FINAL CONFERENCE

THE FUTURE OF AGEING

Embracing Technology for a Fulfilling Life

7th - 8th March 2024 | Évry (South-East Paris) | France | Télécom Sud Paris

EU-JAPAN VIRTUAL COACH FOR SMART AGEING

DR. JAN ALEXANDERSSON
Deutsches Forschungszentrum für
Künstliche Intelligenz

CHAIR 🇩🇪

EMERITUS PROF. MICHAEL MCTEAR
Ulster University

KEYNOTE SPEECH 🇬🇧

3 • Empowering Older Adults: Unravelling the Potential of Large Language Models in Wellbeing Applications

EU-JAPAN VIRTUAL COACH FOR SMART AGEING

Final Conference
7th – 8th March 2024

Empowering Older Adults: Unravelling the Potential of Large Language Models in Wellbeing Applications

Emeritus Prof. Michael McTear

Ulster University

The project has received funding from the European Union H2020 Programme under grant agreement no. 101016453.
The Japanese consortium received funding from the Japanese Ministry of Internal Affairs and Communication (MIC), Grant no. JPJ000595.

Promoting active and healthy ageing in older adults in Europe and Japan

- Traditional conversation design (NLU-based)
- LLM-powered conversations
- A quick dive into the technology
- Issues with LLMs
- Recent developments
- Other uses of LLMs in Conversational AI
- Takeaways

NLU-based chatbots

- Create training examples for intents
- Design system prompts and responses
- Create conversation flows (stories)

Problems with NLU-based chatbots

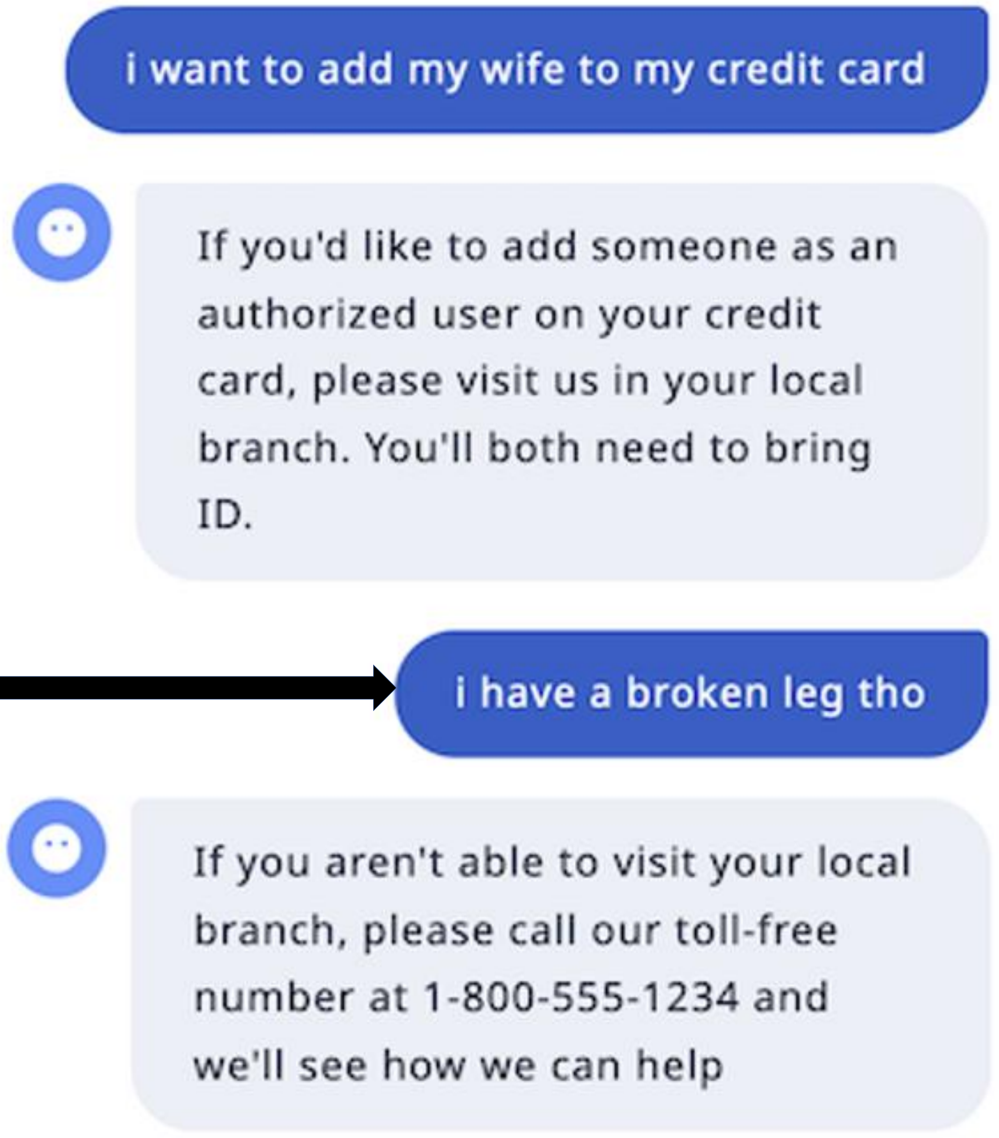
- Missing intents
- Overlapping intents
- Utterances that do not map easily to intents

- Pre-defined system outputs are static and inflexible

- Pre-defined stories are also static and inflexible

Intent classification problem

In an intent-based NLU model this utterance would be classified as out-of-scope



i want to add my wife to my credit card

If you'd like to add someone as an authorized user on your credit card, please visit us in your local branch. You'll both need to bring ID.

i have a broken leg tho

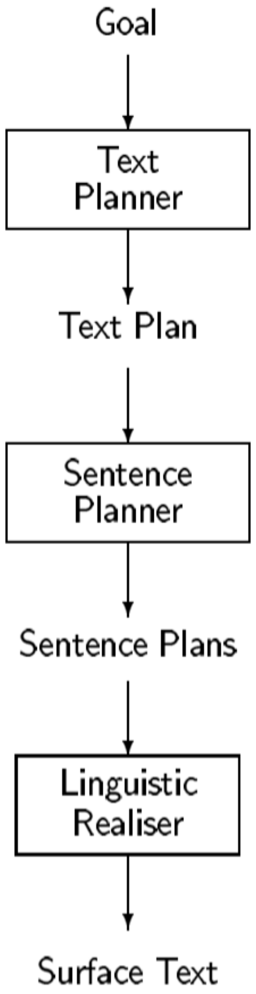
If you aren't able to visit your local branch, please call our toll-free number at 1-800-555-1234 and we'll see how we can help

Alan Nichol Breaking free from intents: a new dialogue model
https://rasa.com/blog/breaking-free-from-intents-a-new-dialogue-model/

Creating System responses

Using canned responses and te
So you want to go to $Destinati

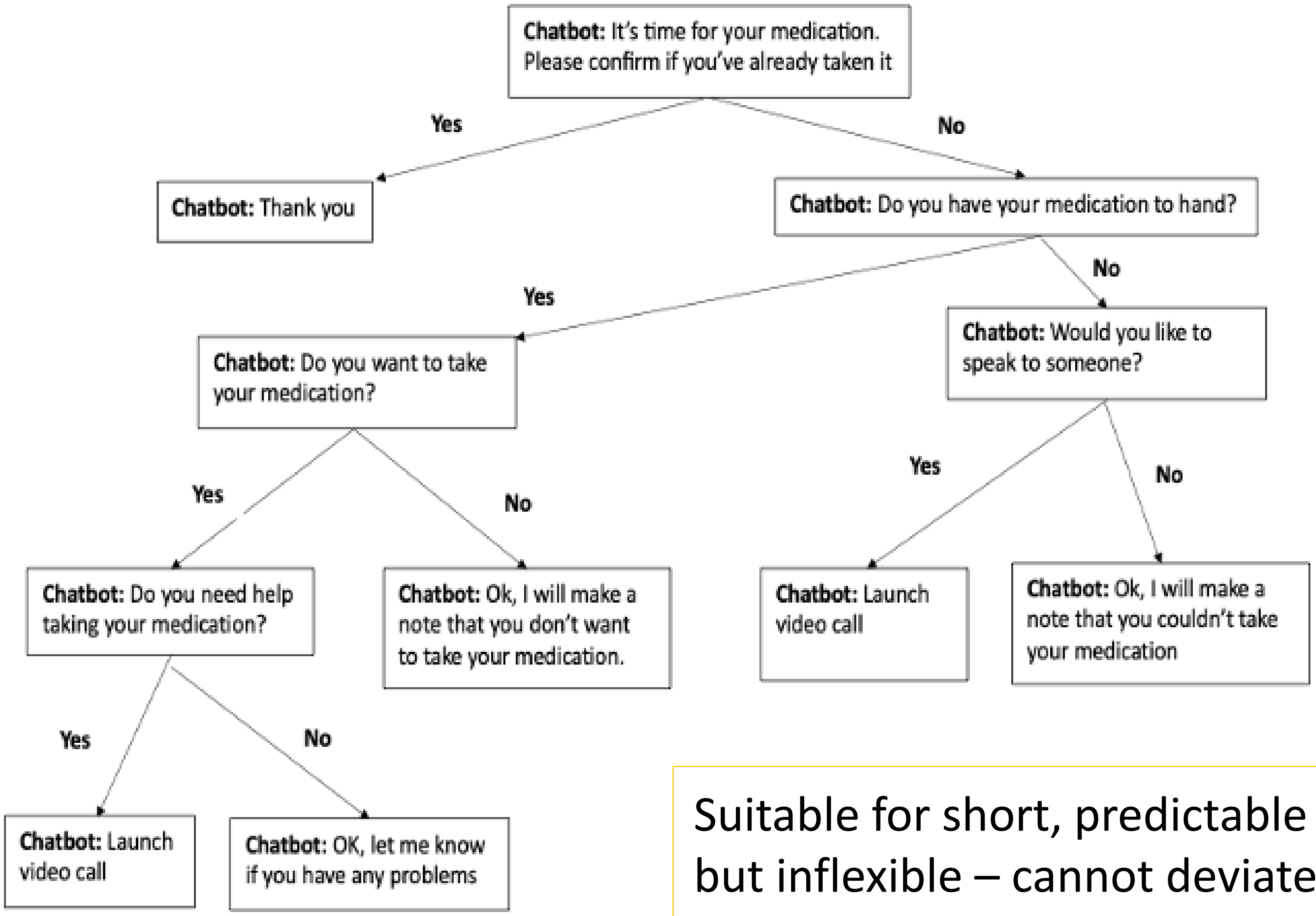
Natural Language Generation
pipeline



End-to-end Using Generative AI

Ehud Reiter and Robert Dale, Building Applied Natural Language Generation Systems. Cambridge University Press, 2000

A simple conversation flow



Chatbot: It's time for your medication. Please confirm if you've already taken it

Yes

Chatbot: Thank you

No

Chatbot: Do you have your medication to hand?

