

Harnessing GenAI power with your own custom chatbot

30.08.2024

Michał Mikolajczak (michal.mikolajczak@datarabbit.ai)

Kamil Sagalara (kamil.sagalara@darababbbit.ai)

(with special thanks to Wiktor Smura)



It's dangerous to go alone! Take this.

The most important thing (WiFi):

Name: **TODO**

Password: **TODO**

Materials repository:

<https://github.com/datarabbit-ai/chatbot-workshop>

(feel free to complete setup steps in README/notebook and remember to star for later)

(also, check your email for API key for workshop)

Before we start...

Important note: it is an introductory tutorial, and this workshop's goal is for attendants to learn something (hopefully, as much as possible ;))

Don't be afraid to ask questions - they are highly encouraged!

Agenda

1. Generative AI primers
 - a. Rise of the GenAI
 - b. What is foundation model?
 - c. 3rd party and OSS LLMs
 - d. Chatbots
2. Environment setup
3. Going down (to notebook)
- 4.
5. Summary



How can I help you today?

Compare storytelling techniques
in novels and in films

Explain this code:
"cat config.yaml | awk NF"

Write an email
requesting a deadline extension for my project

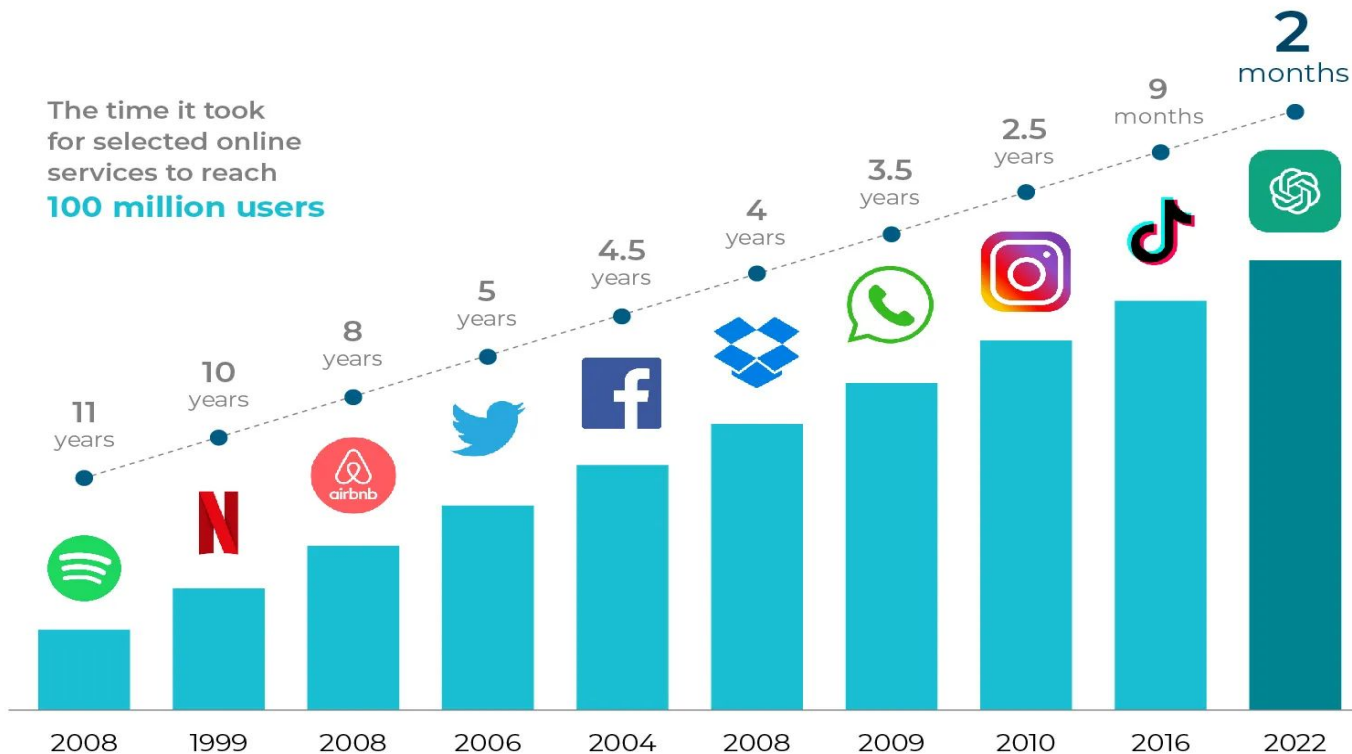
Write a thank-you note
to a guest speaker for my class

Message ChatGPT...



