

DATA5000 Artificial Intelligence Programming in Business Analytics

Introduction to large language models
Workshop #10



DATA5000 Roadmap

Week 1

Artificial Intelligence with Python

Week 2

Al Predictive Models

Week 3

Deep Learning

Week 4

Causal Al

Week 5

A1: Skills-Building (25%)

Week 6

Causal Forests & Meta-Learners Week 7

Generative Al

Week 8

Al in Image & Video Recognition

Week 9

A2: Generative AI Startup (35%) Week 10

Introduction to Large Language Models Week 11

Innovation with Large Language Models 1

Week 12 Innovation with Large Language Models 2

Week 13

A3: Project Report (40%)



COMMONWEALTH OF AUSTRALIA Copyright Regulations 1969

WARNING

This material has been reproduced and communicated to you by or on behalf of Kaplan Business School pursuant to Part VB of the *Copyright Act* 1968 (**the Act**).

The material in this communication may be subject to copyright under the Act. Any further reproduction or communication of this material by you may be the subject of copyright protection under the Act.

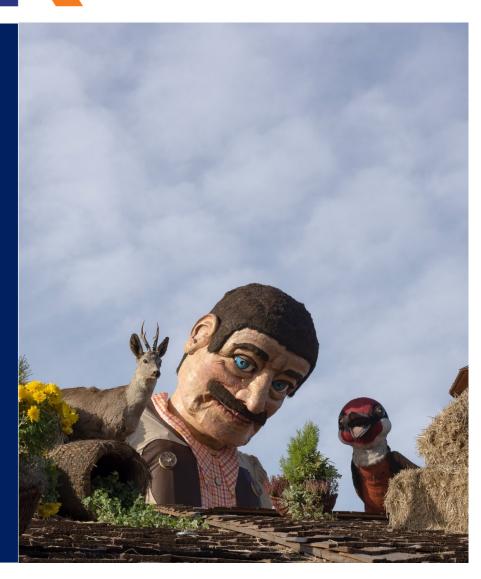
Do not remove this notice.



Lesson Learning Outcomes

1	Explore the capability and limitation of Large language models (LLMs)
2	Investigate the use of Llama with LLMs
3	Integrate Llama with python in order to take advantage of the features of all both technologies





Unleashing the Power of Large Language Models: Revolutionising Business Analytics

Transforming business analytics with large language models



Introduction to Large Language Models

WHAT ARE LARGE LANGUAGE MODELS?



Explore the definition and key characteristics of Large Language Models (LLMs), which are Al systems trained on vast amounts of text data to understand and generate human-like language.

Examples:

- ChatGPT
- Gemma
- Claude
- Cohere
- Gemini
- Llama
- Orca
- BERT

This Photo by Unknown Author is licensed under CC BY-ND





The Journey of LLMs

2011

GPT-1 introduced, demonstrating the potential of language models for text generation **2019**

GPT-3 launched, setting new benchmarks for language understanding and generation with its massive scale 2021

Multimodal models like
DALL-E emerge,
combining language
understanding with visual
generation

2018

GPT-2 released, showcasing improved text generation capabilities and raising concerns about potential misuse **2020**

Transformer models like T5 and BERT become widely adopted, enabling new applications in NLP **2022**

ChatGPT debuts, captivating the public with its ability to engage in natural conversations and tackle complex tasks





- Large language models are a group of AI models trained on enormous amounts of data in order to understand queries indifferent languages and then generate natural language, images and perform other tasks.
- LLMs are a subset of Gen AI, since the latter can do a greater variety of tasks, e.g. generate computer code and audio.
- LLMs are made up of many layers of trained neural networks (a transformer model). Based on context in previous parts of a sentence, the networks learn to predict the next word in a sentence.

https://www.ibm.com/topics/large-language-models



Large Language Models and Business Analytics



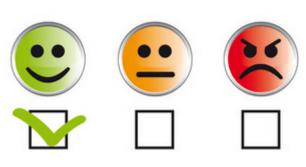
TRANSFORMING BUSINESS ANALYTICS

Discover how LLMs can revolutionise business analytics (BA) by providing intelligent insights, automating report generation, and enabling more natural language interaction with data.

LLMs are not designed to perform analysis of numerical data, however they can provide insight into the context of textual data.

