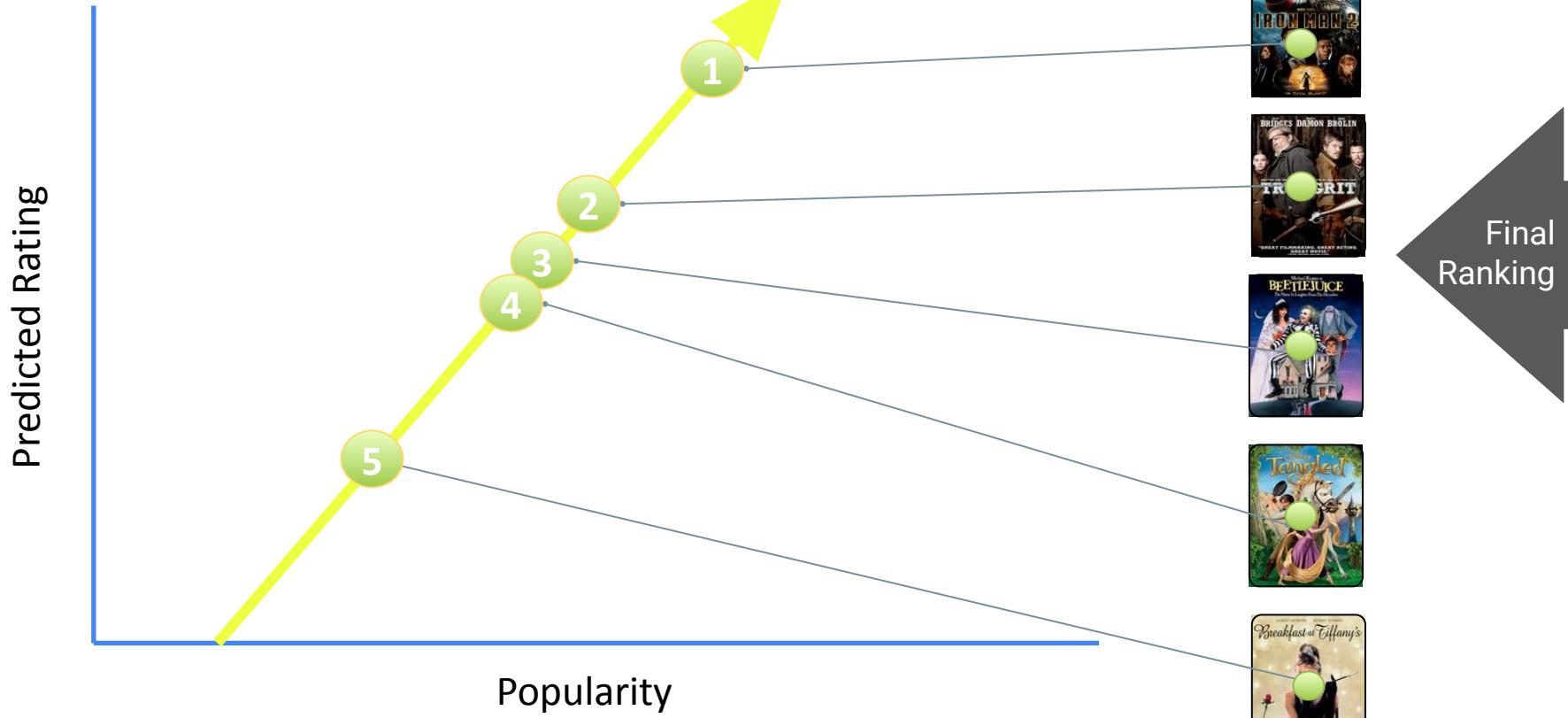


# Machine Learning (the “old AI”)

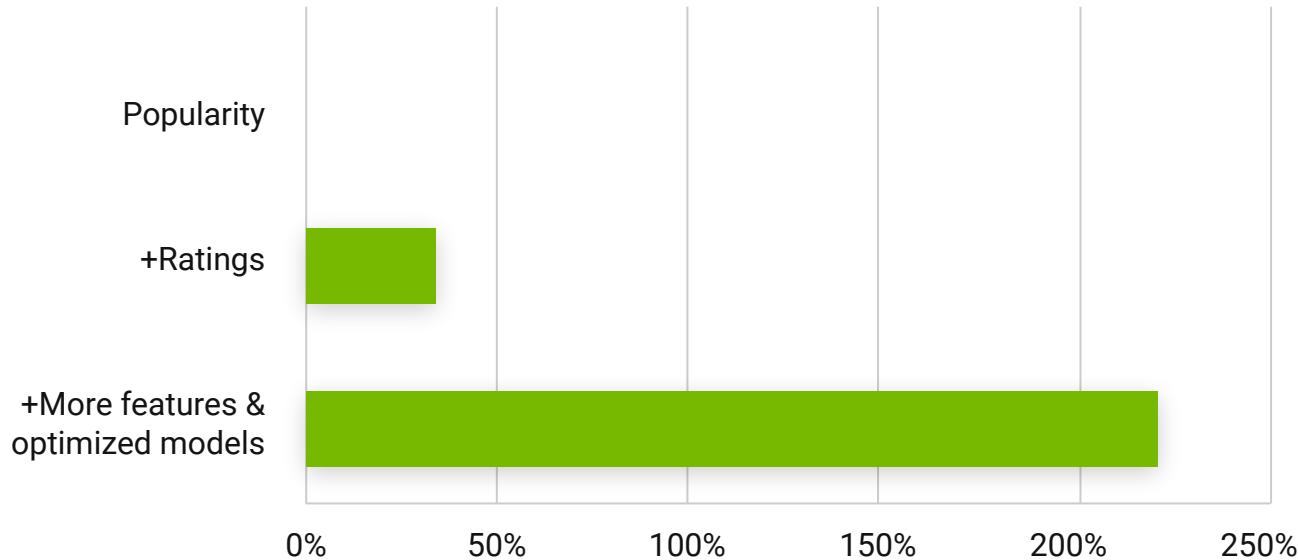


Linear Model:  
 $f_{\text{rank}}(u,v) = w_1 p(v) + w_2 r(u,v) + b$

# Machine Learning (the “old AI”)



# Ranking improvement over baseline



# Other important lessons learned

# Explanations (and UX in general) matters!

• Sarah Smith • Richard Henry and 3 more upvoted this • 7h

**How can I complain about my roommate who is cheating on his Google phone interviews?**

 Ben Garrison, Software Engineer at Google  
304.3k Views • Upvoted by Jeremy Miles, Quantitative analyst at Google, Mayeesha Tahsin, Sarah Smith, and 3 others you follow