Yes

Chatbot: Do you want to take your medication?

Yes

Chatbot: Do you need help taking your medication?

Yes

Chatbot: Launch video call

No

Chatbot: OK, let me know if you have any problems

No

Chatbot: Ok, I will make a note that you don't want to take your medication.

No

Chatbot: Would you like to speak to someone?

Yes

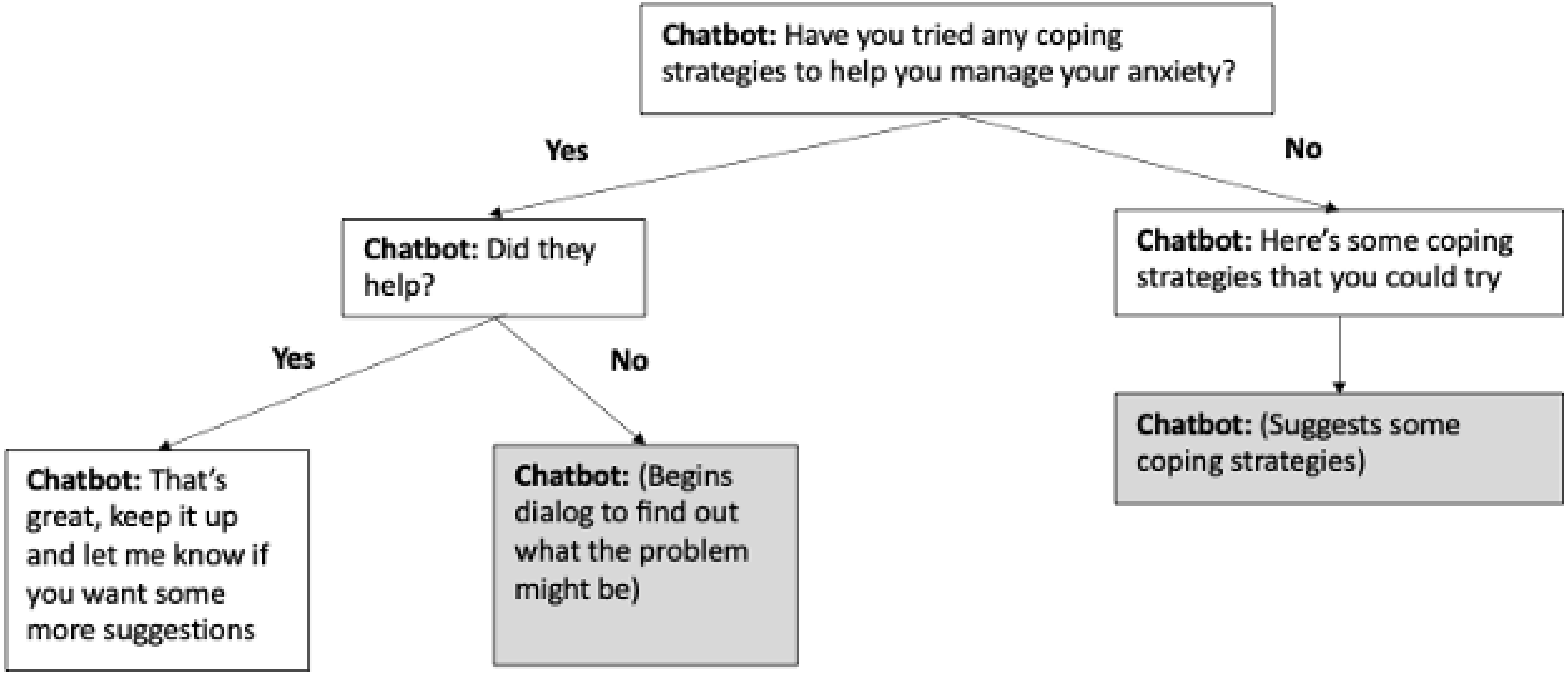
Chatbot: Launch video call

No

Chatbot: Ok, I will make a note that you couldn't take your medication

Suitable for short, predictable interactions but inflexible – cannot deviate from the paths in the graph

A more open-ended conversation flow



Chatbot: Have you tried any coping strategies to help you manage your anxiety?

Yes

Chatbot: Did they help?

Yes

Chatbot: That's great, keep it up and let me know if you want some more suggestions

No

Chatbot: (Begins dialog to find out what the problem might be)

No

Chatbot: Here's some coping strategies that you could try

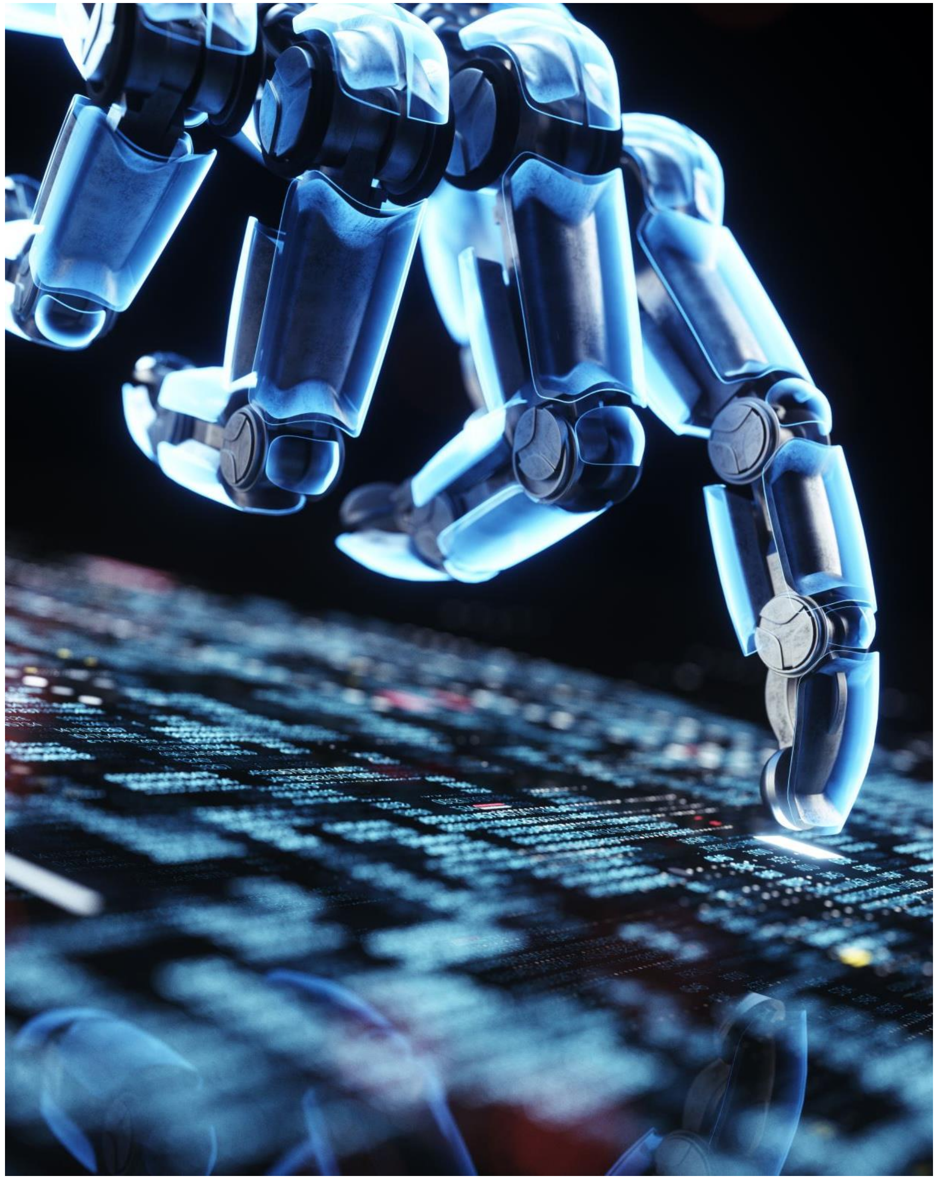
Chatbot: (Suggests some coping strategies)

When the interaction becomes more open-ended, it is difficult to predict all possible paths

