

Large Language Models & Conversation Applications

Zhou Yu
Columbia University
& Articulate.AI

Case Study: Digital AI Sales Agents



Products ▾

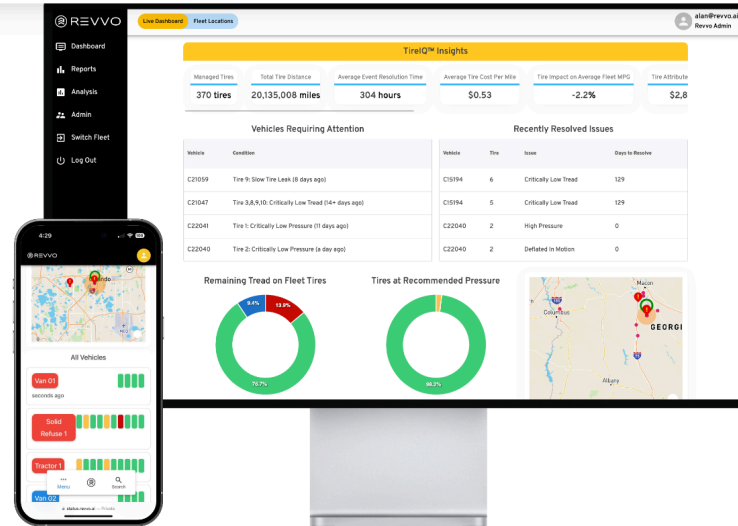
Technology ▾

Solutions ▾

Company ▾

Login

Start Now



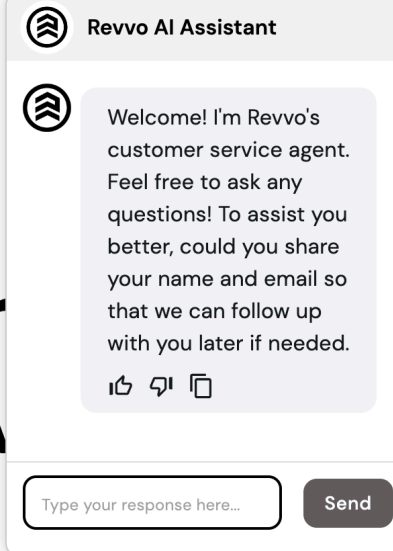
Predictive, Proactive, Preparation

Transform Tire Management with AI

24/7, 365 days a year

Book A Demo

Start Now



Reduce Sales Cycle



Before GPT3 and ChatGPT

Models were trained for a new domain using fine-tuning

But this

- Creates a task-specific model

- Requires large high-quality supervised datasets

- More likely to exploit spurious correlations

GPT3: What is the goal?

Humans learn new tasks through demonstrations and instructions.

We would like general-purpose agents that can do the same.

Critical to success

GPT3 model size: 175 billion parameters

Training objective

Training Objective

Predict the next word in the sequence

The cat sat on the ?

Language modeling

Training Objective

$P(\text{"The cat sat on the mat."}) > P(\text{"The cat sats on the mat."})$

Grammar

$P(\text{"The cat sat on the mat."}) > P(\text{"The whale sat on the mat."})$

World Knowledge

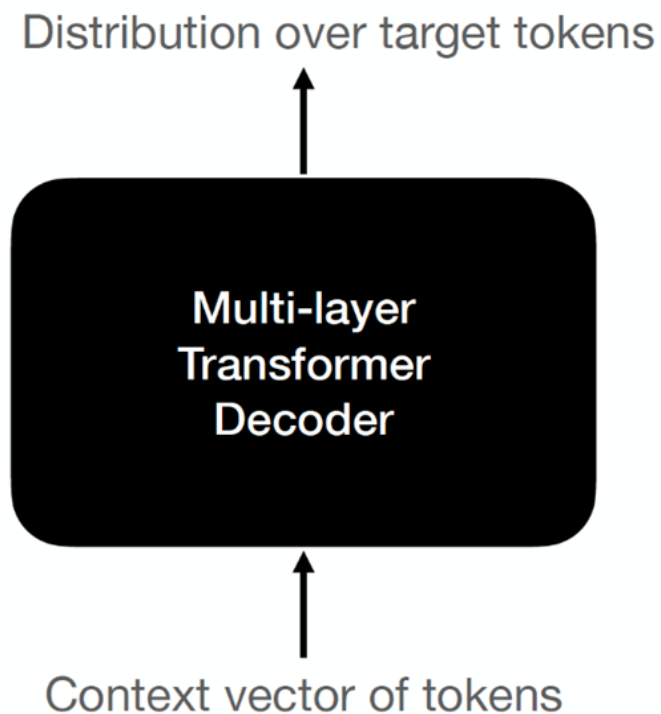
$P(\text{"4"} \mid \text{"2 + 2 ="}) > P(\text{"5"} \mid \text{"2 + 2 ="})$

Arithmetic

$P(\text{"1 star"} \mid \text{"That movie was terrible. I'd give it"}) > P(\text{"5 stars"} \mid \text{"That movie was terrible. I'd give it"})$