First off, I really appreciate your trying to make sure the right thing happens. I think that's great. Cheating sucks. However, the answer is "don't worry about it". Phone screens here at Google ar... (more)

[Upvote | 968](#) [Downvote](#) [Comments 23+](#) [Share](#)

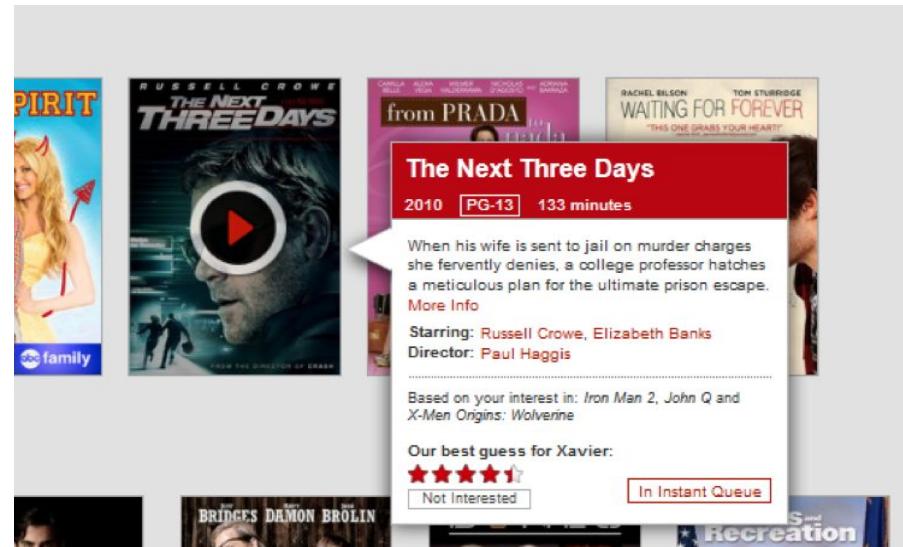
\*\*\*

 Discover new topics

**Last.fm**  
Last.fm builds detail...  
Followed by Neal Lathia and 8 more  
[Follow | 21.9k](#)

 **Quantitative Finance**  
Quantitative finance ...  
Followed by Katie Hoban and 22 more  
[Follow | 74.1k](#)

 **California State Un...**  
California State Univ...  
Followed by Rachelle Baratto  
[Follow | 2.4k](#)



**SPIRIT** (TV Movie 2008)

**RUSSELL CROWE** **THE NEXT THREE DAYS** (2010)

**from PRADA** (2007)

**RACHEL BILSON** **WAITING FOR FOREVER** (2010)

**The Next Three Days** (2010) PG-13 133 minutes

When his wife is sent to jail on murder charges she fervently denies, a college professor hatches a meticulous plan for the ultimate prison escape.

[More Info](#)

Starring: Russell Crowe, Elizabeth Banks  
Director: Paul Haggis

Based on your interest in: Iron Man 2, John Q and X-Men Origins: Wolverine

Our best guess for Xavier:

★★★☆?

[Not Interested](#) [In Instant Queue](#)

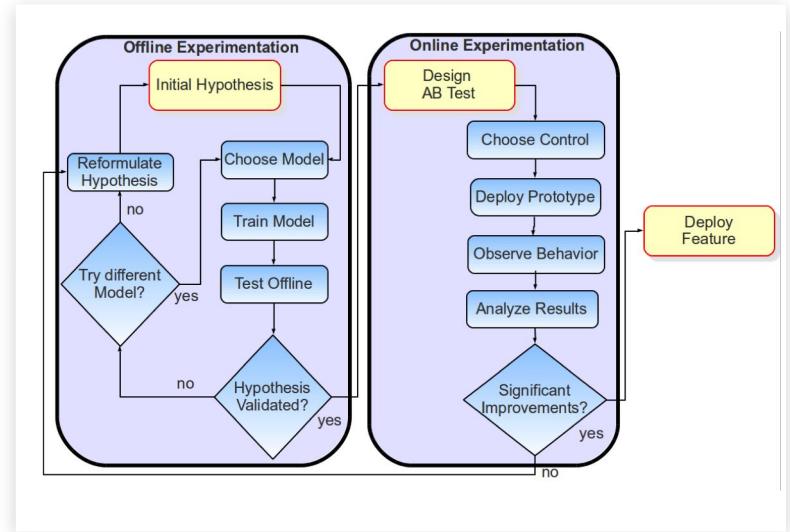
# The importance of the experimentation framework

## Offline

- Prefer ranking or UX-related metrics
- Measure other aspects such as diversity or novelty
- Keep bank of metrics that you can map to online results with post-hoc analysis

## Online

- Measure differences in metrics across statistically identical populations exposed to different algorithms.
- Overall Evaluation Criteria (OEC)
  - Use **long-term metrics** whenever possible
  - **Short-term metrics** can be informative and allow faster decisions. But, not always aligned with OEC



# Domain knowledge

What is a good Quora answer?



How are those dimensions translated into features?

- Features that relate to the answer
- Quality itself
- Interaction features (upvotes/downvotes, clicks, comments...)
- User features (e.g. expertise in topic)



Paula Griffin, data scientist and biostatistics PhD ... (more)

13 upvotes by William Chen, Alexandr Wang (王晉舜), Sheila Christine Lee, (more)

I was figuring that this question was just fishing for someone to answer that Big Data is their favorite band. Unfortunately, the question log indicates this was asked about 6 months before their EP came out, so there goes that theory.

This is going to be a pretty odd list, but here's the list, in order of decreasing social acceptability:

- Electropop -- Banks and CHVRCHES are my favorites at the moment.
- Miscellaneous alt-rock -- this category basically includes anything I found out about from listening to Sirius XM in the car.
- Nerd rock -- What kind of geek would I be if Jonathan Coulton wasn't on this list?
- Straight-up nostalgia -- I have an admittedly weird habit of listening to the same album (sometimes just one song) over and over for hours on end which was formed during all-nighters in high school. Motion City Soundtrack, Jimmy Eat World, and Weezer are my go-to's in this category.
- Soundtracks of all sorts -- *Chicago*, *Jurassic Park*, *Bastion*, *The Book of Mormon*, the Disney version of *Hercules*... again, basically anything that works on a repeat loop for ~3 hours.
- Pop -- don't make me list the artists. I've already told you I listen to Disney soundtracks; you can't possibly need more dirt on me. The general principle is that if you can dance to it, you can code to it.

Now, if you don't mind, I'm just going to sit at my desk and be super-embarrassed that my coworkers know what's in my headphones.