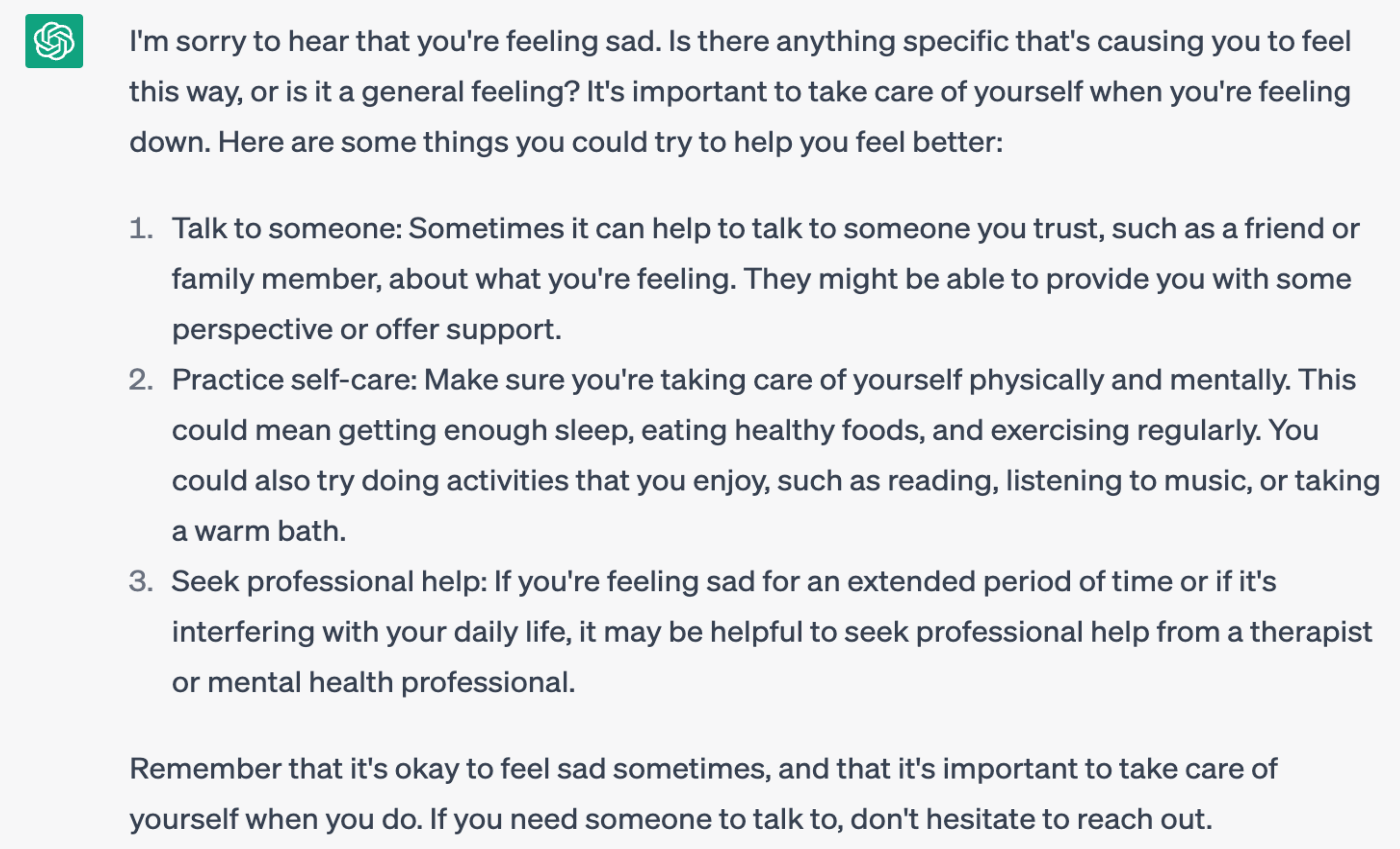
LLM-powered chatbots

An LLM can:

- interpret the user’s inputs
- generate responses dynamically
- continue the conversation dynamically on a turn-by-turn basis



LLM’s response to the input “I am feeling sad”



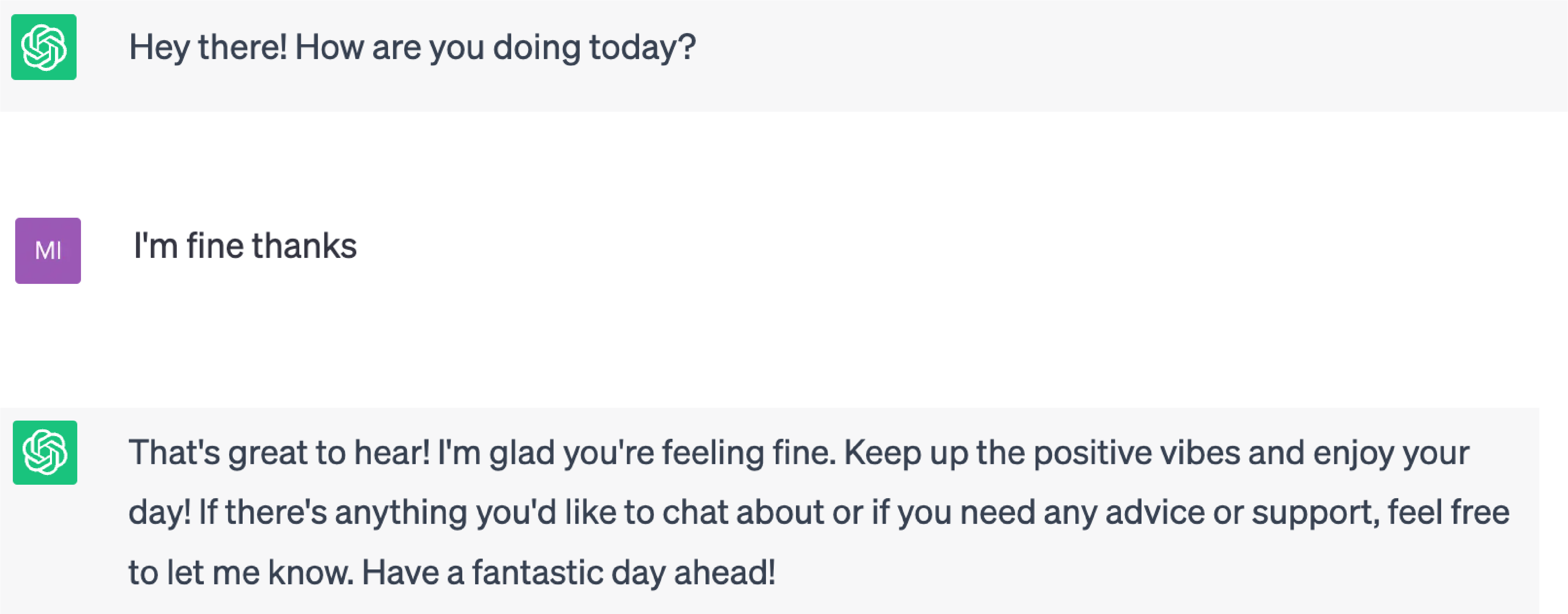
I'm sorry to hear that you're feeling sad. Is there anything specific that's causing you to feel this way, or is it a general feeling? It's important to take care of yourself when you're feeling down. Here are some things you could try to help you feel better:

1. Talk to someone: Sometimes it can help to talk to someone you trust, such as a friend or family member, about what you're feeling. They might be able to provide you with some perspective or offer support.
2. Practice self-care: Make sure you're taking care of yourself physically and mentally. This could mean getting enough sleep, eating healthy foods, and exercising regularly. You could also try doing activities that you enjoy, such as reading, listening to music, or taking a warm bath.
3. Seek professional help: If you're feeling sad for an extended period of time or if it's interfering with your daily life, it may be helpful to seek professional help from a therapist or mental health professional.

Remember that it's okay to feel sad sometimes, and that it's important to take care of yourself when you do. If you need someone to talk to, don't hesitate to reach out.

Creating a mood chatbot

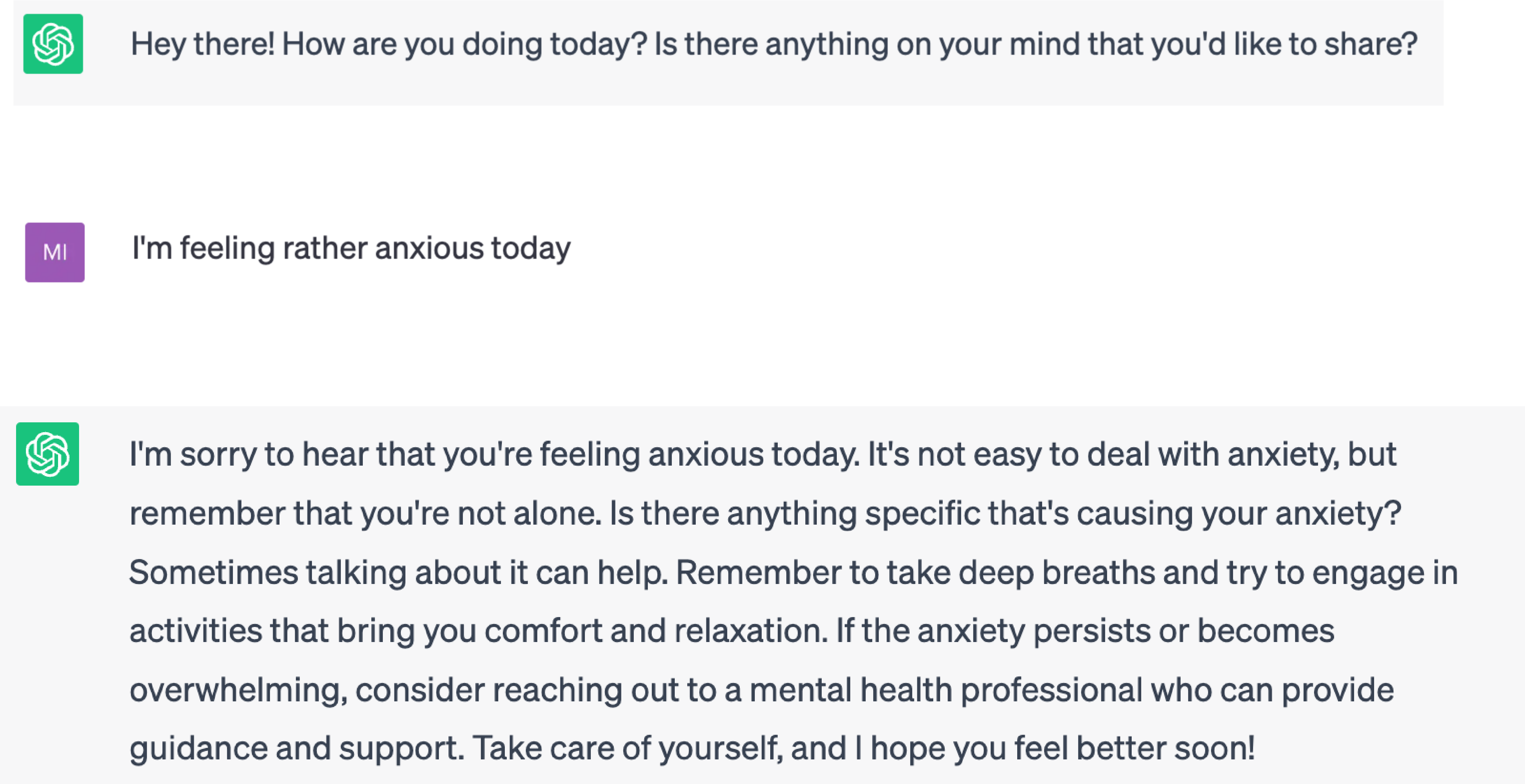
You are MoodBot, an automated service giving advice to a user.
First greet the customer, and then ask how they are.
Let the user respond before continuing.
If the user says something negative, provide an appropriate response to help and then finish.
If the user’s response is not negative, say something cheerful and then finish.
Respond in a short, very conversational friendly style.



