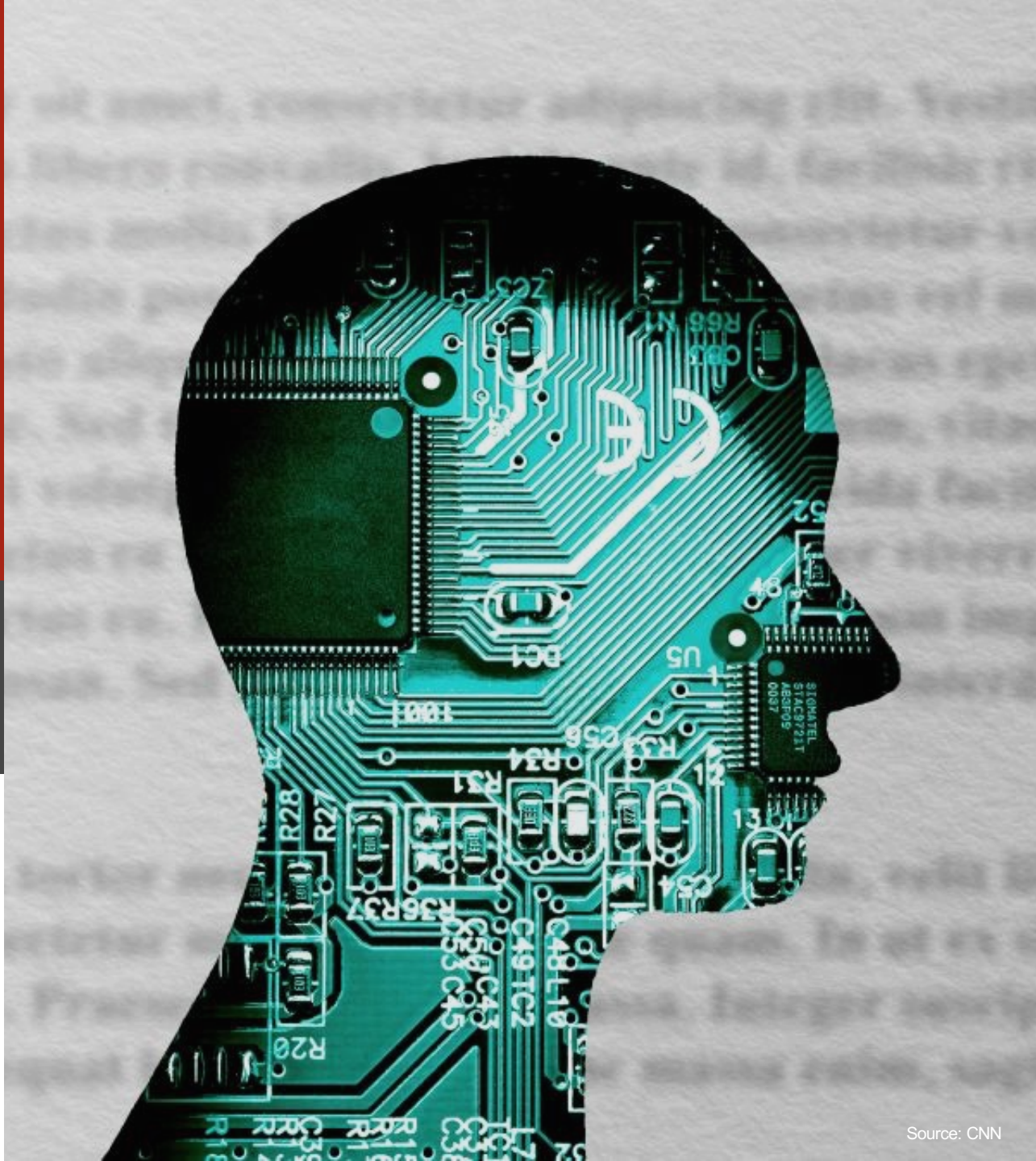
Generative AI – Building a Secure and Local “Mini ChatGPT”