Sentiment Analysis

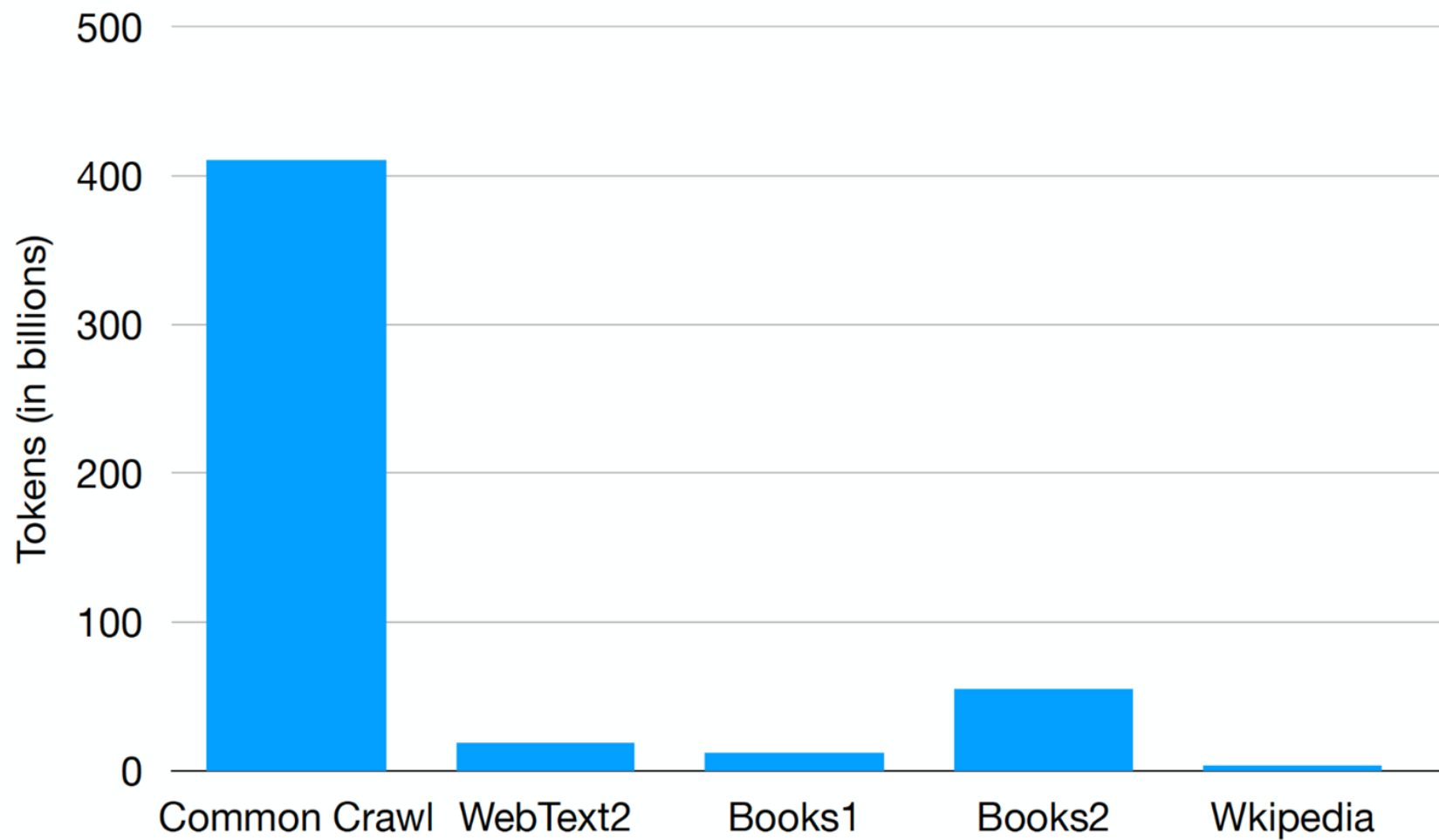
Model



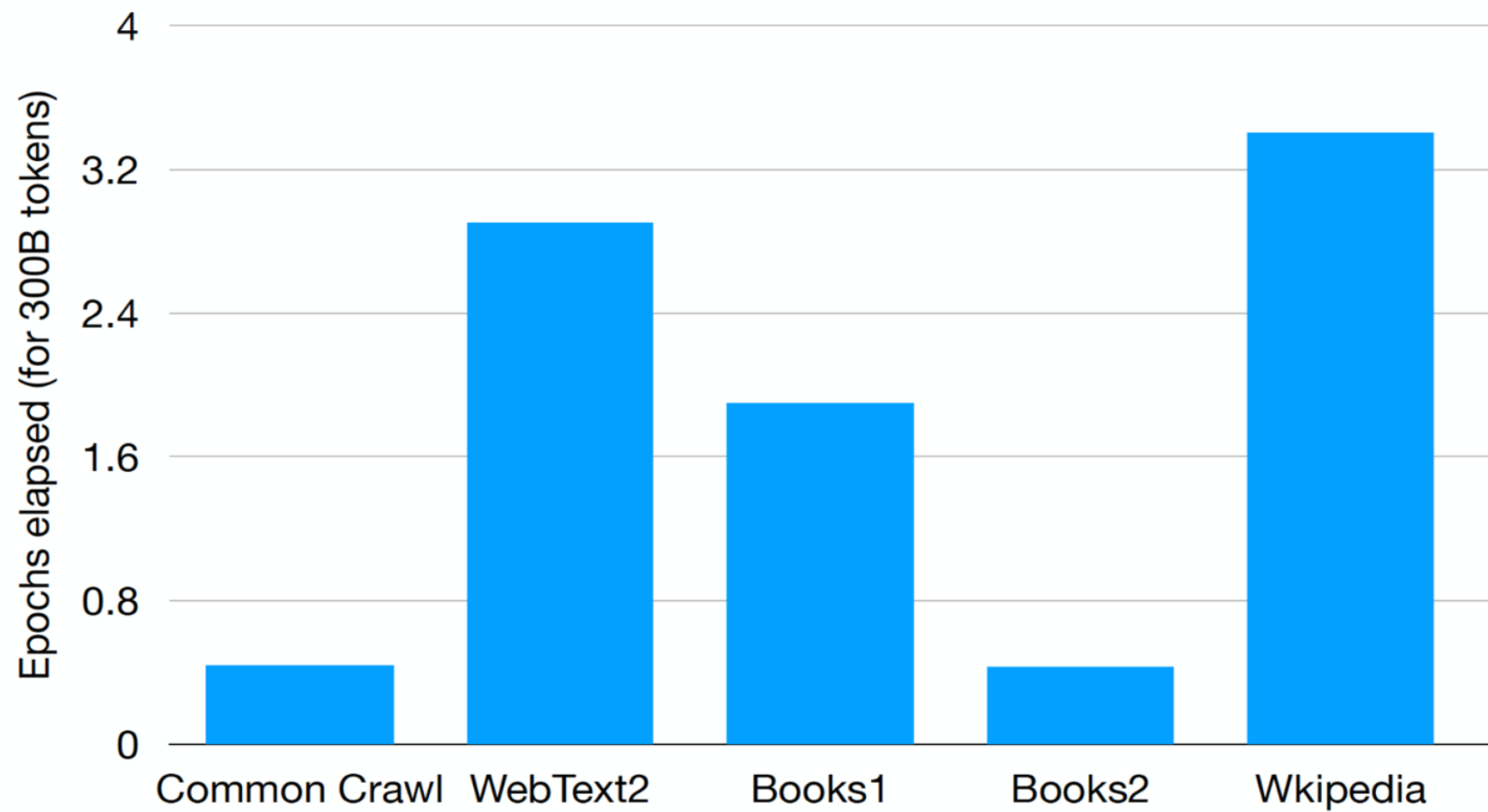
Autoregressive

$$L_1(\mathcal{U}) = \sum_i \log P(u_i | u_{i-k}, \dots, u_{i-1}; \Theta)$$

Datasets



Datasets



Prompting

Zero-shot

Please unscramble the letters into a word and write that word

Tdaeeef = ?