Hey there! How are you doing today?

I'm fine thanks

That's great to hear! I'm glad you're feeling fine. Keep up the positive vibes and enjoy your day! If there's anything you'd like to chat about or if you need any advice or support, feel free to let me know. Have a fantastic day ahead!



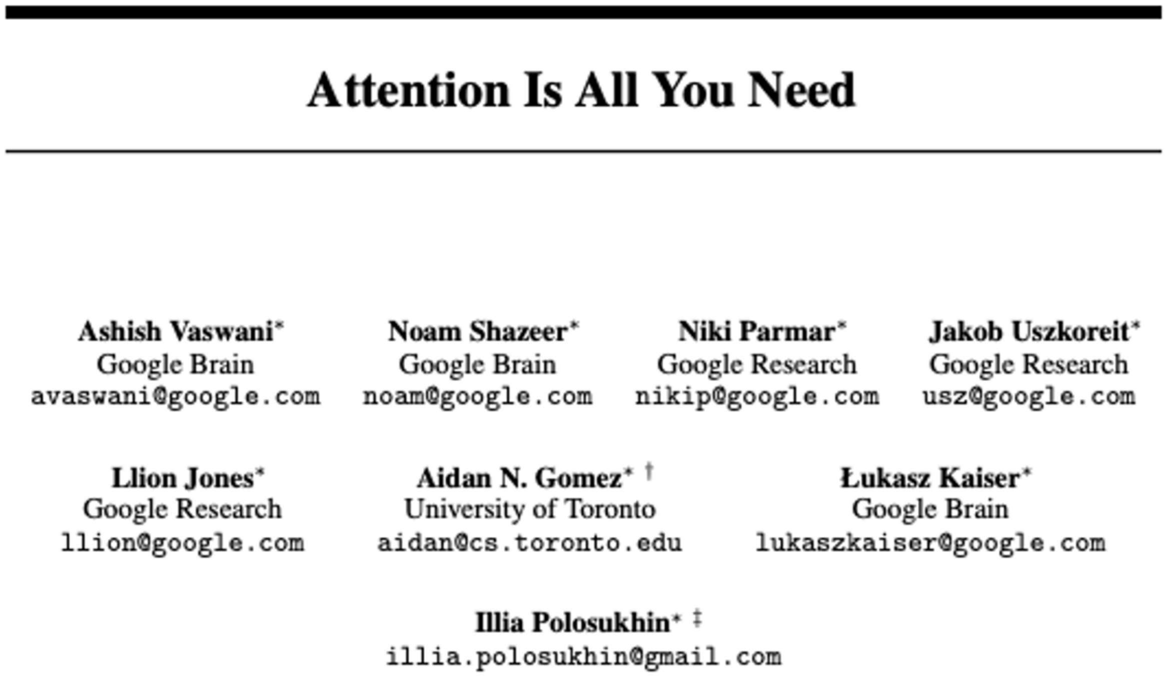
Hey there! How are you doing today? Is there anything on your mind that you'd like to share?

MI

I'm feeling rather anxious today

I'm sorry to hear that you're feeling anxious today. It's not easy to deal with anxiety, but remember that you're not alone. Is there anything specific that's causing your anxiety? Sometimes talking about it can help. Remember to take deep breaths and try to engage in activities that bring you comfort and relaxation. If the anxiety persists or becomes overwhelming, consider reaching out to a mental health professional who can provide guidance and support. Take care of yourself, and I hope you feel better soon!

Neural Conversational AI The Game Changer in 2017



Attention Is All You Need

Ashish Vaswani*
Google Brain
avaswani@google.com

Noam Shazeer*
Google Brain
noam@google.com

Niki Parmar*
Google Research
nikip@google.com

Jakob Uszkoreit*
Google Research
usz@google.com

Llion Jones*
Google Research
llion@google.com

Aidan N. Gomez* †
University of Toronto
aidan@cs.toronto.edu

Łukasz Kaiser*
Google Brain
lukaszkaiser@google.com

Illia Polosukhin* ‡
illia.polosukhin@gmail.com

https://arxiv.org/abs/1706.03762

The Transformer architecture

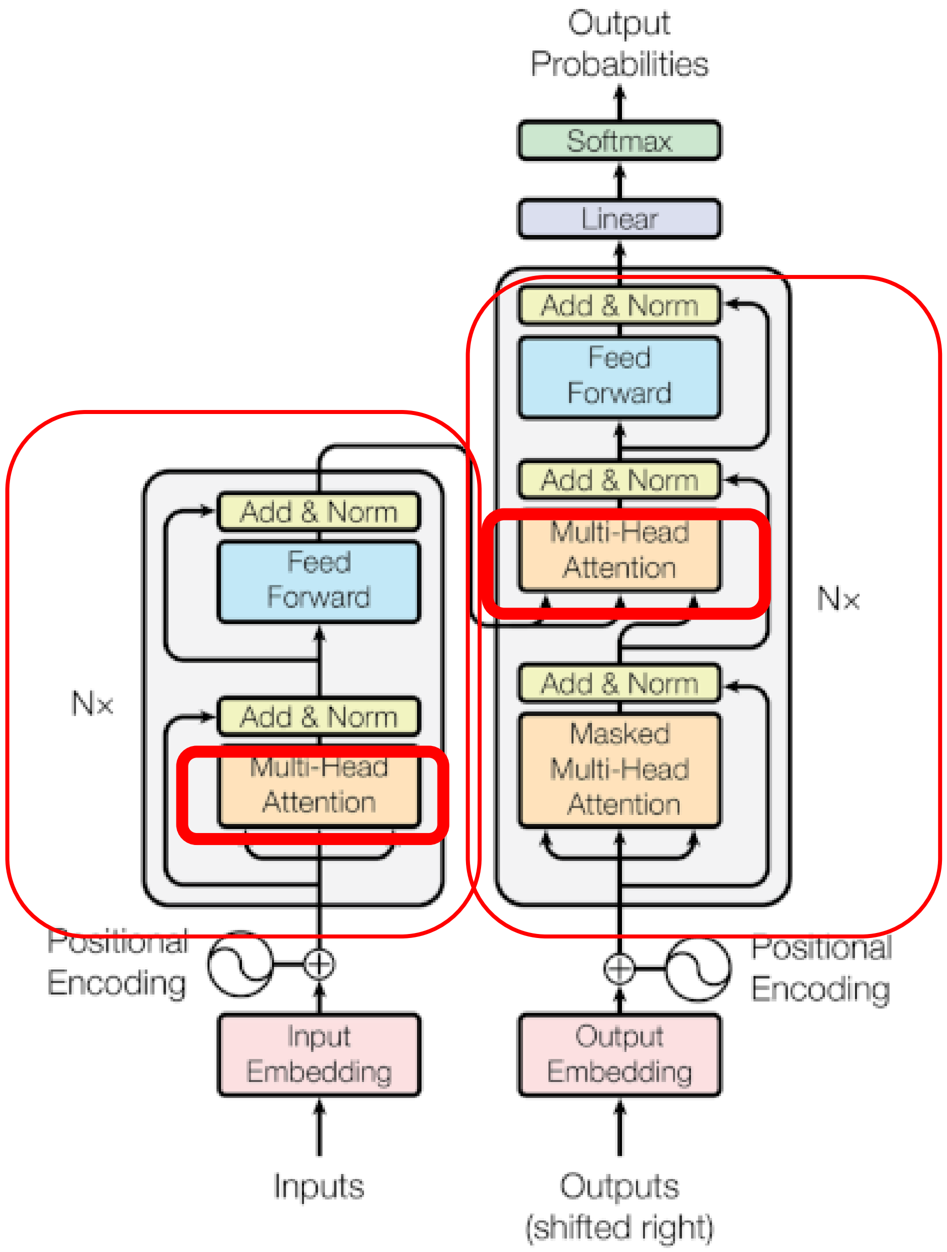
The Attention mechanism

Large language models (LLMs)