Thus, LLMs can improve BA

- Finding relevant features in ML models
- Sentiment analysis (key words, themes)
- General context



<u>This Photo</u> by Unknown Author is licensed under <u>CC BY-SA-NC</u>





Applications of Large Language Models

PA

APPLICATIONS OF LLMS

Discuss the diverse applications of LLMs, including text generation, language translation, sentiment analysis, content creation, and task automation across various industries.

In particular, applications of LLMs include

- Bioinformatics
 - Identifying DNA alterations, drug targets and protein structures
- Customer Service
 - Personalised solutions at scale
- Online searches (in near real-time)

Sources: Sarumi, A & Heider D 2024, 'Large language models and their applications in bioinformatics', *Computational and Structural Biotechnology Journal*, vol. 23, pp. 3498 – 3505.





Challenges and Limitations of Large Language Models



CHALLENGES AND LIMITATIONS

Explore the potential challenges and limitations of LLMs, such as bias, lack of contextual understanding, and the need for careful deployment and oversight.

Other issues LLMs suffer from are

- Hallucinations (erroneous output)
- Lack of long term memory

https://www.labellerr.com/blog/challenges-in-development-of-llms/#:~:text=From%20dealing%20with%20massive%20datasets,struggling%20to%20update%20old%20knowledge.

IN CONCLUSION, UNDERSTANDING THE FUNDAMENTALS OF LARGE LANGUAGE MODELS AND THEIR TRANSFORMATIVE IMPACT ON BUSINESS ANALYTICS IS CRUCIAL FOR ORGANISATIONS TO STAY COMPETITIVE AND LEVERAGE THE POWER OF AI-DRIVEN INSIGHTS.



Activity: LLMs

- Form groups
- Consider one of the following industries
 - Healthcare
 - Mining
 - Retail
- Suggest an application of LLMs that hasn't been mentioned so far in the slides
- Report back to class





 Language models are built to predict the next word based on information so far. They use probabilities to do this which form part of a transformer model.

What is a Transformer?



- ANSWER: A type of neural network model that learns the context of data by transforming one sequence of data into another sequence, paying attention to the order of the data,
 - eg. translating words in a French sentence into those of an English sentence with the same meaning



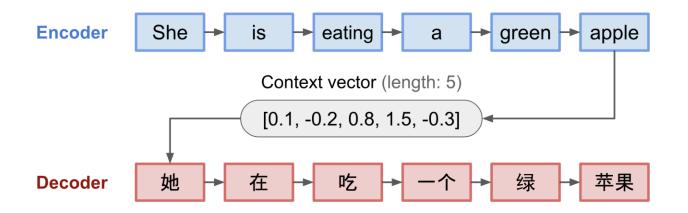


Transformers

- Use mathematics to find an optimal solution
- Differ from previous types of neural nets
- Information such as images, text and databases can be used
- Can run fast due to parallel processing across computers
- Pay attention, but how?
 - Answer: Transformers use encoders and decoders (see next page for definitions) to create a kind of map of how each element in the data relates to other elements
- Self-attention on inputs (one word attending to other words in a sequence) enables the algorithm to find the meaning from words in the context of particular sentences



Encoders	Decoders
Convert raw into coded signals	Convert coded into raw signals
Simple process	Complicated process
Examples: Video encoders, email	Examples: Memory storage devices









Glossary: Latent variable = a variable inferred from observed data =,
e.g. inferring intelligence quotient (IQ), a score relative to others in the
population, based on testing/questioning

A perceiver

- Is a very flexible transformer One model to handle multiple inputs
- Can process and classify multiple data types, such as images, sounds, videos
- Directs the self-attention process on latent variables (not on input)
- Is computationally more efficient than transformers
- Can use an existing algorithm to start with



- Examples of Perceivers
- Google's DeepMind perceiver IO
- Perceiver Al



This Photo by Unknown Author is licensed under CC BY



Meta Al

Meta's data2vec

https://ai.facebook.com/research/data2vec-a-generalframework-for-self-supervised-learning-in-speechvision-and-language/

https://perceiver.ai/

https://huggingface.co/blog/perceiver

https://www.zdnet.com/article/metas-data2vec-is-the-next-step-toward-one-neural-

network-to-rule-them-all/

Activity: LLMs

- Form groups
- Access one of the following webpages on the benefits and/or limitations of LLMs:

https://aithority.com/machine-learning/benefits-and-limitations-ofllm/#:~:text=LLMs%20offer%20unparalleled%20benefits%20in,need%20 for%20vast%20computational%20resources.