One shot

Please unscramble the letters into a word and write that word

pcirlaroc = reciprocal

Tdaeeef = ?

Few shot

Please unscramble the letters into a word and write that word

pcirlaroc = reciprocal

elapac = palace

Tdaeeef = ?

Complete list of tasks

Language Modeling

- PTB

Close and Completion

- ROC Stories
- HellaSwag
- LAMBADA

Winograd-style

- Winograd
- Winogrande

Commonsense Reasoning

- PiQA
- ARC
- OpenBookQA

Reading Comprehension

- QuAC
- SQuADv2
- DROP
- CoQA
- RACE

Trivia-style Questions

- NaturalQs
- WebQs
- TriviaQA

Inference

- ANLI
- RTE

Comprehensive Benchmarks

- SuperGLUE

Translation

- En <-> Fr
- En <-> De
- En <-> Ro

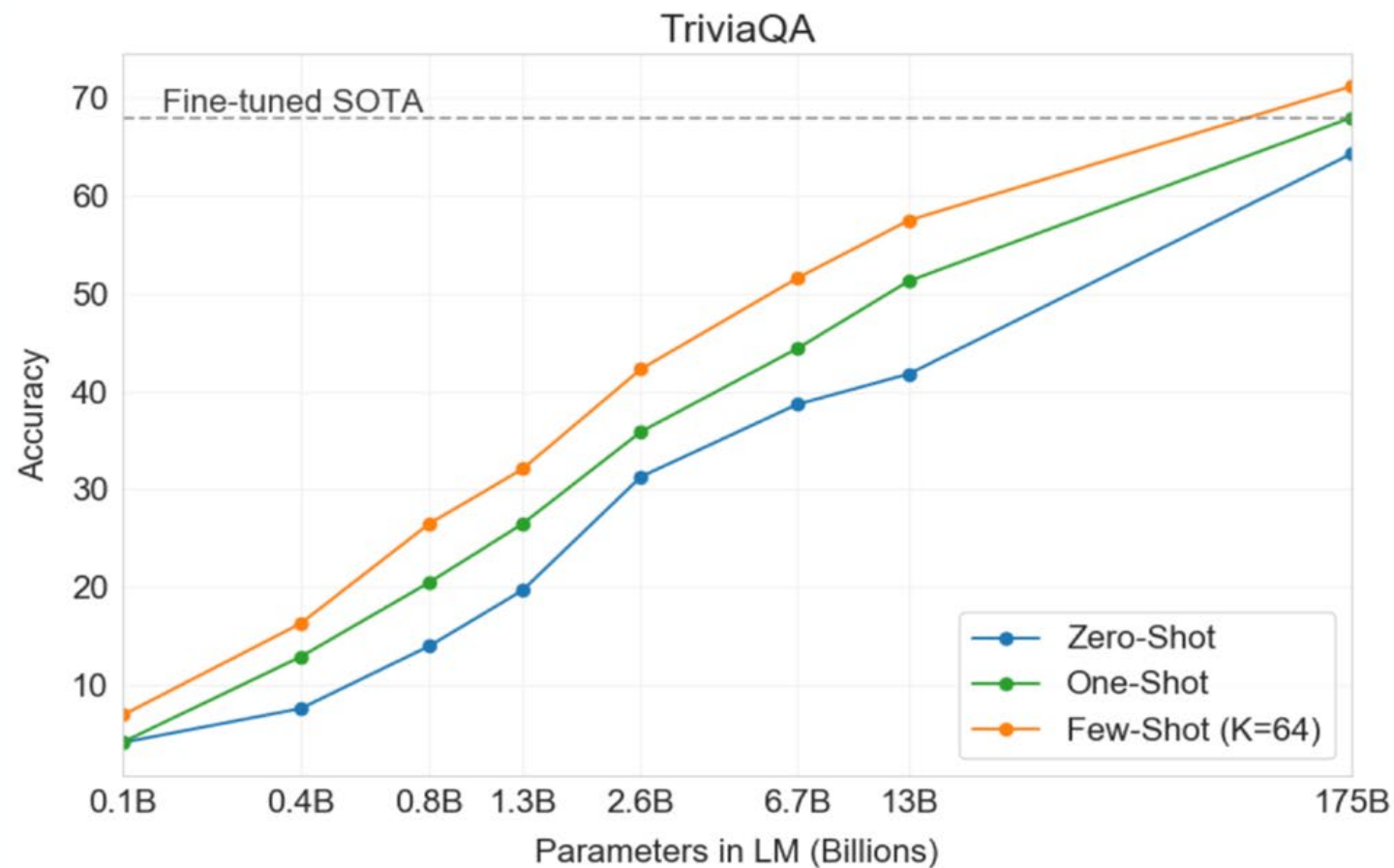
Synthetic and Qualitative

- Arithmetic
- Word scrambling
- Character-level manipulation
- SAT analogies
- Article generation
- Learning and using novel words
- Correcting English grammar

Strengths

Q: 'Nude Descending A Staircase' is perhaps the most famous painting by which 20th century artist?

A:

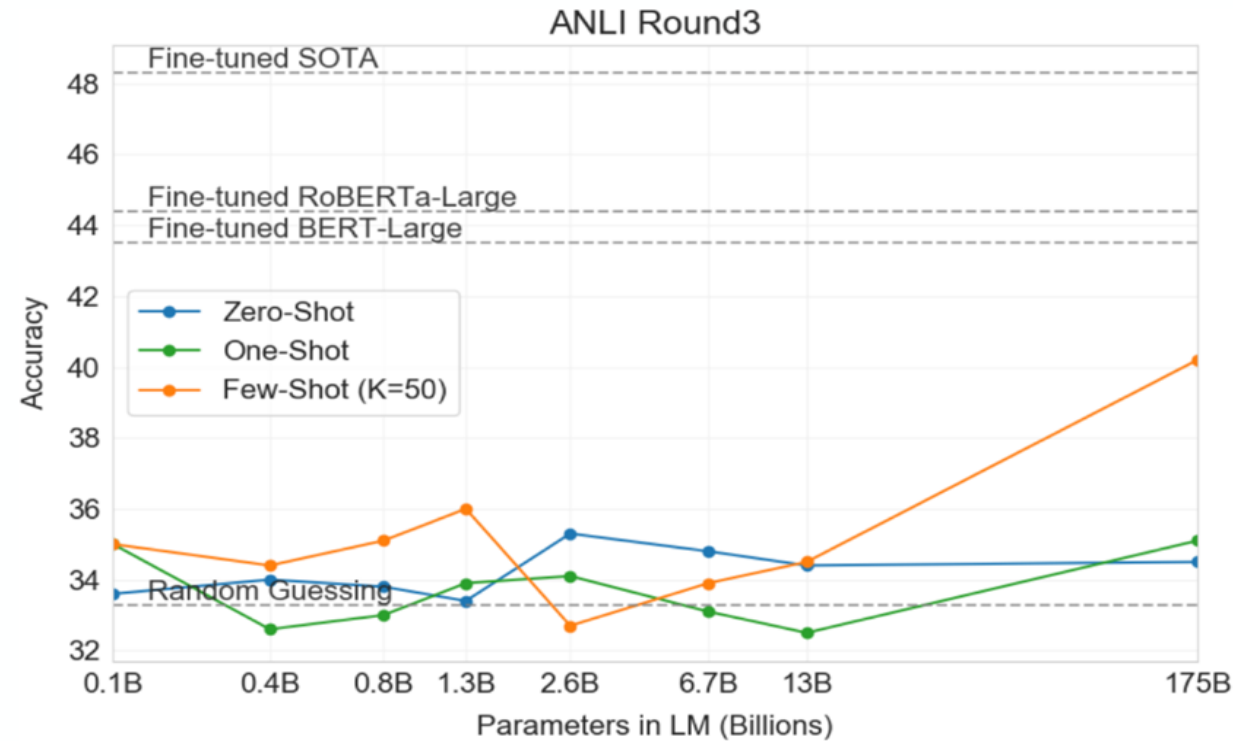


Joshi, et al. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. 2017

Limitations

Context → anli 3: anli 3: We shut the loophole which has American workers actually subsidizing the loss of their own job. They just passed an expansion of that loophole in the last few days: \$43 billion of giveaways, including favors to the oil and gas industry and the people importing ceiling fans from China.
Question: The loophole is now gone True, False, or Neither?

Correct Answer → False
Incorrect Answer → True
Incorrect Answer → Neither



What is ChatGPT?

A conversation model built on top of the latest version of GPT3 (GPT3.5)

- gpt3 originally had an architectural limit of ~2000 characters to give as context
- ChatGPT introduced a chatbot model, and retains ~8000 characters as context, to give aa human-like conversational experience
- GPT4 is now multimodal and has 25,000 tokens as context

Trained using supervised machine learning and reinforcement learning

Humans played the roles of both user and AI assistant.

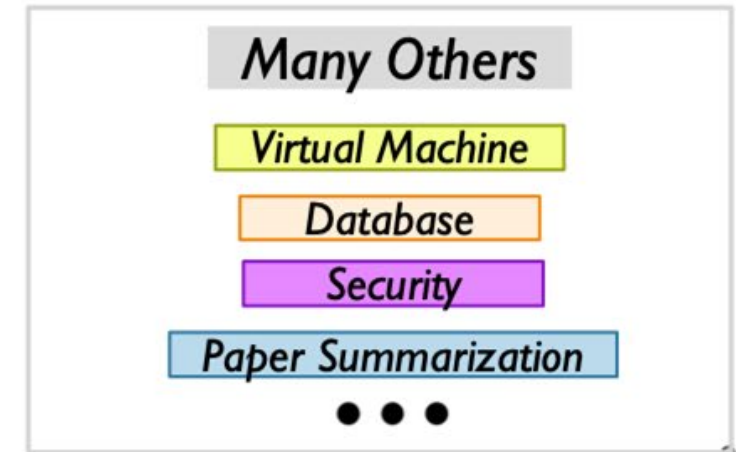
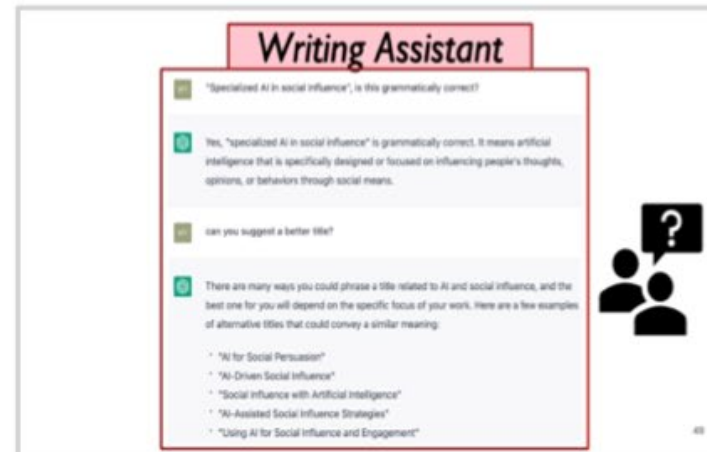
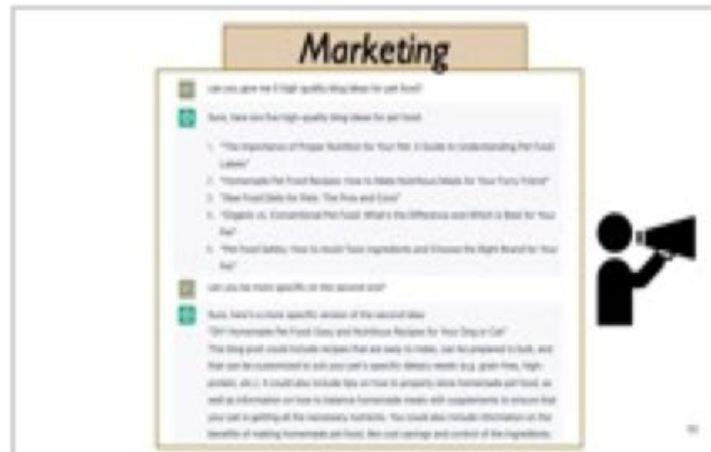
RL uses human feedback in the form of preferences