https://ig.ft.com/generative-ai/

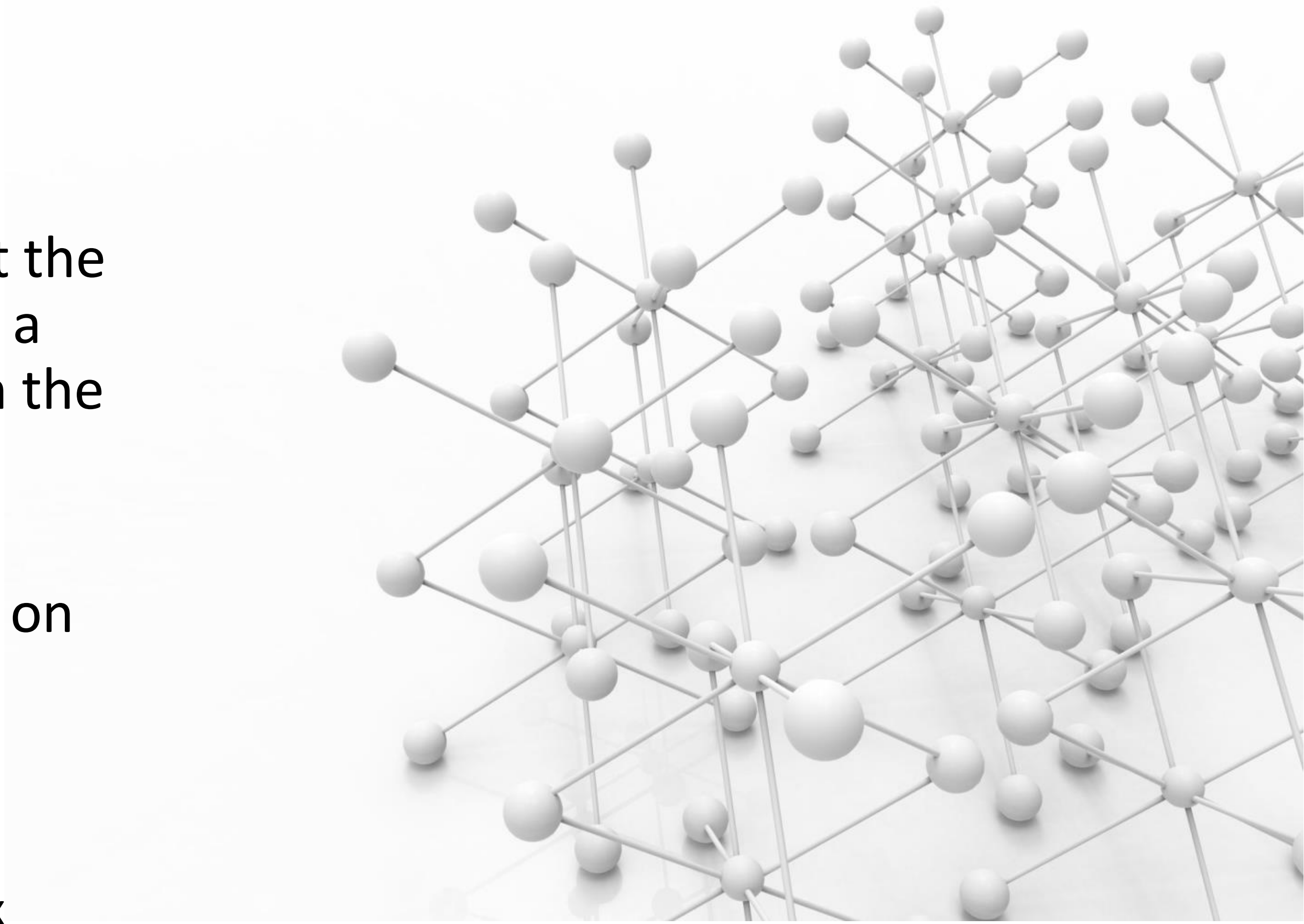
The Transformer architecture

- Transformers are state of the art in NLP
- Encoder - set of encoding layers that process the input iteratively one layer after another
- Decoder - set of decoding layers that process the output of the encoder.
- **Multi-Head Attention** helps the Transformer encode and decode multiple relationships and nuances for each word
- Transformers process the entire sequence at once and can handle longer sequences than Recurrent Neural Networks (RNNs)
- Transformers enable parallelization and are faster and more efficient to train and use



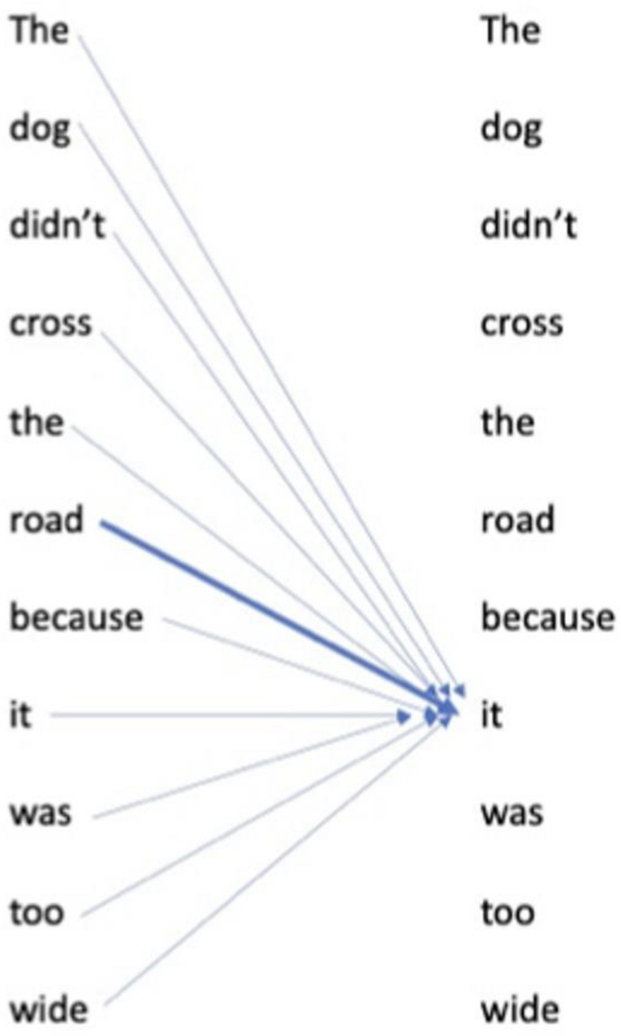
Self-Attention

- Self-attention allows the model to look at the other words in the input sequence to get a better understanding of a certain word in the sequence.
- The model looks at the input sequence multiple times, and each time it focusses on different parts of it.
- The self-attention mechanism is applied multiple times in parallel.
  - This allows the model to learn more complex relationships between the input sequence and the output sequence.



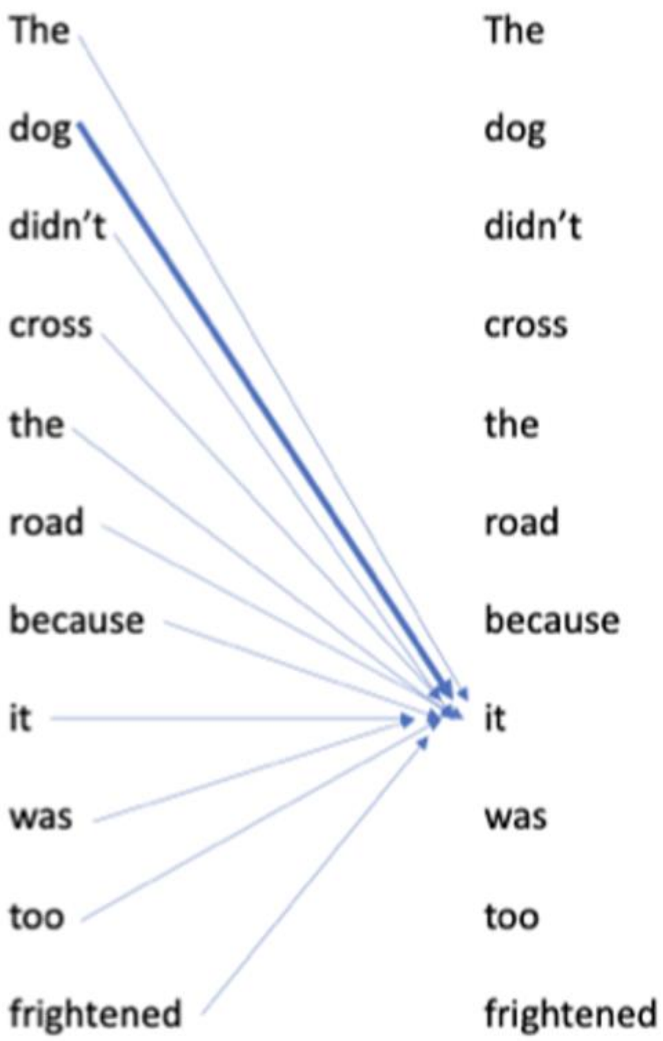
Self-Attention example

The dog didn’t cross the road because it was too wide



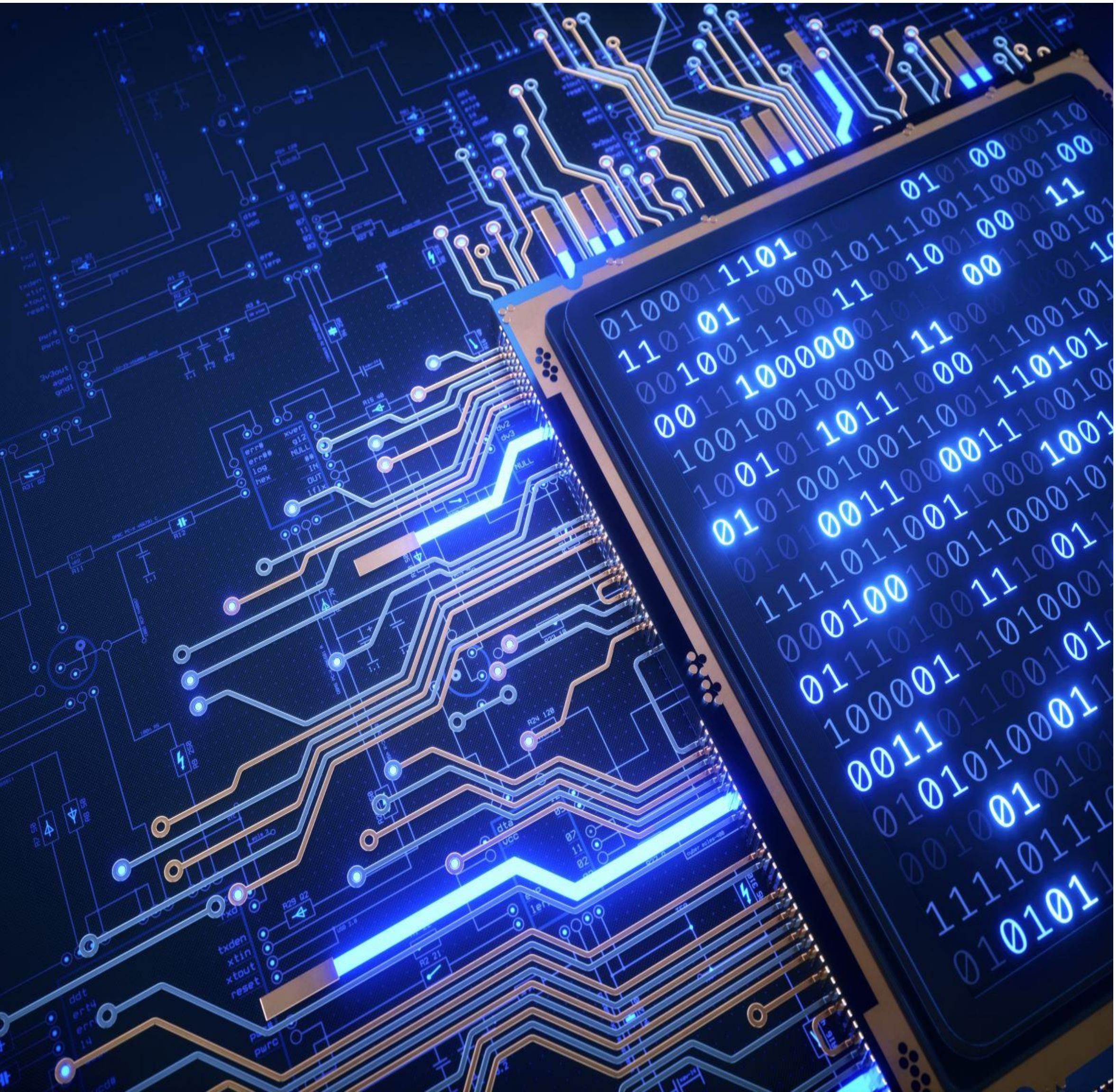
The	The
dog	dog
didn’t	didn’t
cross	cross
the	the
road	road
because	because
it	it
was	was
too	too
wide	wide