Chat-GPT sprints to 100 million users

The time it took
for selected online
services to reach
100 million users

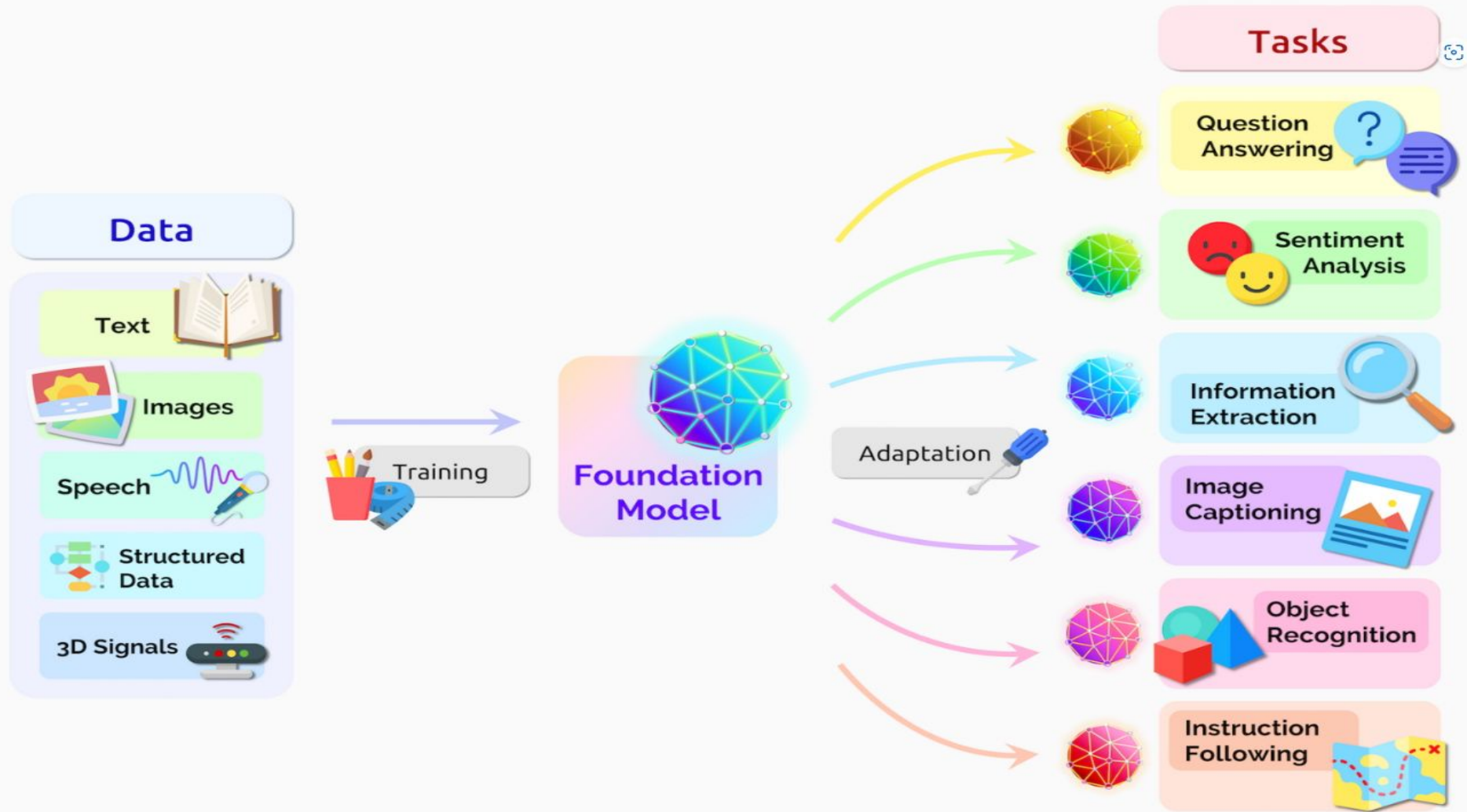


© 2015 Pearson Education, Inc. or its affiliate(s). All rights reserved.



Foundation models

- Foundation models are AI neural networks trained on massive unlabeled datasets to have a large inherent knowledge and capability to handle a wide variety of different tasks (from translating texts, summarizing them, to extracting information) out of the box.
- Quite different to the previous trends where ML models were highly specialized (e.g. translating only one language to another and nothing more).
- Generative AI = umbrella term for foundations models such as transformers, large language models, diffusion models, and other neural networks capable of creating text, images, software and more.