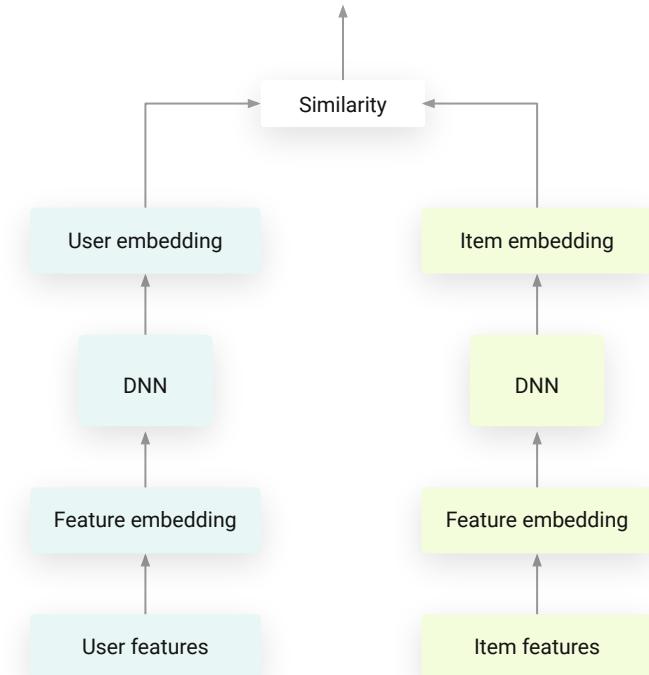
Written 4 Dec. 353 views. Asked to answer by William Chen.

[Upvote](#) 13 [Downvote](#) [Comment](#) [Share](#)

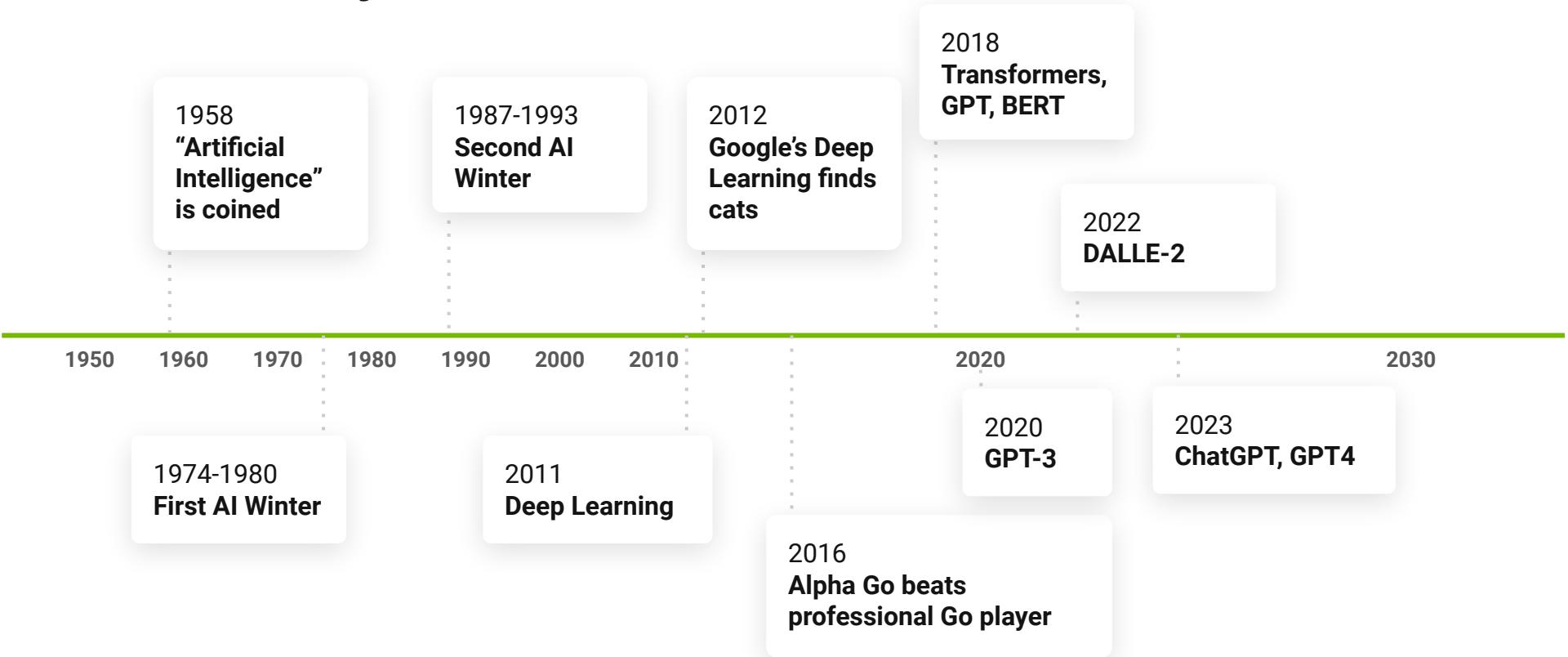
...

01

# Yesterday: ML -> Deep Learning



# The history of AI



# Deep learning

## Applying deep learning to AirBnB search

Malay Haldar, Mustafa Abdool, Prashant Ramanathan, Tao Xu, Shulin Yang, Huizhong Duan, Qing Zhang, Nick Barrow-Williams, Bradley C. Turnbull, Brendan M. Collins and Thomas Legrand  
Airbnb Inc.  
malay.haldar@airbnb.com

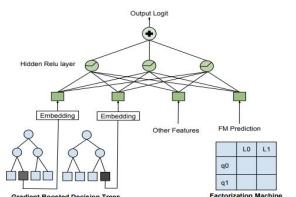
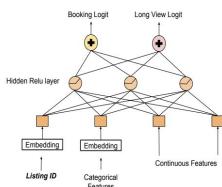


Figure 3: NN with GBDT tree nodes and FM prediction as features



## Wide and deep learning for recommender systems

Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushare Chandra, Hrishi Aradhye, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Ispir, Rohan Anil, Zakaria Haque, Lichen Hong, Vihan Jain, Xiaobing Liu, Hemal Shah  
Google Inc.

Google Inc.

malay.haldar@airbnb.com

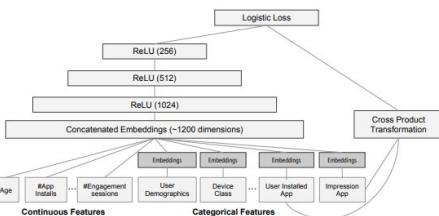


Figure 4: Wide & Deep model structure for apps recommendation.

## Deep Neural Networks for YouTube Recommendations

Paul Covington, Jay Adams, Emre Sargin  
Google  
Mountain View, CA  
[pcovington, jka, msargin]@google.com

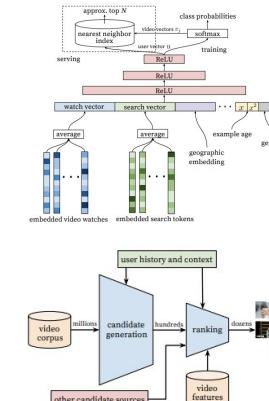


Figure 2: Recommendation system architecture demonstrating the “funnel” where candidate videos are retrieved and ranked before presenting only a few to the user.