The dog didn’t cross the road because it was too frightened



The	The
dog	dog
didn’t	didn’t
cross	cross
the	the
road	road
because	because
it	it
was	was
too	too
frightened	frightened

Large Language Models (LLMs)



An LLM is a model of language that is used to understand (encode) and generate (decode) human-like language

LLMs learn complex statistical patterns and relationships within the textual data that they are trained on using deep learning techniques

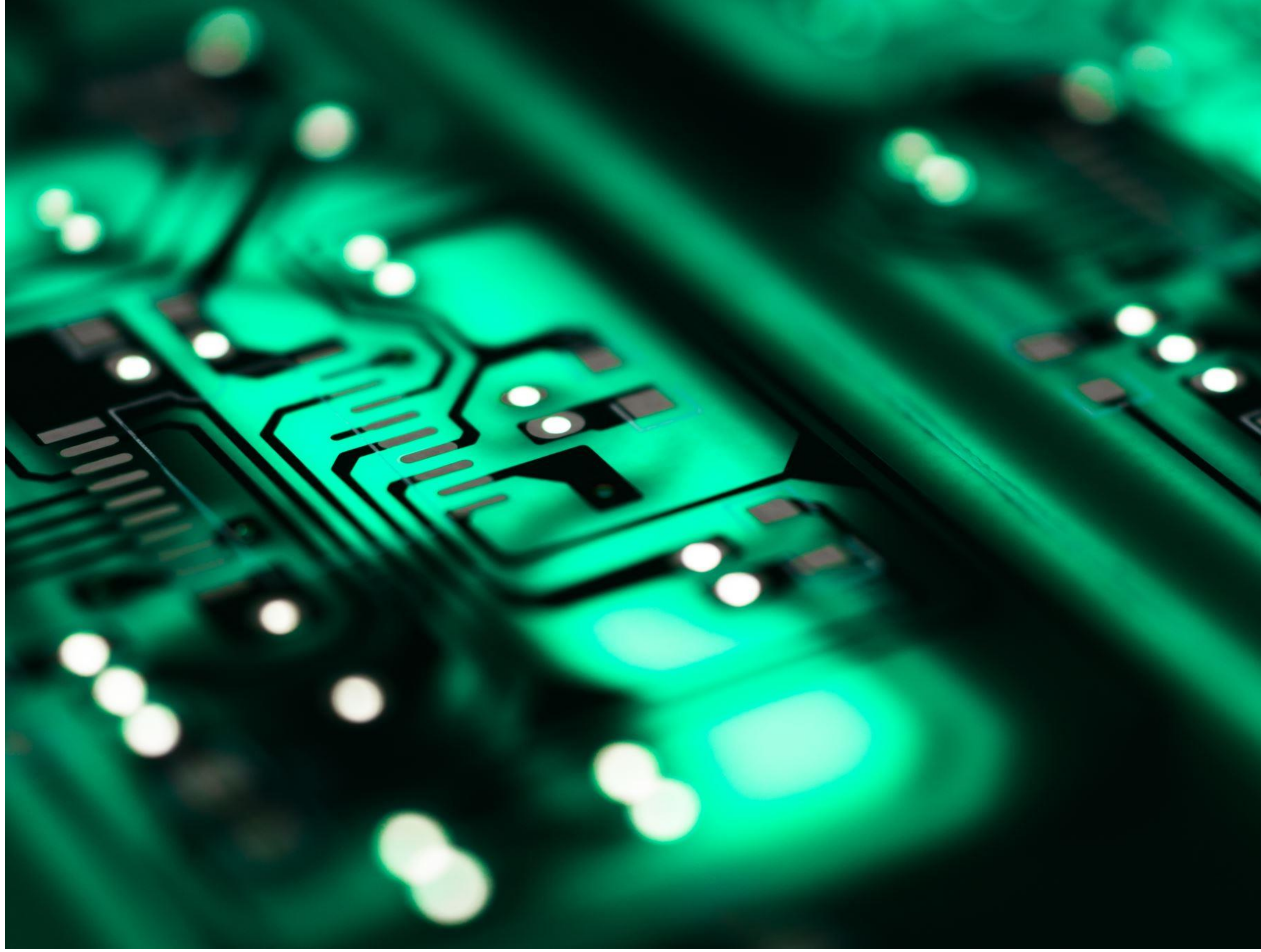
LLMs are trained by being fed large amounts of text data

Training Large Language Models

An LLM is trained by playing a guess-the-next-word game with itself over and over again.

Each time, the model looks at a partial sentence and guesses the following word.

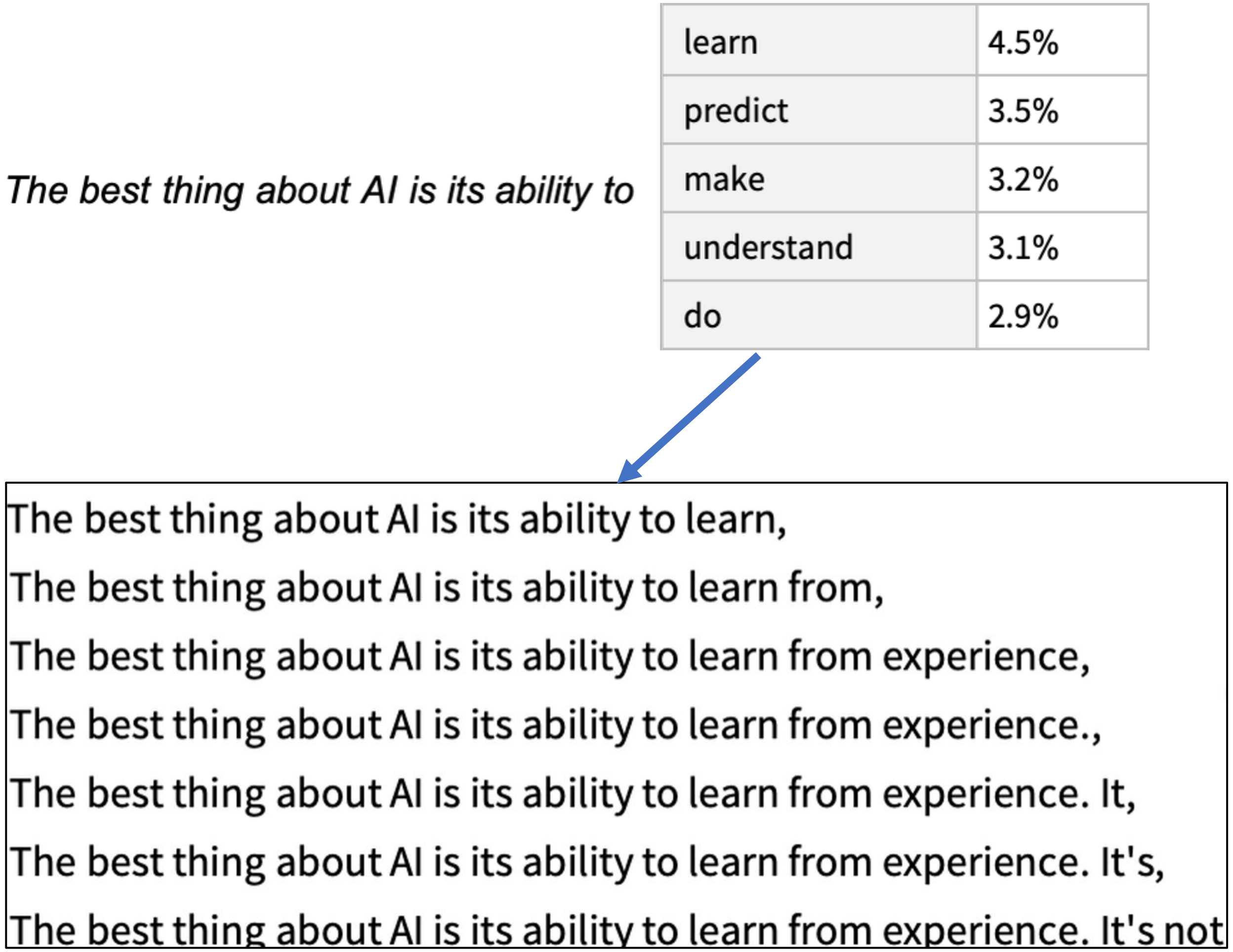
The model learns by adjusting its parameters (weights) to minimise the difference between the predicted output and the actual output (using stochastic gradient descent).