https://enterprisetalk.com/featured/benefits-limitations-of-using-large-language-models-llms

https://www.pairaphrase.com/llm-translation-advantages-disadvantages/

Report back on some of your conclusions from these reports



Llama 3.2

Llama 3.2

https://www.llama.com/

- Llama is a state-of-the art family of large language models and large "multimodal" models (Al models able to understand various modalities, e.g. videos, images, equations, text and audio).
- Meta has made the models free for anyone to use for research/commercial purposes
- Llama uses transformer architecture
- Llama 3.2 11B-Vision model adds "visual reasoning" to existing models, i.e. identify images, extract information from charts and plots and hand writing.



Setting up Llama



Setting Up Llama and Running a Simple Query

Learn how to set up and use the Llama language model for natural language processing tasks.



Activity

In this activity, we will explore the process of setting up the powerful Llama language model and running a simple query to generate a response. Llama is a cutting-edge large language model developed by Anthropic, capable of producing human-like text on a wide range of topics.



Prerequisites

Minimum 8 GB RAM

Ensure your system has at least 8 GB of RAM to provide sufficient memory for running Ollama with Llama 3.2.

At least 10 GB of free disk space

Make sure you have at least 10 GB of free storage space on your system to accommodate the installation and operation of Ollama with Llama 3.2.

Python 3.8 or newer

Ollama with Llama 3.2 requires Python version 3.8 or later to be installed on your system.

Administrator/root access

You will need administrative or root-root-level access to your system to install and configure Ollama with Llama Llama 3.2 properly.

Windows 10 or later

Ollama with Llama 3.2 is supported on supported on Windows 10 and later later versions of the operating system. system.

macOS 11 (Big Sur) or later

Ollama with Llama 3.2 is also supported supported on macOS 11 (Big Sur) and and later versions of the operating system.

Running the Code



This code example demonstrates how to use the Ollama library to generate a response from the Llama 3.2 language model. The system prompt sets the behaviour of the assistant, while the user prompt provides the input for the model to generate a response. The generated response is then printed to the console.



Installing Llama on Windows

Install Llama

Visit the ollama Downloads website, website, download and run the Windows Installer, and follow the prompts to complete the installation.

Verify Installation

Open the Command Prompt and run run the 'llama version' command to to verify the installation and check the the Llama version.

Start Ilama Server

In the Command Prompt, run the 'llama serve' command to start the Llama server.

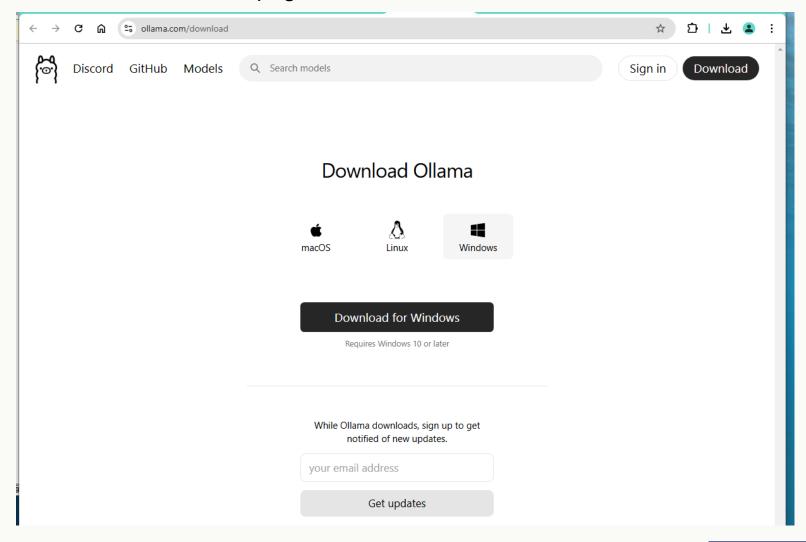
Install Python Package

Open the Command Prompt and run the 'pip install llama' command to install the llama Python package.