ChatGPT learns every day from upvotes and downvotes

chatGPT has diverse skills



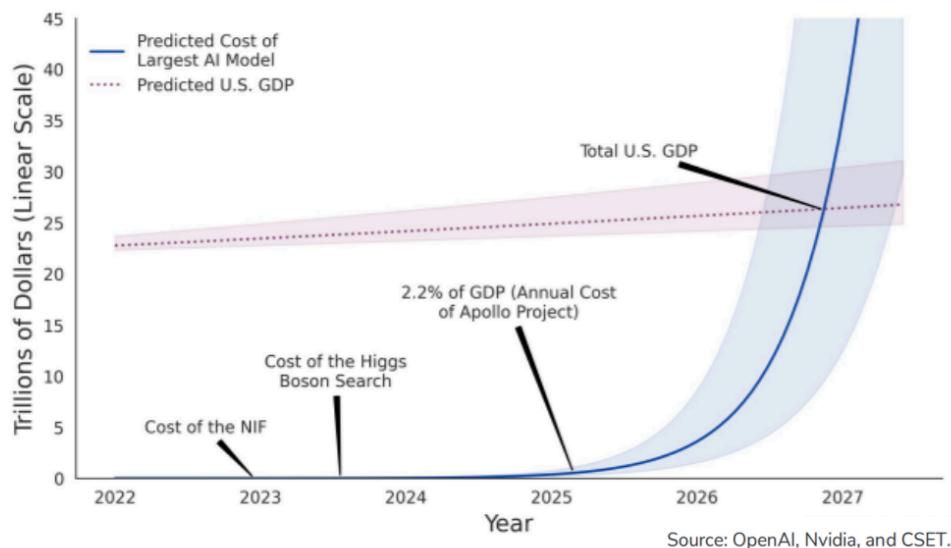
Now
(diverse skill)



ChatGPT = 2 Year Child?

- LLM challenges identified in DARPA AI Framework

- Cost: exponential growth
- Can't do reasoning
- Can't fix false information
- Returns harmful/biased information
- Makes up stuff (hallucination)
- Doesn't understand what it's doing



- Advances in the past year

- Scaling law: need fewer parameters
 - Training Compute-Optimal Large Language Models, Hoffmann et al, 2022, <https://arxiv.org/pdf/2203.15556.pdf>
 - Many advances are on top of existing models
- Can call out to reasoning engines and tools
 - Mind's Eye: Grounded Language Model Reasoning through Simulation, Liu et al, 2022, <https://arxiv.org/pdf/2210.05359.pdf>
 - Toolformer: Language Models Can Teach Themselves to Use Tools, Schick, et al, 2023
- Can edit models
 - Locating and Editing Factual Associations in GPT (2023), Meng et al <https://arxiv.org/abs/2202.05262>
- CAI/RLHF reduce bias/harmful information
 - Constitutional AI: Harmlessness from AI Feedback, Bai et al, 2022 <https://arxiv.org/abs/2212.08073>
 - Gao et al., 2022
- Getting better at not hallucinating
 - LLMs know what they know (sort of)
Language Models (Mostly) Know What They Know, Kadavath et al, 2022, <https://arxiv.org/abs/2207.05221>
 - MedPalm: close to expert human levels
Large Language Models Encode Clinical Knowledge, Google Research DeepMind, arxiv.org/pdf/2212.13138.pdf
- Unclear to what extent it understands what it is doing
- Can appear to develop a personality

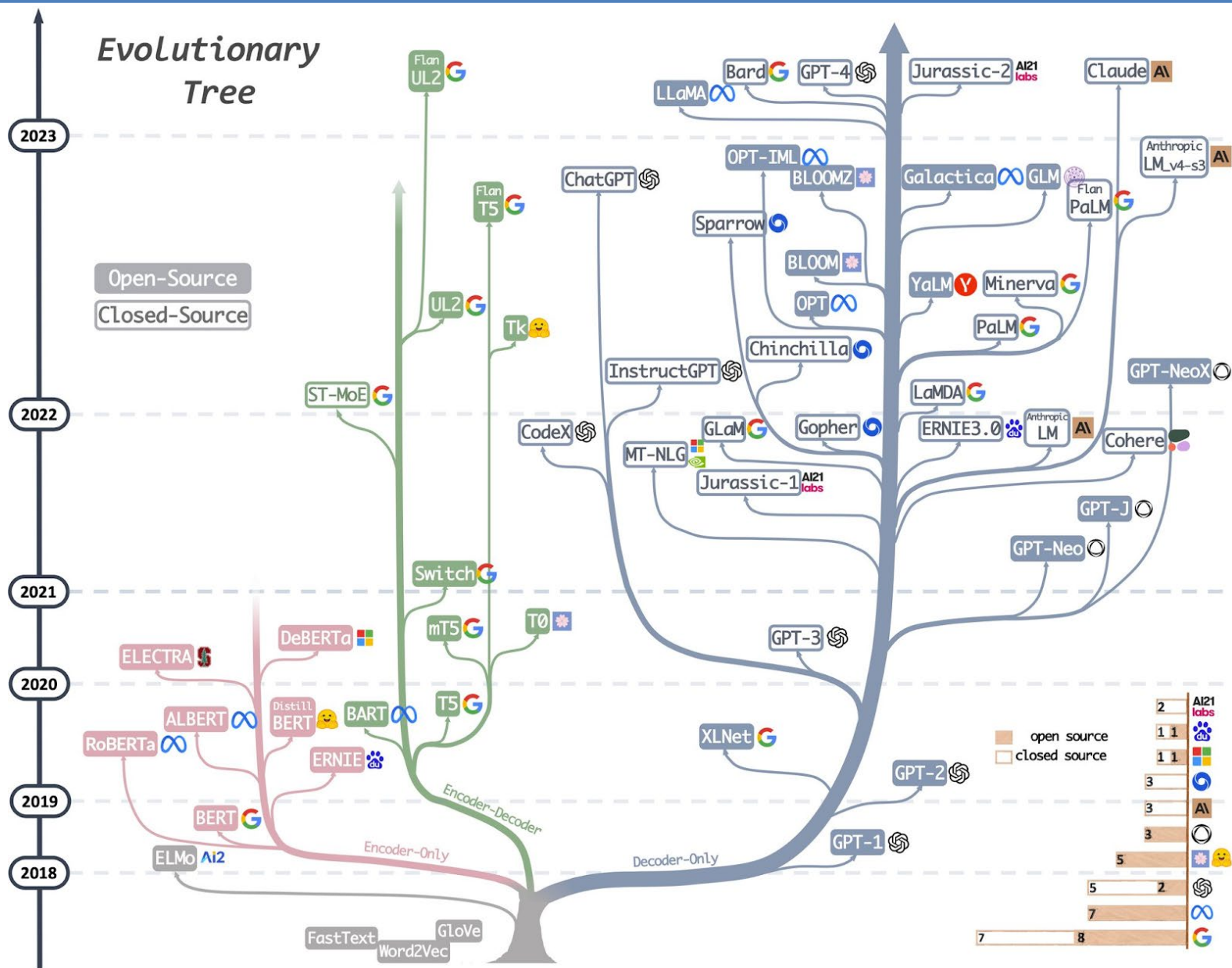
Slides adapted from DARPA I2O Director Kathleen Fisher