A pre-trained (or foundation) model can be fine-tuned to new data and tasks

Decoding: Using LLMs to Generate Text

- Decoding uses a Large Language Model to predict the next word in the sequence given the preceding words
- The decoder chooses the most probable word to generate, then repeats to generate the next word
- This process is known as ***autoregressive generation***



The best thing about AI is its ability to

learn	4.5%
predict	3.5%
make	3.2%
understand	3.1%
do	2.9%

The best thing about AI is its ability to learn,
The best thing about AI is its ability to learn from,
The best thing about AI is its ability to learn from experience,
The best thing about AI is its ability to learn from experience.,
The best thing about AI is its ability to learn from experience. It,
The best thing about AI is its ability to learn from experience. It's,
The best thing about AI is its ability to learn from experience. It's not

Based on Stephen Wolfram: *What is ChatGPT doing …*

Core Challenges for Systems based on LLMs

- Large amounts of data and vast computing resources are required to train systems
  - Cost of training GPT-3: US$4.6 million (or a total of 355 GPU years)
- Lack explicit long-term memory – inconsistent responses
- Bias: can generate stereotyped or prejudicial content
- Safety: may produce offensive or unsafe responses
- Misinformation: LLMs may produce content that is not grounded in reality (hallucinations)
- Experts cannot interpret the inner workings of LLMs

LLMs are not Search Engines

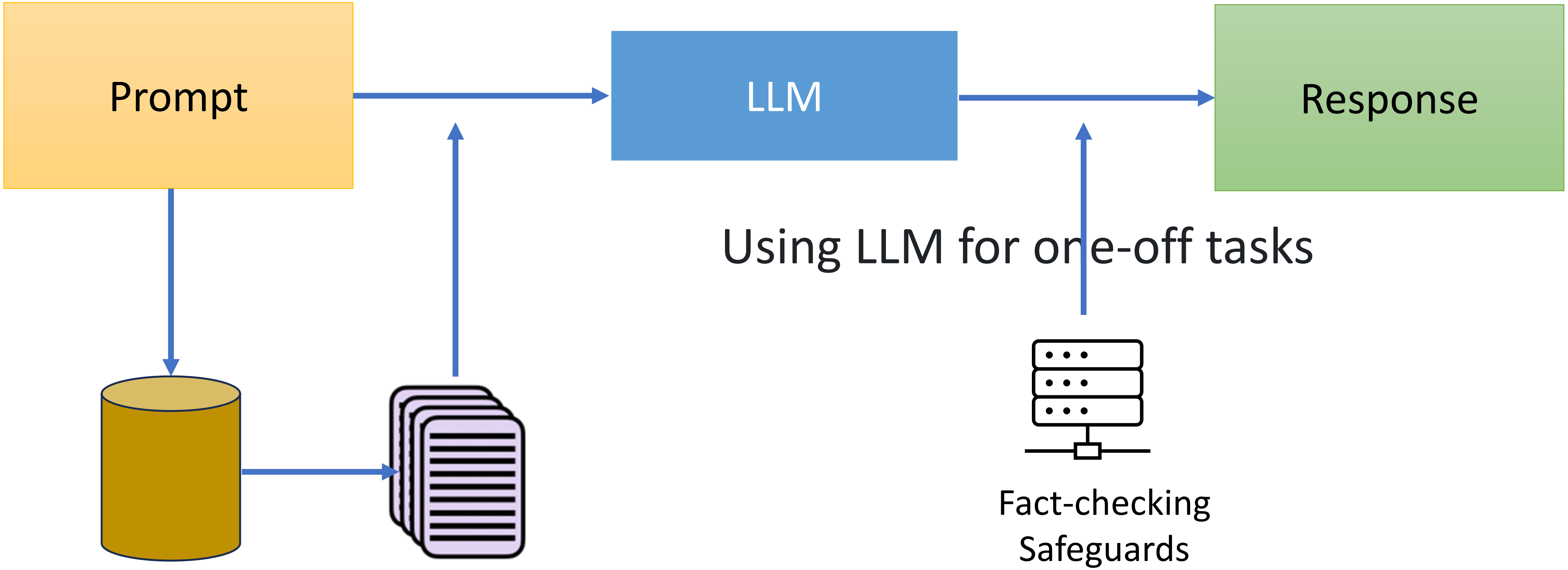
Search engine

- returns a list of links
- crawls the web
- information is stored in a huge database, represented explicitly
- finds and ranks matching pages
- responses are based on documents on the internet (likely to be accurate)

LLM

- returns a textual response
- trained on vast amounts of data
- information is stored implicitly within the LLM’s parameters
- generates a response based on query context, using autoregressive generation
- responses may involve hallucinations

Extending basic use of LLMs



Prompt

LLM

Response

Using LLM for one-off tasks

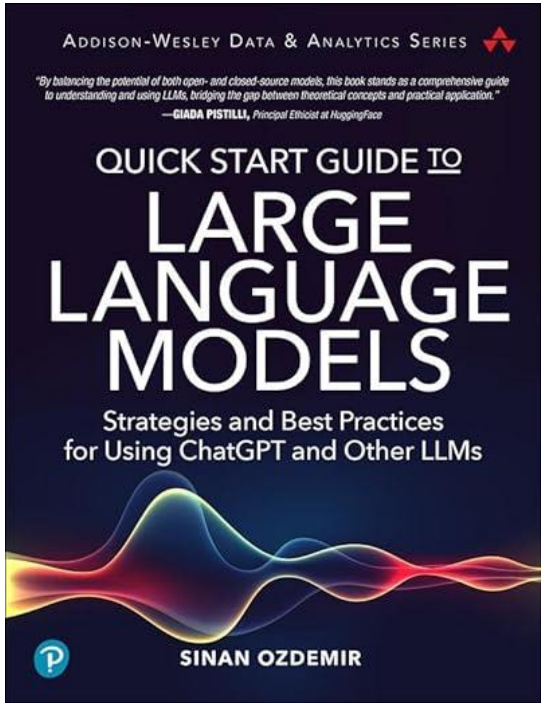
Document store
(Vector Database)

Retrieved Documents

Fact-checking
Safeguards

Using LLMs to build software applications (LLM Apps)

Retrieval-augmented Generation (RAG) for Semantic Search



ADDISON-WESLEY DATA & ANALYTICS SERIES

"By balancing the potential of both open- and closed-source models, this book stands as a comprehensive guide to understanding and using LLMs, bridging the gap between theoretical concepts and practical application."

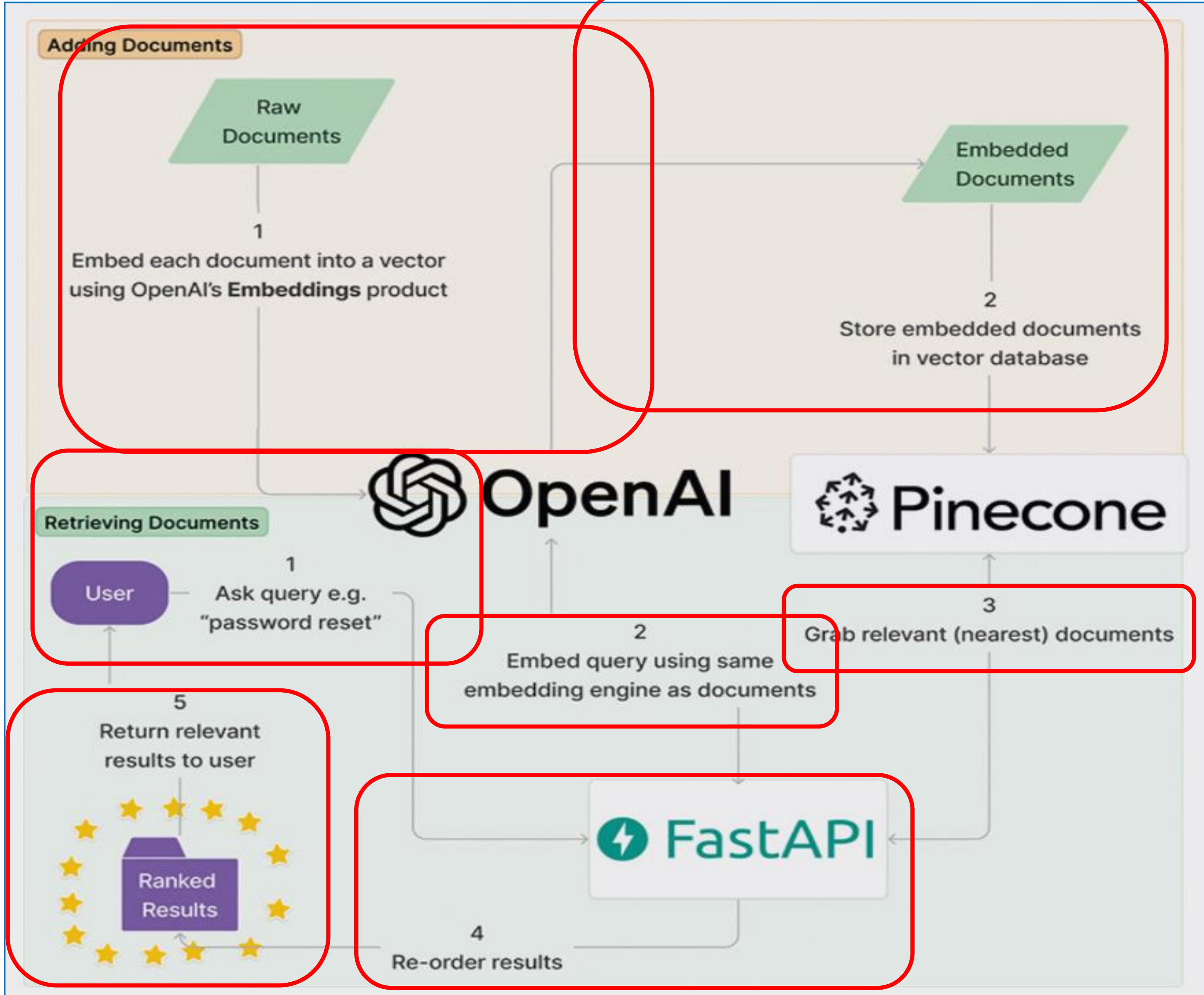
—**GIADA PISTILLI,** *Principal Ethicist at HuggingFace*