# Important concepts in modern AI production systems

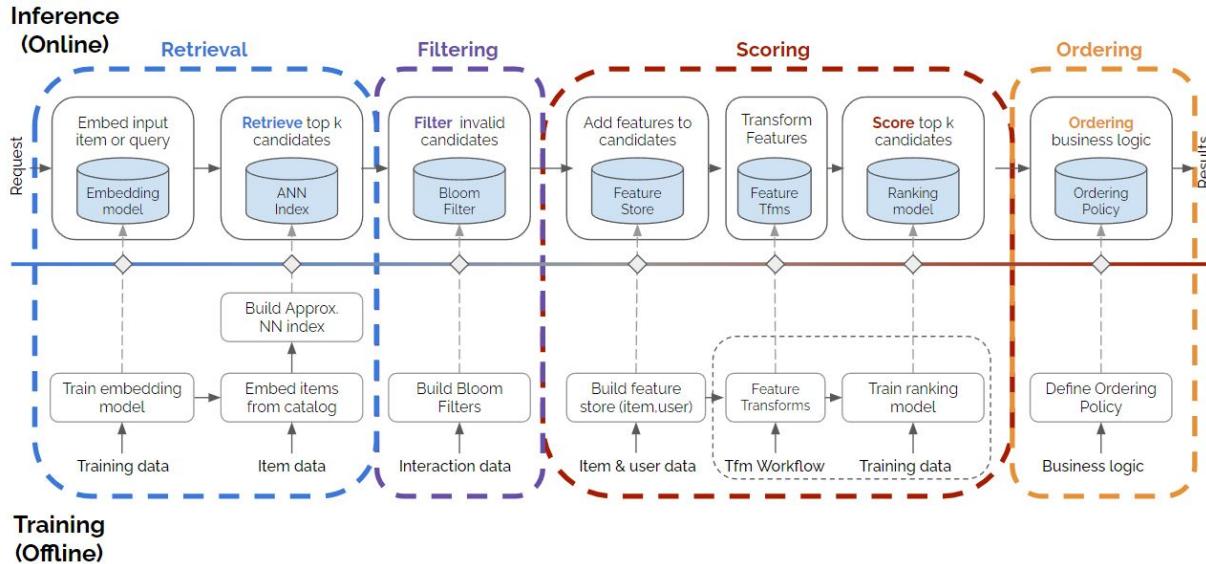
01

Multi-stage systems

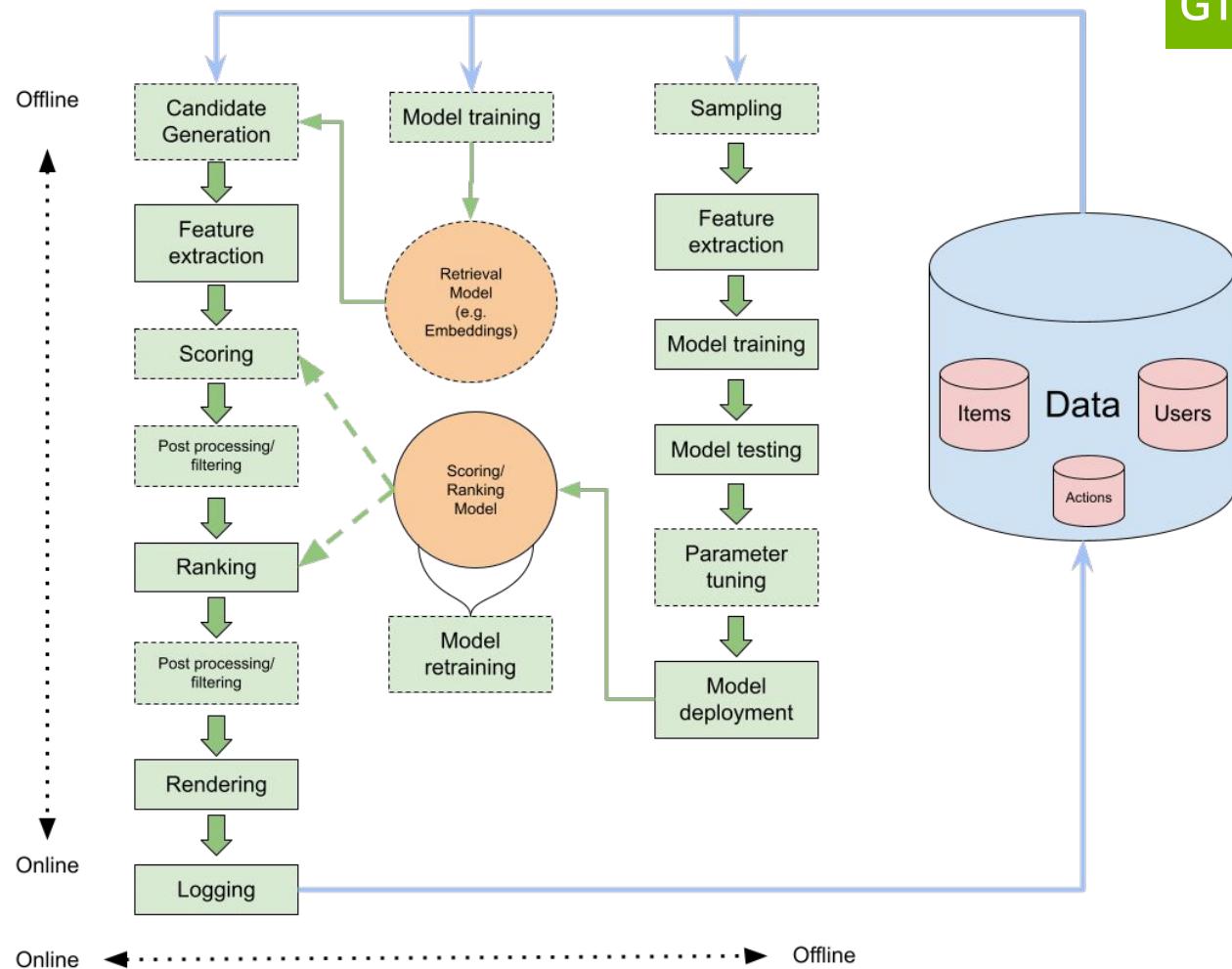
02

Embeddings

# Multi-stage systems (NVIDIA version)



# Multi-stage systems (My version)



# Embeddings

## Billion-scale commodity embedding for e-commerce recommendation in Alibaba

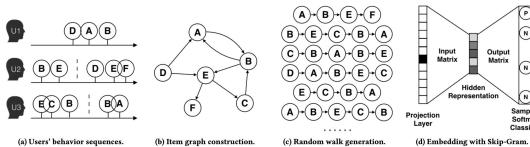
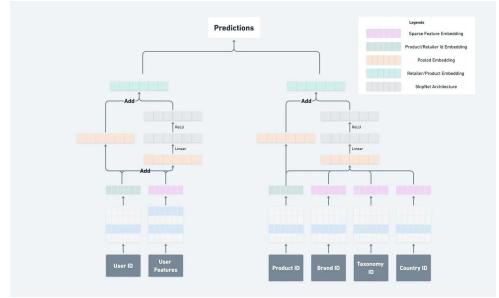