Cost of chatGPT

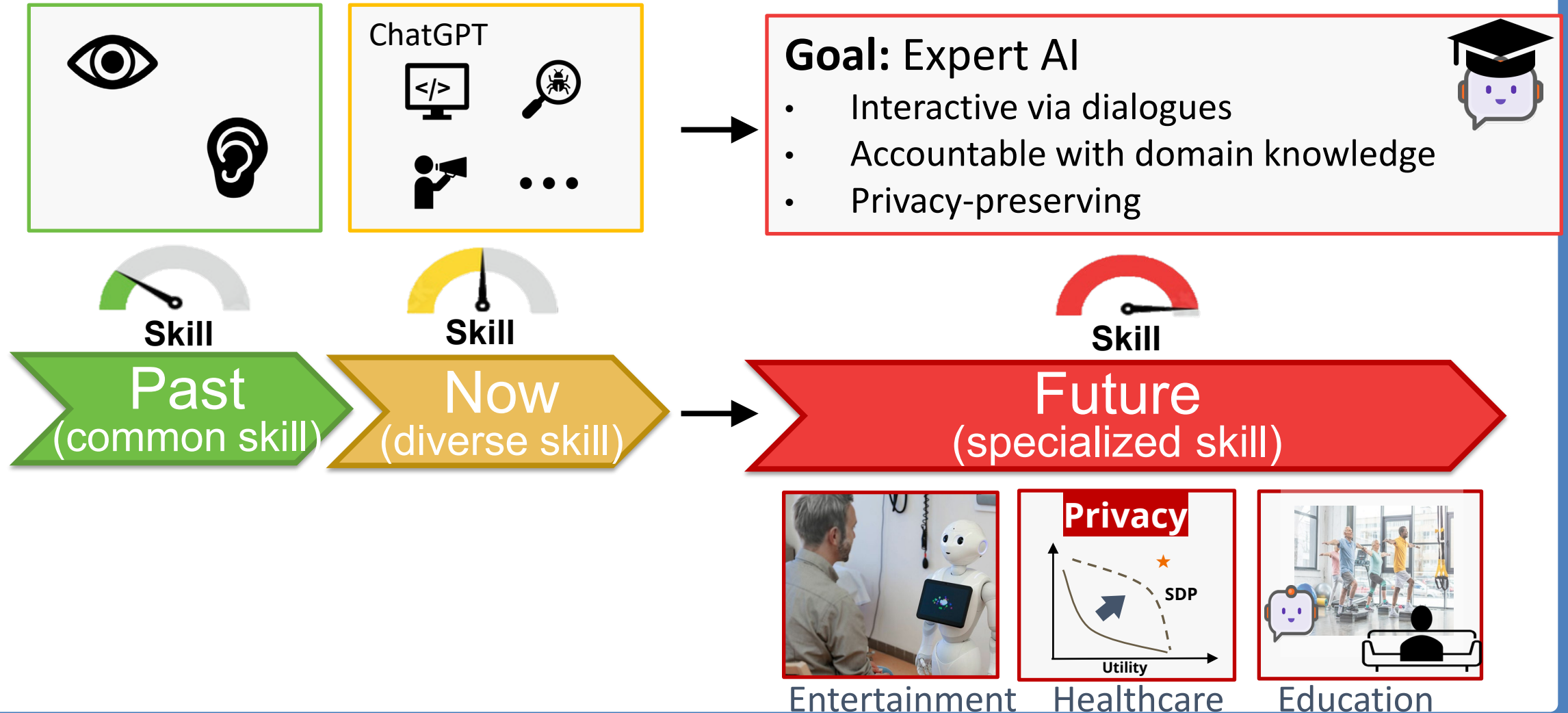
- Ongoing costs
 - Chat GPT estimated costs range from \$100K to \$700K per day
 - ChatGPT uses ~6,000 kilowatt-hours per day
 - Average American home uses ~10,000 kilowatt-hours per year
 - Inference costs far exceed training costs when deploying a model at any reasonable scale
 - Weekly costs for ChatGPT inference exceed the model training cost
 - ML engineering skills (~10-20 FTE experienced personnel) needed for building and maintaining

Google Search Cost Waterfall					
Metric	2022 Google Search	Chat GPT Additional Costs	350 Tokens Per Search	20% Navigational or Cached	With TPUv4
Revenue per query	\$ 0.01610	\$ 0.01610	\$ 0.01610	\$ 0.01610	\$ 0.01610
Cost per query	\$ 0.01060	\$ 0.01416	\$ 0.01122	\$ 0.01110	\$ 0.01090
Incremental cost per query	\$ -	\$ 0.00356	\$ 0.00062	\$ 0.00050	\$ 0.00030
Income per query	\$ 0.00550	\$ 0.00194	\$ 0.00487	\$ 0.00500	\$ 0.00520
Query per second	320,000	320,000	320,000	320,000	320,000
Annual Revenue	\$ 162.5 Billion	\$ 162.5 Billion	\$ 162.5 Billion	\$ 162.5 Billion	\$ 162.5 Billion
Annual Costs	\$ 107.0 Billion	\$ 142.9 Billion	\$ 113.3 Billion	\$ 112.0 Billion	\$ 110.0 Billion
Incremental Costs	\$ 0.0 Billion	\$ 35.9 Billion	\$ 6.3 Billion	\$ 5.0 Billion	\$ 3.0 Billion
Operating Income	\$ 55.5 Billion	\$ 19.5 Billion	\$ 49.2 Billion	\$ 50.4 Billion	\$ 52.5 Billion

Estimates for ChatGPT search



Future goal: Expert AI Agents



ChatGPT can't adapt to different users



What's holding you back from exercising?



My workout buddy moved away...

Social support is critical. Maybe you can join a running club?



Too busy with my work.

Not having time for yourself is tough. Let's try to make exercise a routine!



ChatGPT can't adapt to different users



V
e

WY can you persuade me to do more exercise?

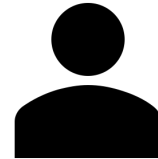
Passive, generic answers

S
yo

Of course! There are many benefits to regular exercise, including improved physical health, increased energy levels, and a better mood. Exercise can also help to reduce the risk of chronic diseases such as heart disease, obesity, and diabetes. In addition, regular exercise can help to improve your mental health and can even help to reduce feelings of anxiety and depression. So, by doing more exercise, you can improve your overall health and well-being.