Personal Project & Portfolio
October 2023

By: Achmad Arviandito Caessara

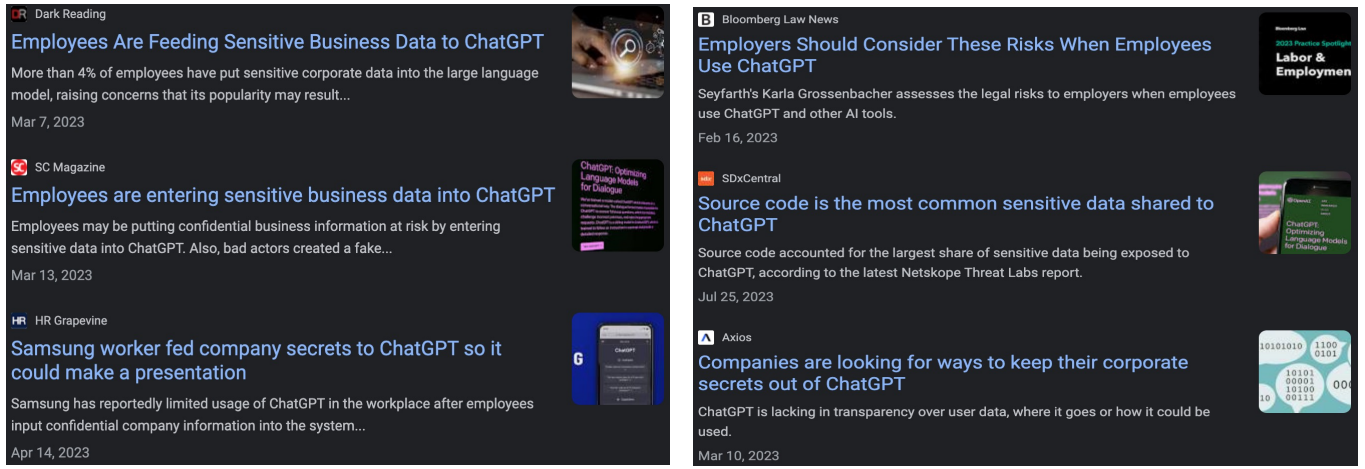


Source: CNN

The project objective is to create a locally hosted and secure “Mini ChatGPT” to address the essential need for protecting sensitive data

Problem Statement and Project Objective

Current situation



- **Confidential data should never** be input into ChatGPT, as numerous cases have demonstrated instances where individuals input sensitive information, resulting in **leaks of confidential data**.
- Nevertheless, generative AI models, such as ChatGPT, prove **highly valuable in enhancing productivity** and expediting task completion.

Needs

To address these concerns, we need to take several **factors into consideration**:

1. Develop our **own model** with features similar to the original ChatGPT
2. Train the model using our **internal documents and database** to incorporate **“personalized” knowledge.**
3. Ensure it operates within our **local premises** to prevent data leakage**.**

Therefore, this project aims to construct a simple model, referred to as "Mini ChatGPT," capable of being trained on our proprietary data and operated within our local environment

Source: Google News

In this project, we leverage Nike information from the Wikipedia, followed by preprocessing the data into full-text and Q&A structures

Data Utilized for Model Development

WIKIPEDIA
The Free Encyclopedia

Search Wikipedia | Search

Create account Log in

Nike, Inc.

76 languages

Article Talk | Read View source View history Tools

From Wikipedia, the free encyclopedia

Coordinates: 45.5093°N 122.8299°W

Contents [hide]

(Top)
- History
- Products
- Headquarters
- Controversies
- Environmental record
- Marketing strategy
- Sponsorship
- Ties with the University of Oregon
- Causes
- Program
- Research
- See also
- Notes
- References
- Further reading
- External links

This article is about the company. For other uses, see Nike (disambiguation).

Nike, Inc.[note 1] (stylized as ***NIKE***) is an American athletic footwear and apparel corporation headquartered near Beaverton, Oregon, United States.[4] It is the world's largest supplier of athletic shoes and apparel and a major manufacturer of sports equipment, with revenue in excess of US$46 billion in its fiscal year 2022.[5][6]

The company was founded on January 25, 1964, as "Blue Ribbon Sports", by Bill Bowerman and Phil Knight, and officially became Nike, Inc. on May 30, 1971. The company takes its name from Nike, the Greek goddess of victory.[7] Nike markets its products under its own brand, as well as Nike Golf, Nike Pro, Nike+, Air Jordan, Nike Blazers, Air Force 1, Nike Dunk, Air Max, Foamposite, Nike Skateboarding, Nike CR7,[8] and subsidiaries including Air Jordan and Converse (brand). Nike also owned Bauer Hockey from 1995 to 2008, and previously owned Cole Haan, Umbro, and Hurley International.[9] In addition to manufacturing sportswear and equipment, the company operates retail stores under the Niketown name. Nike sponsors many high-profile athletes and sports teams around the world, with the highly recognized trademarks of "Just Do It" and the Swoosh logo.

As of 2020, it employed 76,700 people worldwide.[10] In 2020, the brand alone was valued in excess of $32 billion, making it the most valuable brand among sports businesses.[11] Previously, in 2017, the Nike brand was valued at $29.6 billion.[12] Nike ranked 89th in the 2018 Fortune 500 list of the largest United States corporations by total revenue.[13]

History

See also: Nike timeline

Nike, originally known as Blue Ribbon Sports (BRS), was founded by University of Oregon track athlete Phil

Nike, Inc.

Headquarters near Beaverton, Oregon, U.S.