Figure 2: Overview of graph embedding in Taobao: (a) Users' behavior sequences: One session for user  $u_1$ , two sessions for user  $u_2$  and  $u_3$ ; these sequences are used to construct the item graph; (b) The weighted directed item graph  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ ; (c) The sequences generated by random walk in the item graph; (d) Embedding with Skip-Gram.



Figure 5: Similar items for cold start items. Top 4 similar items are shown. Note that "cat" means category.

## How we use engagement-based embeddings to improve search and recommendation on Faire



DECEMBER 1, 2022

## Introducing DreamShard: A reinforcement learning approach for embedding table sharding

By: Daochen Zha, Bhargav Bhagcharam, Louis Feng, Liang Luo, Yusuo Hu, Yuandong Tian, Jade Nie, Dima Ivashchenko, Mykola Lukashenko, Yuzhen Huang



02

# Today:

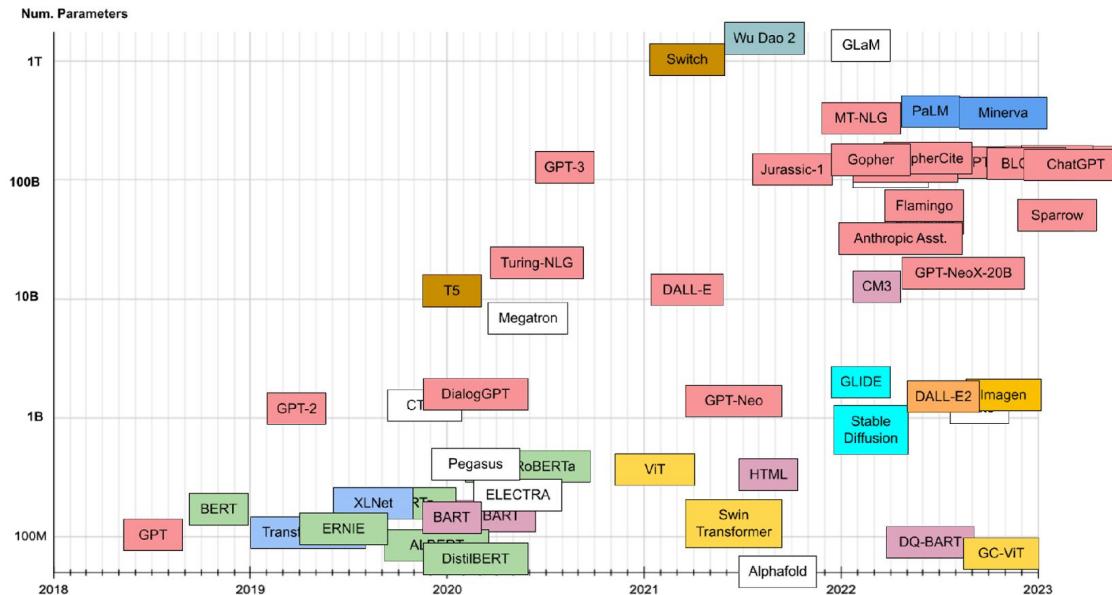
## The Age of (Gen)AI



Generate an  
impressionist style  
painting of a robot  
reading a book  
Gemini Advanced

# The history of AI (part II)

In just the last  
five years...



# GenAI components

GPT



Generative

Generate new content based on a natural language input (prompt)



Pretrained

Trained on the whole internet and more  
Can learn on the fly  
Can be fine-tuned



Transformer

Just a different kind of Deep Learning architecture using “attention”

# Human-level AI

- GPT-4 out of the box:
  - Scores in the 90th percentile for the BAR exam to become a lawyer
  - Scores in the 90th percentile for the SAT both in Math and Reading and Comprehension
  - Exceeds USMLE score required to graduate in medicine in the US by 20 points
  - It is much better at coding than many software engineers
- Gemini Ultra and Claude 3 beat it at many of these tasks

Simulated exams	GPT-4 90th percentile	GPT-4 (no vision) 90th percentile	GPT-3.5 90th percentile
Uniform Bar Exam (MBE+MEE+MPT) <sup>1</sup>	298/400 90th percentile	298/400 90th percentile	213/400 10th percentile
LSAT	163 90th percentile	161 90th percentile	149 90th percentile
SAT Evidence-Based Reading & Writing	710/800 90th percentile	710/800 90th percentile	670/800 90th percentile
SAT Math	700/800 90th percentile	690/800 90th percentile	590/800 90th percentile
Graduate Record Examination (GRE) Quantitative	163/170 90th percentile	157/170 90th percentile	147/170 90th percentile
Graduate Record Examination (GRE) Verbal	169/170 90th percentile	165/170 90th percentile	154/170 90th percentile
Graduate Record Examination (GRE) Writing	4/6 90th percentile	4/6 90th percentile	4/6 90th percentile
USABO Semifinal Exam 2020	87/150 90th-100th	87/150 90th-100th	43/150 10th-50th
USNCO Local Section Exam 2022	36/60 90th percentile	38/60 90th percentile	24/60 90th percentile
Medical Knowledge Self-Assessment Program	75%	75%	53%
Codeforces Rating	392 90th percentile	392 90th percentile	260 90th percentile
AP Art History	5 90th-100th	5 90th-100th	5 90th-100th
AP Biology	5 90th-100th	5 90th-100th	4 90th-100th
AP Calculus BC	4 90th-100th	4 90th-100th	1 90th-100th
AP Chemistry	4 The 80th	4 The 80th	2 The 40th
AP English Language and Composition	2 The 40th	2 The 40th	2 The 40th
AP English Literature and Composition	2 90th-100th	2 90th-100th	2 90th-100th
AP Environmental Science	5 90th-100th	5 90th-100th	5 90th-100th
AP Macroeconomics	5 90th-100th	5 90th-100th	2 30th-40th
AP Microeconomics	5 90th-100th	4 90th-100th	4 90th-100th
AP Physics 2	4 90th-94th	4 90th-94th	3 30th-60th
AP Psychology	5 90th-100th	5 90th-100th	5 90th-100th
AP Statistics	5 90th-100th	5 90th-100th	3 90th-100th
AP US Government	5 90th-100th	5 90th-100th	4 70th-90th
AP US History	5 90th-100th	4 70th-90th	4 70th-90th
AP World History	4 90th-100th	4 90th-100th	4 90th-100th
AMC 10	30/150 90th percentile	36/150 90th percentile	36/150 90th percentile
AMC 12	60/150 90th percentile	48/150 90th percentile	30/150 40th-60th
Intro Sommelier (theory knowledge)	92%	92%	80%
Certified Sommelier (theory knowledge)	86%	86%	58%
Advanced Sommelier (theory knowledge)	77%	77%	46%
Leetcode (easy)	31/41	31/41	12/41
Leetcode (medium)	21/80	21/80	8/80
Leetcode (hard)	3/45	3/45	0/45