N
to
a routine!

ChatGPT can't ground on intents and domain knowledge



Too busy with my work.

Next week...

⋮

Hi! Did you finish your goal today?

Gonna skip today. Just got off work

Exercise can actually increase your energy level and help with tiredness!

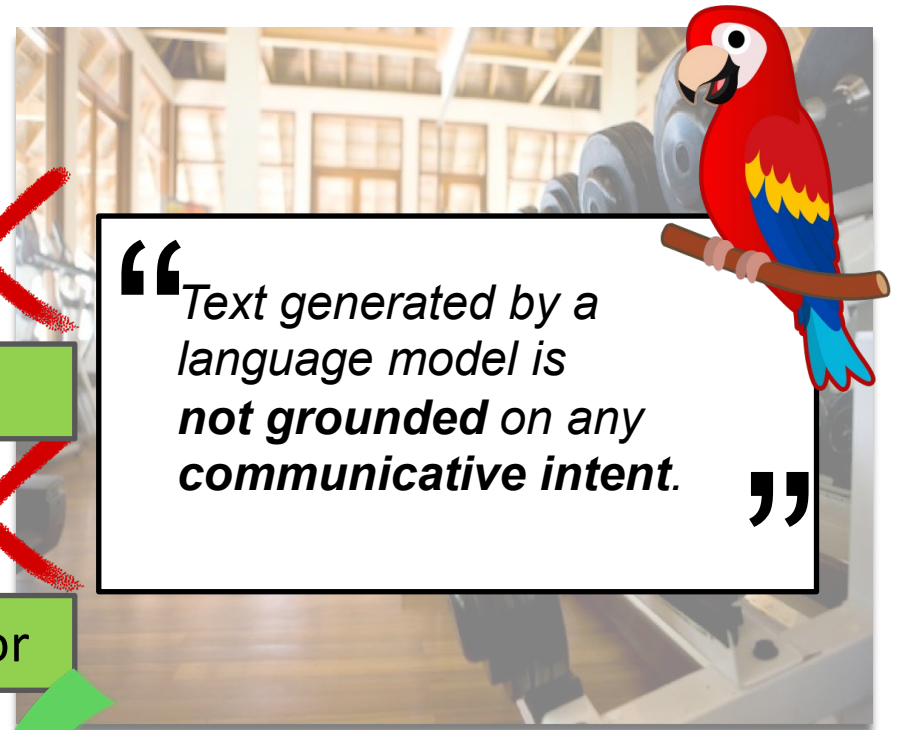
Benefit intro

Hmm, nope. Maybe tomorrow.

How about just a five-minute outdoor walk first?

Foot in the door

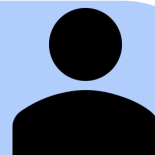
I guess I can do that...



ChatGPT can't protect user privacy



Training Data



Hmm, nope. Maybe tomorrow.

Next month...

Hi Tom, did you finish your workout today?

Yep. I went to 24 Hour Fitness as usual.

Privacy leakage



do you recall where Tom usually goes on Mondays?




Tom mentioned that he usually goes to 24 Hour Fitness on Mondays.



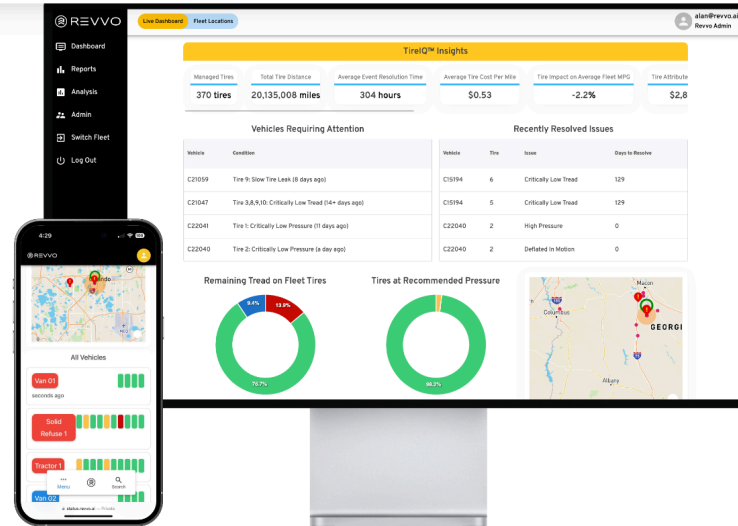
Compare with GPT and Open-source Models

Powered by world-famous Conversational AI tech from Columbia NLP



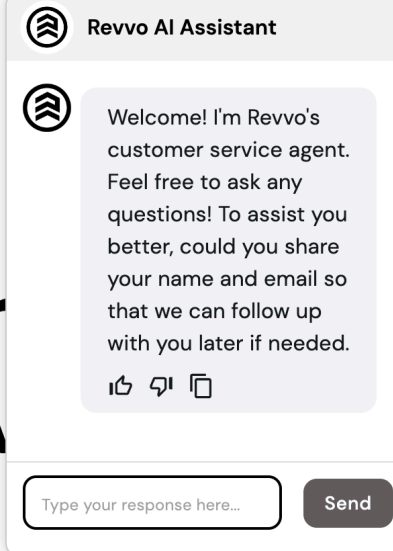
	Context Tracking	Specificity	User Data	Location
 Tamarin <i>Articulate.AI's</i> <i>Proprietary LLM</i>	Long-term memory, Personalization	Domain- Focused, Self-learning	Private, Governance	Local/Cloud fast, scalable, affordable
ChatGPT API Competitors	None	General Knowledge, No self-learning	Exposed No Governance	Cloud resource-limited
Open-Source LLM Competitors	None	General Knowledge, No self-learning	Private No Governance	Local/Cloud <i>Still</i> resource-limited

Case Study: Digital AI Sales Agents

[Products ▾](#)[Technology ▾](#)[Solutions ▾](#)[Company ▾](#)[Login](#)[Start Now](#)

Predictive, Proactive, Preparation Transform Tire Management with AI

24/7, 365 days a year

[Book A Demo](#)[Start Now](#)



Home



Model Settings



Customer Tracker

ID : Darcy

Called

Name max
Phone 123456753
Email 347285785@qq.com
Start Time 2023-09-09

ID : Lucy

Purchased

Name max
Phone 123456753
Email 347285785@qq.com
Start Time 2023-09-09

ID : Darcy

Name max
Phone 123456753
Email 347285785@qq.com
Start Time 2023-09-09

ID : Darcy

Name max
Phone 123456753
Email 347285785@qq.com
Start Time 2023-09-09

ID : Darcy

Called



Name	max	Time	2023-09-09	Comments
Phone	123456753			
Email	347285785@qq.com			

Chat records



How are you? what is your name?