Formerly	Blue Ribbon Sports, Inc. (1964–1971)
Type	Public
Traded as	NYSE: NKE (Class B) DJIA component S&P 100 component S&P 500 component
ISIN	US6541061031
Industry	Apparel

Source: *https://en.wikipedia.org/wiki/Nike,_Inc.*

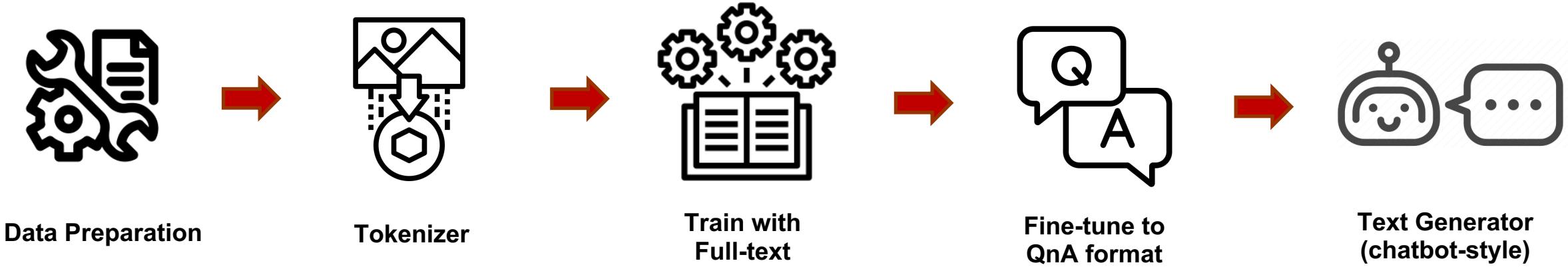
We will **treat the Nike information as our "confidential data"** and preprocess it into two types for training and fine-tuning purposes:

- **Full-Text** (Training)
  Contains 7,097 words in total with 2,780 unique words.
- **Question and Answer** (Fine-Tuning)
  We prepare 200 sets of questions and answers.

Note: You may try using any other data of your choice!

Using the Transformer library, tokenize the data, train initially with the full text, and fine-tune it in a Q&A format to have a chatbot-style model

High-level Flow of Model Development



In this project, there are five main steps, and only need to run four Python commands in the project files:

1. **Data Preparation –** Prepare the dataset, both the full text and Q&A, then standardize the format. -> run: **$python data/prepare.py**
2. **Tokenizer –** Tokenize every word using GPT2Tokenizer.
3. **Train with Full-text –** Train the model using the pre-trained model 'GPT-2' with defined parameters. -> run: **$python train.py**
4. **Fine-tune to QnA format –** Fine-tune the model to yield results in a question-and-answer format. -> run: **$python fine_tune.py**
5. **Text Generator –** Input a question then run the model to generate the answer. ->run: **$python generate.py**

textGenerator

EXPLORER

OPEN EDITORS

TEXTGENERATOR

- model_finetuned_qanda
  - checkpoint-500
  - runs
  - config.json
  - merges.txt
  - pytorch_model.bin
  - special_tokens_map.json
  - tokenizer_config.json
  - training_args.bin
  - vocab.json
- model_fulltext
  - runs
  - config.json
  - merges.txt
  - pytorch_model.bin
  - special_tokens_map.json
  - tokenizer_config.json
  - training_args.bin
  - vocab.json
- .gitignore
- ~$Project Context.pptx
- delete.py
- fine_tune.py
- generate.py M
- Project Context.pptx
- README.md
- train.py

OUTLINE

TIMELINE

train.py | generate.py M

generate.py > ...

```python
max_len = 40

user_input = input("Enter question: ")
sequence1 = "[Q] " + user_input

# Functions
def load_model(model_path):
```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS

zsh

```
(base) macbookpro@MacBooks-MacBook-Pro textGenerator % ls
Project Context.pptx    __pycache__    delete.py    generate.py    train.py
README.md               data           fine_tune.py model          ~$Project Context.pptx
(base) macbookpro@MacBooks-MacBook-Pro textGenerator % python3 gener
```

main* 0 0 0 | Ln 11, Col 13 | Spaces: 4 | UTF-8 | LF | Python 3.10.5 64-bit | Go Live | Prettier

Several next steps for improvement: tuning the parameters, exploring different training methods, or experimenting with other infrastructure

Next Step

1. Fine-tune the model parameters
2. Use GPUs or TPUs to speed up the model training
3. Try using another dataset
4. Add further training methods: supervised policy, reward model, PPO
5. Explore other tokenizers, models, and infrastructure.
6. Try it yourself!

   https://github.com/ArvianditoCaessara/textGenerator

   With only four commands required.

Special acknowledgment to these incredible individuals who inspired the development of this project through their previous work

List of Acknowledgements

- **Daniel Guetta -** Associate Professor at Columbia Business School / CBS, PhD
- **Andrej Karpathy -** Sr. Director of AI at Tesla / Stanford, PhD
- **Sophia Yang -** Sr Data Scientist at Anaconda / University of Texas at Austin, PhD
- **Sreenivas Bhattiprolu -** Director of Digital Solutions at Zeiss / Michigan Technological Univ., PhD

Note: Please check their GitHub and YouTube channels for other amazing projects!

Thank you !