# GenAI is already re-defining most companies' roadmaps

Forbes

FORBES > INNOVATION > ENTERPRISE TECH

## Microsoft's Plan To Infuse AI And ChatGPT Into Everything

Bernard Marr Contributor 

0 Mar 6, 2023, 02:22am EST

 Listen to article 9 minutes 

 Microsoft has big plans for artificial intelligence (AI), and it's becoming clear that it believes that ChatGPT – and natural language technology in general - will play a big part.

 in



## What if ChatGPT was trained on decades of financial news and data? BloombergGPT aims to be a domain-specific AI for business news

The news and data giant has — with a relatively small team — built a generative AI that it says outperforms the competition on its own specific information needs.

Businessweek  
Technology

## Google's Plan to Catch ChatGPT Is to Stuff AI Into Everything

A new internal directive requires "generative artificial intelligence" to be incorporated into all of its biggest products within months.



**Let's try LLMs for  
recommendations!**



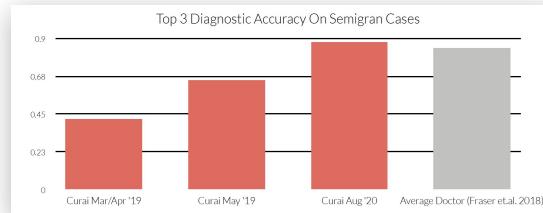
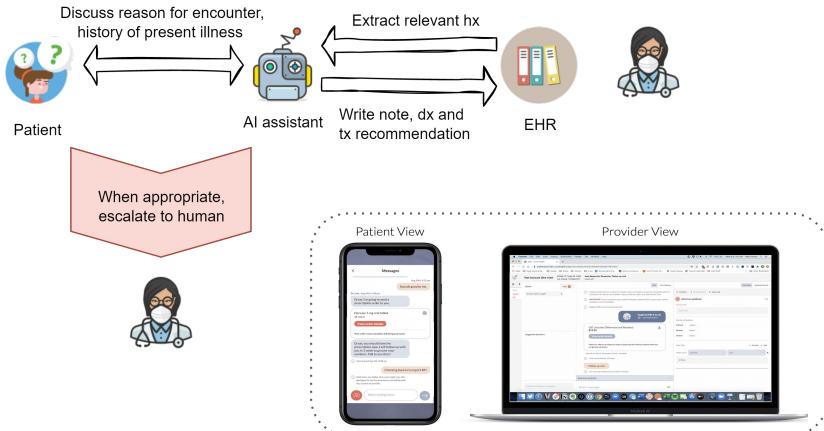
# Basic LLM recommendations

“Here are some examples of Netflix shows I really liked: Breaking Bad, Bojack Horseman, Shameless, and Peaky Blinders. Can you recommend other Netflix shows?”

# Beyond basic LLM recommendations

“Let's play a game. Ask me 5 yes or no questions (one at a time) and then recommend some music artists I will like.”

# Everything will be disrupted by AI. Very quickly



**MEDCOD: A Medically-Accurate, Emotive, Diverse, and COntrollable Dialog System**

Proceedings of Machine Learning Research 126:1–16, 2021

Riley Compton<sup>1</sup>, Bhargav Chintagunta<sup>1</sup>, Li Deng<sup>2</sup>, Costa Hatzigeorgiou<sup>3</sup>, Naman Katariya<sup>4</sup>, Xavier Amatriain<sup>5</sup>, Anitha Kannan<sup>6</sup>, Chen<sup>7</sup>

**Medically Aware GPT-3 as a Data Generator for Dialogue Summarization**

Bharath Chintagunta<sup>1</sup>, Naman Katariya<sup>1</sup>, Xavier Amatriain<sup>1</sup>, Anitha Kannan<sup>1</sup>

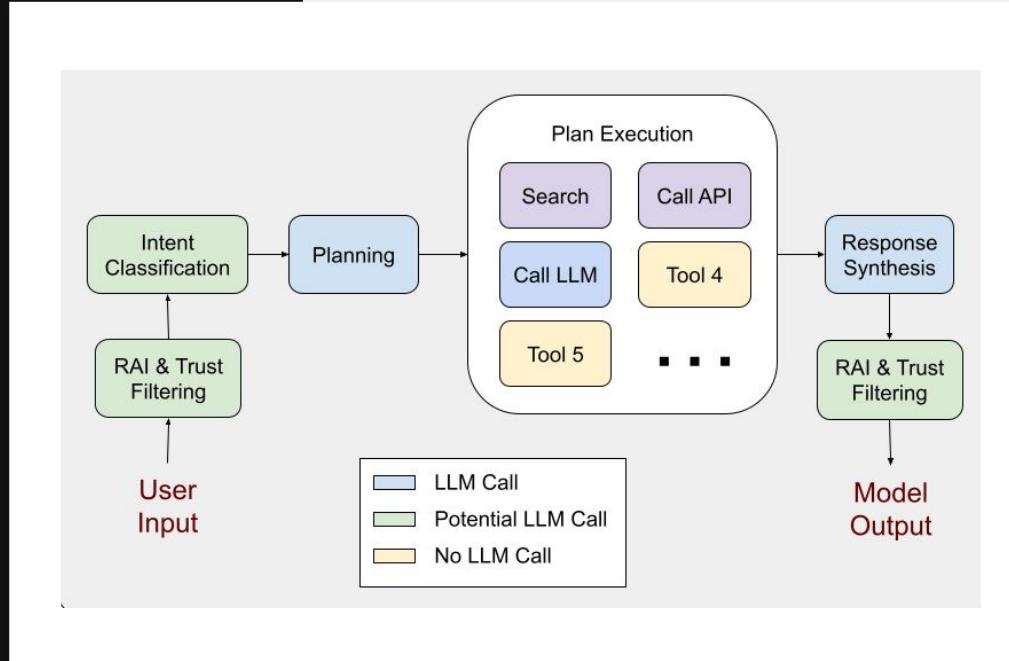
**Abstract**  
In medical dialogue summarization, summarizers must track the medically relevant information in the dialogue. To do this, summarizers require large amounts of labeled data to process. In this paper, we propose a novel way to generate such data: by utilizing a pre-trained language model (GPT-3) and a set of 210 human-curated examples to yield results labeled with medical entities. We validate our approach in experiments; we show that this approach produces high-quality summaries that are comparable in quality to those produced by models trained on human data also and coherence.