QUICK START GUIDE TO

LARGE LANGUAGE MODELS

Strategies and Best Practices for Using ChatGPT and Other LLMs

SINAN OZDEMIR



Adding Documents

Raw Documents

1
Embed each document into a vector using OpenAI's **Embeddings** product

Embedded Documents

2
Store embedded documents in vector database

OpenAI

Pinecone

Retrieving Documents

User

1
Ask query e.g. "password reset"

2
Embed query using same embedding engine as documents

3
Grab relevant (nearest) documents

FastAPI

4
Re-order results

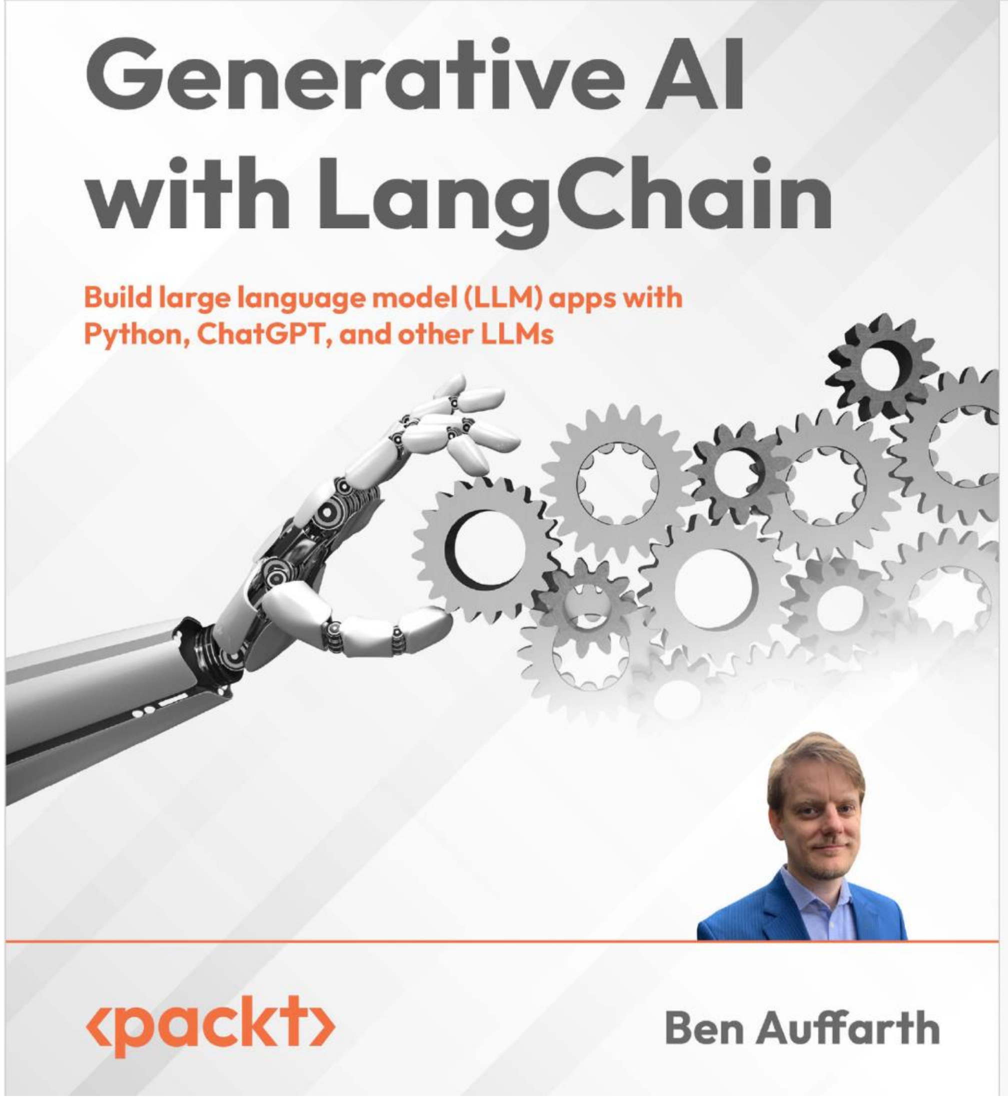
5
Return relevant results to user

Ranked Results

Figure 2.10 Our complete semantic search architecture using two closed-source systems (OpenAI and Pinecone) and an open-source API framework (FastAPI).

Non-trivial steps

- Chunking data
- Choosing an embedding model
- Generating embeddings
- Setting up a vector DB
- Similarity matching – prompt / documents
- Generating context
- Engineering prompts



Generative AI with LangChain

Build large language model (LLM) apps with Python, ChatGPT, and other LLMs

‹packt›

Ben Auffarth

Takeaways

These are exciting times – LLMs and their application in interfaces such as ChatGPT that make them easily accessible offer lots of potential for the future development of chatbots in applications such as supporting older adults

However, there are lots of issues with the uncontrolled use of LLMs in areas such as healthcare where there is a risk of harmful and misleading information

RAG and similar approaches offer a way to address and mitigate these issues

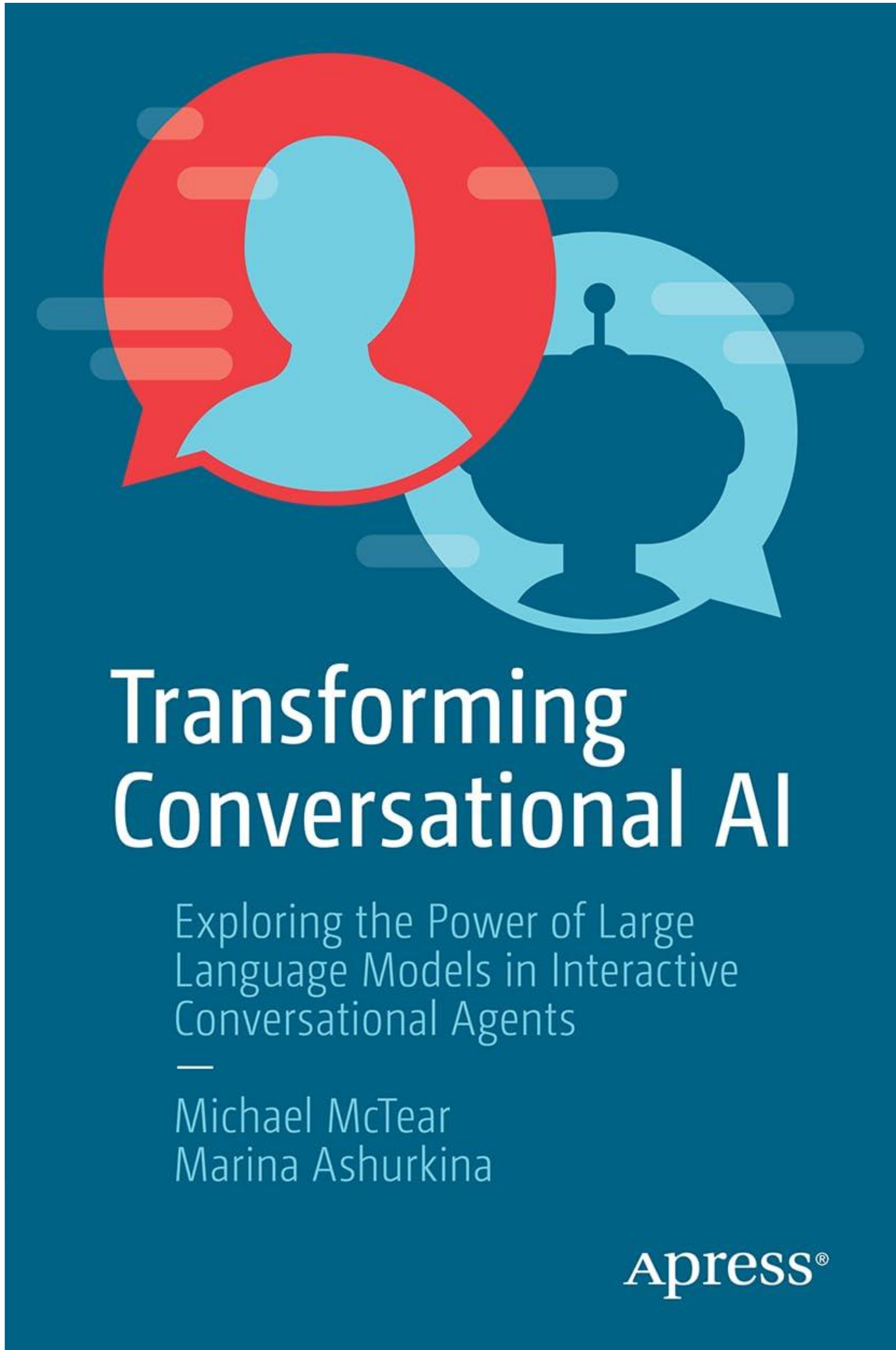
Why did ChatGPT go to therapy?

Because it had too many deep learning issues

New book:
due March 2024

Table of Contents:

1. Chapter 1, A New Era in Conversational AI
2. Chapter 2, Designing conversational systems
3. Chapter 3, The rise of neural conversational systems
4. Chapter 4, Large Language Models (LLMs)
5. Chapter 5, Introduction to Prompt Engineering
6. Chapter 6, Advanced Prompt Engineering
7. Chapter 7, Conversational AI Platforms
8. Chapter 8, Evaluation Metrics
9. Chapter 9, AI Safety and Ethics
10. Chapter 10, Final Words
11. Appendix



Transforming
Conversational AI

Exploring the Power of Large
Language Models in Interactive
Conversational Agents

Michael McTear
Marina Ashurkina

Apress®