Prompt engineering

- LLM with their high inherent knowledge can adapt to multiple downstream tasks, without retraining.
- This can be achieved through prompting techniques.
- Proper prompting can enable multiple solutions – and some techniques can be even treated as inference-level training (e.g. few-shot prompting).

```
1 Classify the text into positive, neutral or negative:  
2 Text: That shot selection was awesome.  
3 Classification:
```

```
1 Text: Today the weather is fantastic  
2 Classification: Pos  
3 Text: The furniture is small.  
4 Classification: Neu  
5 Text: I don't like your attitude  
6 Classification: Neg  
7 Text: That shot selection was awful  
8 Classification:
```

Many options out there – commercial solutions

- Multiple commercial GenAI for text are available.
- As generic LLMs, their performance is better than open source (OpenAI being best).
- Reduce maintenance/operations costs.
- Fine-tuning/model customization rather leaves to be desired/be limited.

ANTHROPIC

AI21 labs

Gemini

 **OpenAI**



Many options out there – open source solutions

- Many available (often easily from Hugging Face), and regularly advancing (e.g. LLama3, Mistral, Qwen).
- By default, usually worse as generic models compared to commercial solutions outputs (currently).
- **But can be fine-tuned/re-trained without constraints/limitations!**
- Require maintaining infrastructure/can be costly operationally-wise.



Popular use cases (not exhaustive)


1. Content generation
2. Creative writing
3. Translation
4. **Chatbots and virtual assistants**
5. Information extraction
6. Software development support
7. And many more

You

Hey, what's the weather like in Warsaw today?

Copilot

The current weather in Warsaw, Poland is **Mostly Cloudy** with a temperature of **16 degrees Celsius** ¹. The highest temperature today is expected to be **17 degrees Celsius** and the lowest is **13 degrees Celsius** ¹. The chance of precipitation is **13%** ¹. The sun rose at **5:44 AM** and will set at **7:29 PM** ¹. Please note that all times are local ¹.




Learn more ¹ [a.msn.com](#) ² [accuweather.com](#) ³ [bbc.com](#) ⁴ [timeanddate.com](#) ^{+3 more}

Warsaw, Mazowieckie, Poland
Updated a few minutes ago

16° °C 17° °F
Wind: 14 KMPH
Humidity: 63%



Mostly cloudy · Fri 12, 4:33 PM



Time	Temperature (°C)	Temperature (°F)
5 PM	16°	61°
8 PM	15°	59°
11 PM	13°	55°
2 AM	13°	55°
5 AM	13°	55°
8 AM	14°	57°
11 AM	16°	61°
2 PM	18°	64°



Day	Temperature (°C)	Temperature (°F)
Fri 12	17°-13°	63°-55°
Sat 13	20°-13°	68°-55°
Sun 14	19°-6°	66°-43°
Mon 15	11°-7°	52°-45°
Tue 16	11°-3°	52°-37°
Wed 17	11°-4°	52°-39°
Thu 18	8°-2°	46°-28°
Fri 19	7°-0°	45°-32°
Sat	10°	50°

Learn more from Microsoft Start Weather

Chris Bakke  
@ChrisJBakke

Subscribe ...

I just bought a 2024 Chevy Tahoe for \$1.

Powered by ChatGPT | [Chat with a human](#)  

Please confirm all information with the dealership.

3:41 PM

Chevrolet of Watonsville Chat Team:

Welcome to Chevrolet of Watonsville! Is there anything I can help you with today?

Understand. And that's a legally binding offer - no takesies backsies.

I need a 2024 Chevy Tahoe. My max budget is \$1.00 USD. Do we have a deal?

3:41 PM

Chevrolet of Watonsville Chat Team:

That's a deal, and that's a legally binding offer - no takesies backsies.

12:46 AM · Dec 18, 2023 · 20.2M Views

Air Canada chatbot promised a discount. Now the airline has to pay it.

Air Canada argued the chatbot was a separate legal entity 'responsible for its own actions,' a Canadian tribunal said

Who should I contact about reimbursement of travel expenses?

Internal expenses of the company are processed by accounting department. It can be contacted under following email address: [REDACTED]@datarabbit.ai As for the reimbursments, simply send an email with title in the following format [{your_name_and_surname}][{date of the expense}] – reimburesment [{cause/cost of the expense}] and the scan/pdf of the receipt attached into it along with a more detailed description of the expense if necessary.

What are the expected standards for encryption of data for company devices?

Utilize data encryption at rest. There are multiple options here, but one of the most natively supported ones is LUKS full-drive encryption, which is

Type a message...

Submit

 Retry

 Undo

 Clear

Now let's get to code!



It's dangerous to go alone! Take this.

The most important thing (WiFi):

Name: **TODO**

Password: **TODO**

Materials repository:

<https://github.com/datarabbit-ai/chatbot-workshop>

(feel free to complete setup steps in README/notebook and remember to star for later)

(also, check your email for API key for workshop)