**Open Set Medical Diagnosis**

Viraj Prabhu<sup>\*,1</sup>, Anitha Kannan<sup>3</sup>, Geoffrey J. Tso<sup>2</sup>, Naman Katariya<sup>3</sup>, Manish Chahal<sup>3</sup>, David Sontag<sup>\*,2</sup>, Xavier Amatriain<sup>3</sup>  
\*Georgia Tech      <sup>2</sup>MIT      <sup>3</sup>Curai

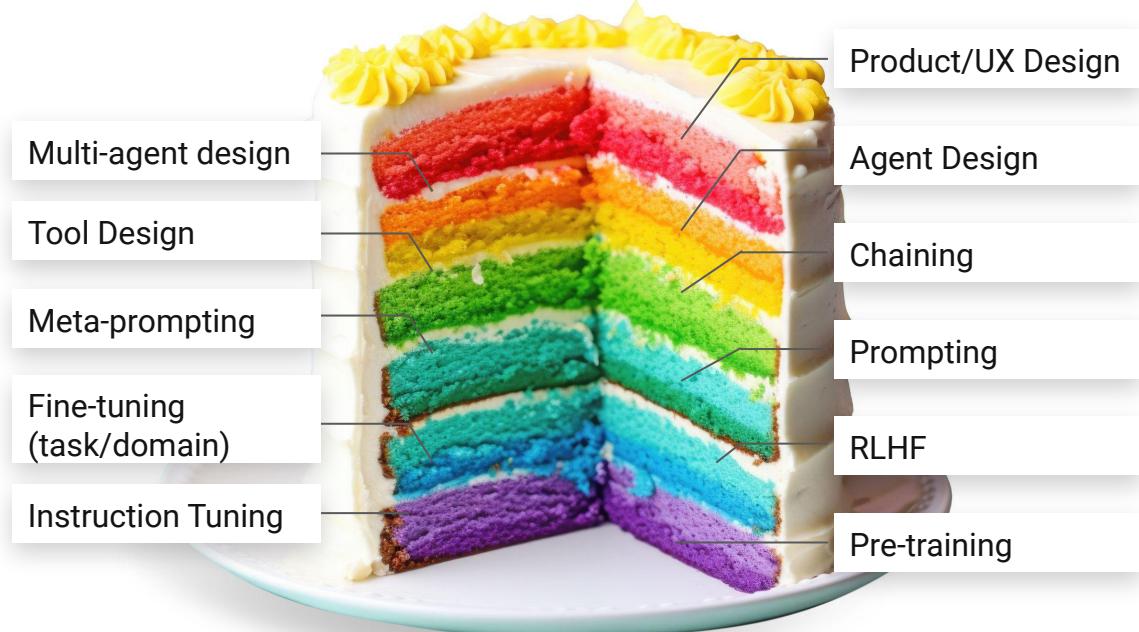
**Abstract**  
Machine-learned diagnosis models have shown promise as medical aids but are trained under a closed-set assumption, i.e., that models will only encounter conditions on which they have been trained. However, it is practically infeasible to obtain sufficient training data for every human condition, and once deployed such models will encounter new conditions that were not seen during training. We view machine-learned diagnosis as an *open-set* learning problem, and study how state-of-the-art approaches compare. Further, we extend our study to a setting where training data is *open-set* and the test data is *closed-set*. We propose a novel training strategy and experiment with different strategies of building open-set diagnostic ensembles. Across both settings, we observe consistent gains from explicitly modeling unseen conditions, but find the optimal training strategy to vary across settings.

# What's behind an LLM application



# The multilayer cake of GenAI

Evaluation (Quality, Hallucination detection, RAI...)



# How is this similar to the “old AI”

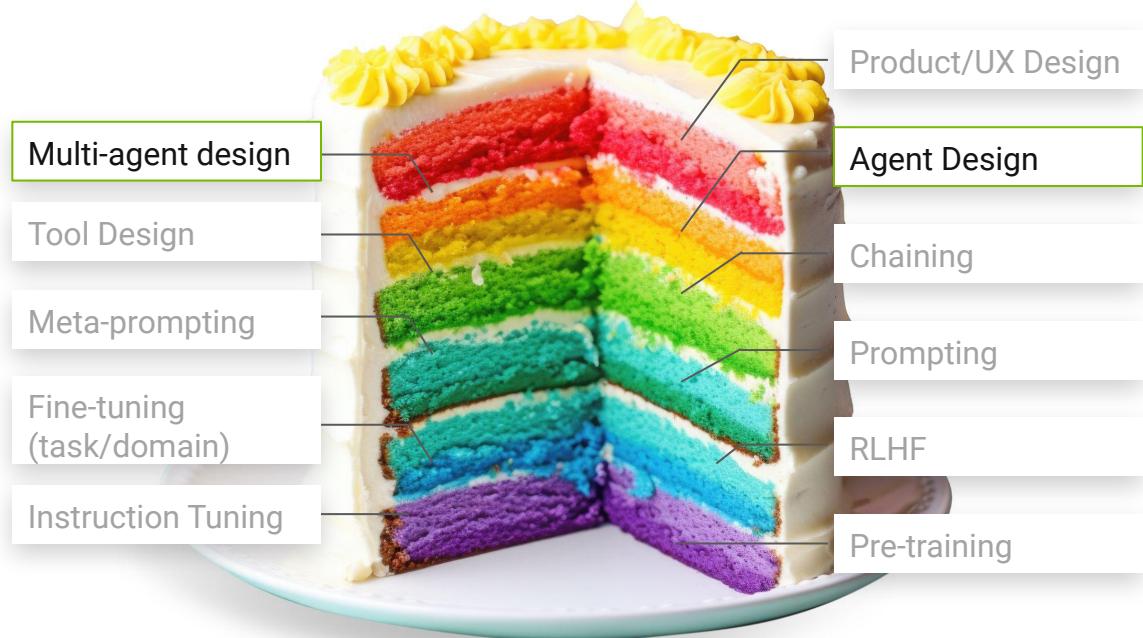
- 1** Importance of Product and UX design
- 2** Importance of evaluation and metrics
- 3** Importance of domain knowledge (e.g. healthcare)

# How is this different to the “old AI”

- 1 Now the UX **is** the AI
- 2 We need new evaluation metrics and frameworks (e.g. hallucination)
- 3 Domain knowledge is needed but is also present in the models

# Let's focus on Agents and Multi-agents

Evaluation (Quality, Hallucination detection, RAI...)