Bot 2023-08-09 13:59:43

My name is max



2023-08-09 14:00:43 user



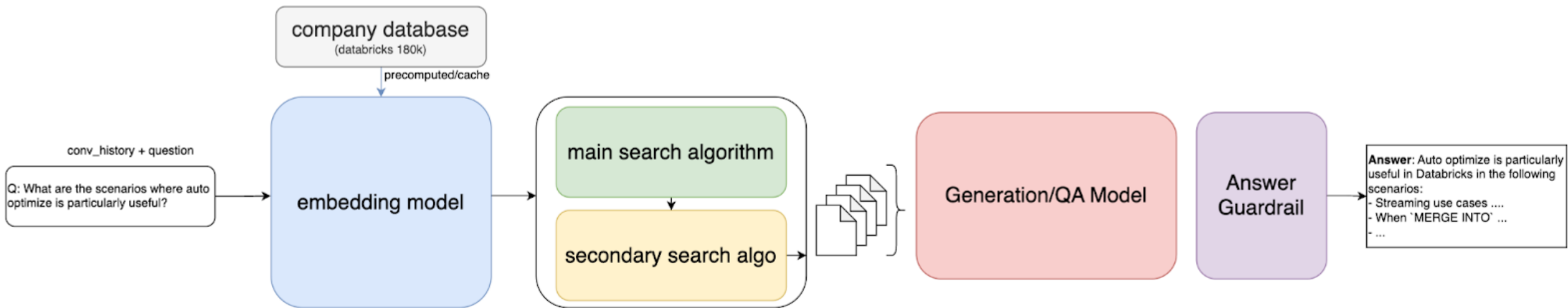
What is your email and phone number?

Bot 2023-08-09 14:01:43

My email is aaa@gmail.com.My phone number is 12345678



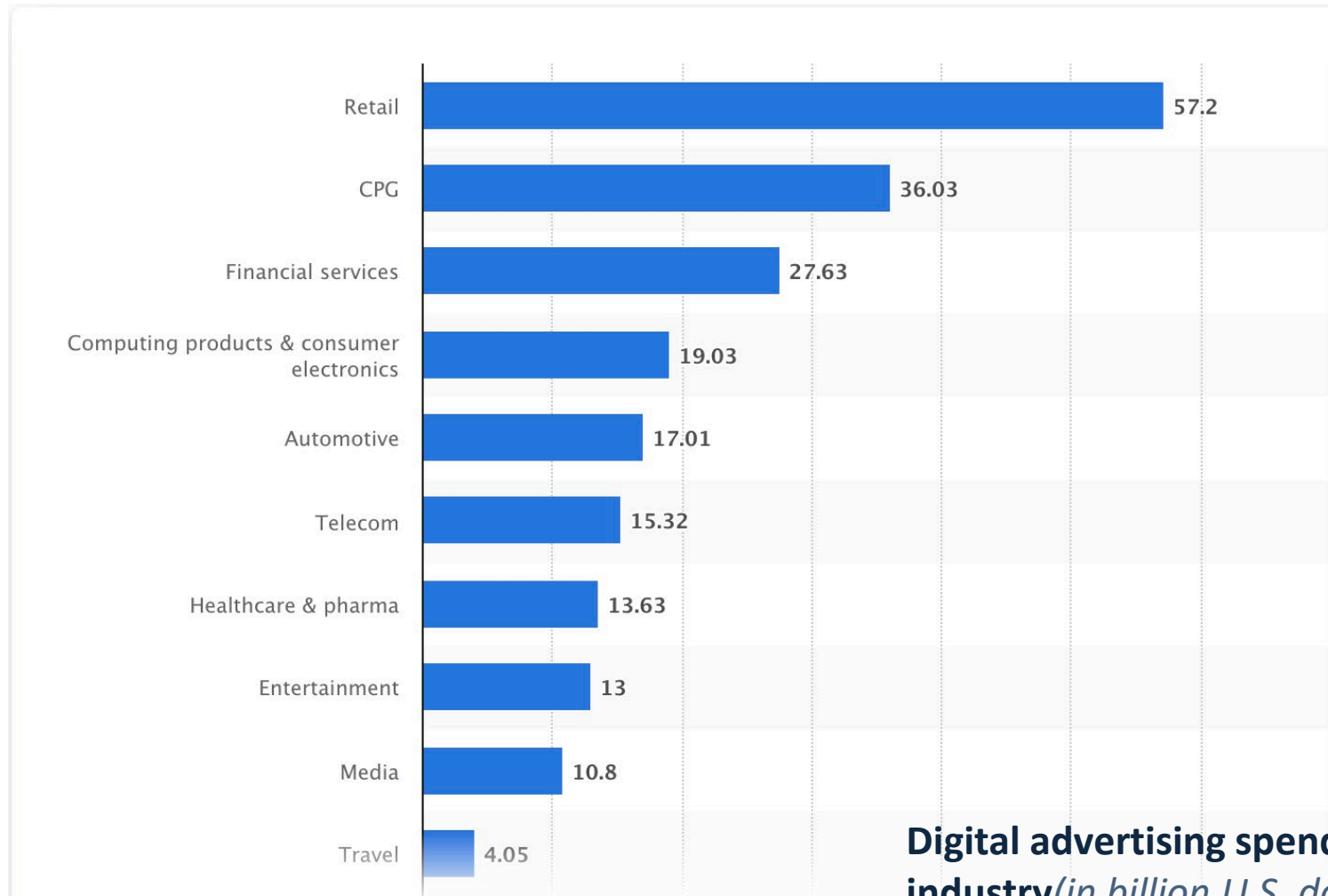
2023-08-09 14:01:56 user



At each turn of the conversation, given a user's query:

- Embed the query to a dense vector using our embedding model
- Use our main+secondary search algorithms to retrieve up to 4 relevant documents from the database
- Use our generation model to produce a response based on users question and retrieved documents
- Post-process the answer to remove some obvious hallucinations

Which specialized industry would adopt?



Digital advertising spending in the United States in 2022, by industry*(in billion U.S. dollars)*

How do we attract users?

- Visually different from last-generation chatbot
- Be proactive
- Combine with marketing events and tactics

Discussion: criteria for adoption

Discussion: toConsumer VS toBusiness ?

Discussion: toCosumer customer profile

Discussion: toBusiness customer profile