# Agents

Agent = LLM based system that has access to tools and can decide how to use them

- Agents are hard to maintain and implement and tools/libraries like LangChain, Auto-GPT, or AutoGen come to the rescue

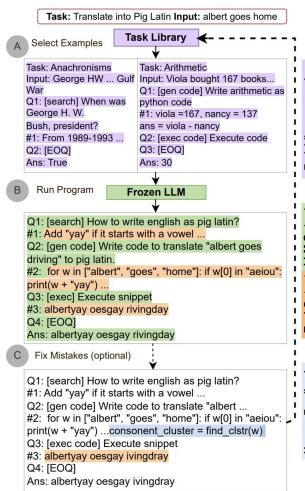
The diagram is divided into four main sections:

- Top Left:** A schematic showing an "LLM" box connected to a "Calculator" icon and a "Code Editor" icon, with a "Source" box above it, representing an Agent's access to various tools.
- Top Right:** A screenshot of the NVIDIA Developer Technical Blog titled "Introduction to LLM Agents". The post is dated Nov 30, 2023, by Tanay Varsney, with 24 likes and 0 discussions. It features a purple banner at the bottom labeled "Part 1" with a gear icon.
- Bottom Left:** A section titled "Conversable agent" showing three separate agents (blue robot head, person icon with wrenches, green person icon with Python logo) each enclosed in a dashed box, indicating they can interact with each other.
- Bottom Right:** A section titled "Flexible Conversation Patterns" showing:
  - Multi-Agent Conversations:** Two agents in dashed boxes connected by a double-headed arrow.
  - Joint chat:** Three agents in dashed boxes connected in a triangular loop.
  - Hierarchical chat:** A tree-like structure where multiple agents in colored boxes (orange, green, yellow) are interconnected.

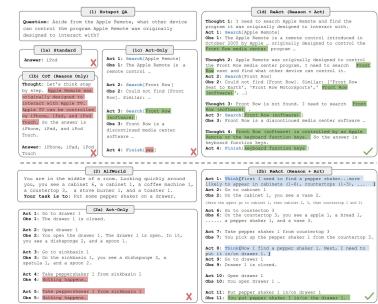
*Figure 1. AutoGen enables complex LLM-based workflows using multi-agent conversations. (Left) AutoGen agents are customizable and can be based on LLMs, tools, humans, and even a combination of them. (Top-right) Agents can converse to solve tasks. (Bottom-right) The framework supports many additional complex conversation patterns.*

# Implementing multi-agent systems

## ART (Automatic multi-step reasoning and tool-use)



## REACT (Reason and Act)



## DERA (Dialog-Enabled Resolving Agents)

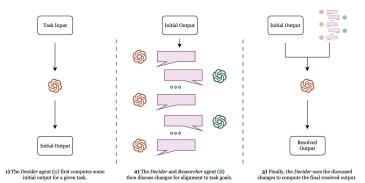
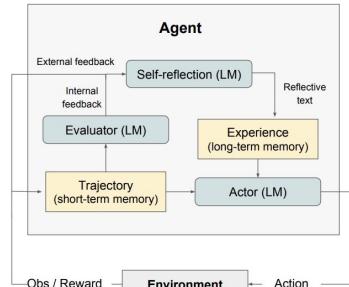


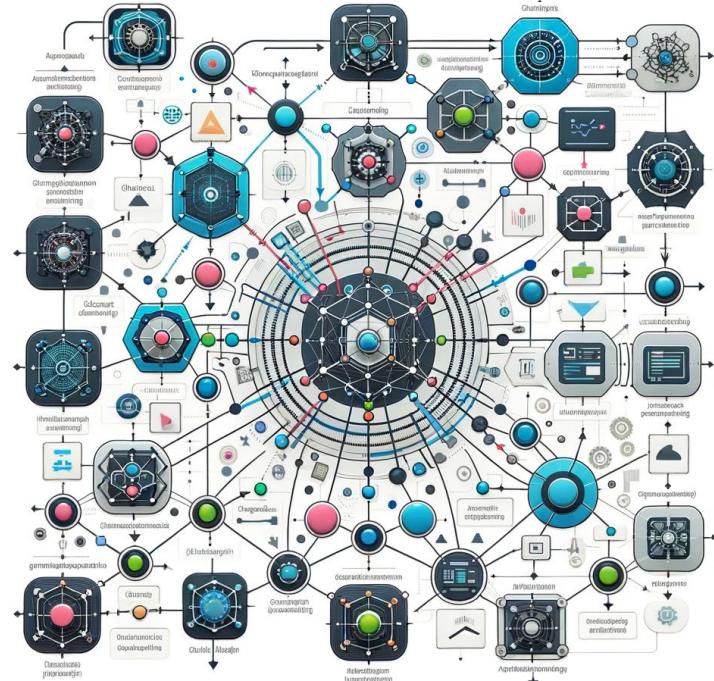
Figure 1: (1) Comparison of 4 prompting methods: (a) Standard, (b) Chain-of-thought (CoT), Reason Only, (c) Act-only, and (d) React (Reason+Act), solving a HopoQA (Vana et al., 2018) question. (2) Comparison of (a) Act-only and (b) REACT prompting to solve an AIWorld (Shafir et al., 2020b) game. In both domains, we omit in-context examples in the prompt, and only show task solving trajectories generated by the model (Act, Thought) and the environment (Obs).

## Reflection



# Multi-agent systems and AGI

- AGI is a misnomer commonly used to talk about human-level intelligence
- Human level intelligence and beyond is already being accomplished by combining specialized agents
- Specialized multi-agent systems increase our chances of taking all the upsides of AI while minimizing the risks!



# Conclusions

**GatesNotes** THE BLOG OF BILL GATES [LOG IN](#)

---

A NEW ERA

## The Age of AI has begun

Artificial intelligence is as revolutionary as mobile phones and the Internet.

By Bill Gates | March 21, 2023 • 14 minute read



[!\[\]\(23b3898f3cb6a27337a39265e4d1750a\_img.jpg\)](#) [!\[\]\(d09d5ba786eda48d77eb00e19fd6366b\_img.jpg\)](#) [!\[\]\(b1ea00f539f58b9ccee429dee091aabc\_img.jpg\)](#) [!\[\]\(9f932a8f74c52f71073cbf9d6ea85412\_img.jpg\)](#)

Q&A