Downlading Ollama

The Ollama webpage should look like this







Windows Example of Ollama with llama3.2:1b from the command line

The Ollama download

- 이 OllamaSetup.exe
- Set up the software from the .exe and you will se it in your apps list
- Open the command line using the lower left search window (on Windows), by typing in 'command prompt'
- Once you have the command window, (as below), type in 'ollama run llama3.2:1b'

```
Command Prompt - ollama run llama3.2:1b — X

Microsoft Windows [Version 10.0.19045.5247]
(c) Microsoft Corporation. All rights reserved.

C:\Users\gjcra>ollama run llama3.2:1b
```

You will get various lines of code related pulling llama3.2

```
C:\Users\gjcra>ollama run llama3.2:1b

pulling manifest

pulling 74701a8c35f6... 100% E

pulling 966de95ca8a6... 100% E

cespulling fcc5a6bec9da... 100% E

pulling a70ff7e570d9... 100% E

pulling 4f659a1e86d7... 100% E

vonifying sba356 digest
```

Ask a question

- You will see 'Send a message'
- Write a prompt to to set the behaviour of the assistant, e.g. 'You are a helpful and curious AI assistant'

```
>>> You are a helpful and curious AI assistant
I can provide information on a wide range of topics, from science and history to
entertainment and culture. I can also help with language-related tasks, such as
writing and proofreading, and even generate text based on a prompt or topic.

If you have a specific question or topic in mind, feel free to ask me anything.
I'll do my best to provide a helpful and accurate response. What's on your mind
today?
```

Ask a question, for example "What is the colour of love"

```
>>> What is the colour of love?
That's a beautiful and thought-provoking question.
The color of love can be subjective and vary from person to person, culture to
culture, and context to context. However, I can offer some insights.
In many spiritual and poetic traditions, the color of love is often associated
with red or pink. Red symbolizes passion, energy, and warmth, which are all key
components of romantic love. Pink, on the other hand, represents tenderness,
nurturing, and vulnerability – qualities that are also present in loving
relationships.
Some people might associate the color of love with orange or coral, as these
colors can evoke feelings of excitement, enthusiasm, and playfulness – all
positive emotions that are often associated with loving connections.
Ultimately, the color of love is a personal and emotional experience. It's a
shade that each individual brings to their relationship, shaped by their unique
experiences, values, and perspectives.
Can you tell me more about what you mean when you think of the color of love?
What comes to mind for you?
 >>> Send a message (/? for help)
```



Running ollama in python

- One method is provided in the Word document associated with this workshop.
- Another method using Google colab is shown within the next few slides



Troubleshooting



Command not found

Ensure Llama is installed and added to your system's PATH environment variable.



Python package not installed

Run 'pip install llama' to install the required Python package for Llama.



Model not found

Run 'llama pull llama3.2' to download and install the Llama 3.2 model.



Server not running

Ensure the 'llama serve' command is running in a separate terminal to start the Llama

SERVER

If you encounter any of these issues, refer to the solutions to troubleshoot and get your Llama setup running



Example from within colab

- Open Google colab
- Start setting up a terminal

```
!pip install colab-xterm #this will install colab-xterm %load_ext colabxterm
```

%xterm

- This will launch a terminal
- Then at separate prompts(which look like /content#) you can type pip install ollama
- Type python
- The prompt will change to >>>
- Type import ollama



Example from within colab contd

```
# System prompt to set the behaviour of the assistant
system prompt = 'You are a helpful and curious AI assistant.'
# User prompt
user_prompt = 'What is the color of love?'
# Generate the response using the local LLM
response = ollama.generate(
 model='llama3.2',
  prompt=f'{system prompt}\n\nUser: {user prompt}\nAssistant:'
```

You should see some interesting output



Installing Llama on macOS

Install Llama

Visit the Llama Downloads website, download the macOS Installer (DMG), and open the downloaded file to drag the Llama icon to the Applications folder.

Verify Installation

Open the Terminal application and run the 'llama version' command to verify that the Llama installation was successful.

Start Llama Server

In the Terminal, run the 'llama serve' command to start the Llama server, which will allow you to interact with the Llama model.

Install Python Package

Open the Terminal and run the 'pip install llama' command to install the Llama Python package, which will enable you to use the Llama model in your Python scripts.



Additional Resources

Llama Documentation

Comprehensive documentation for documentation for Llama, covering covering installation, usage, and and advanced features.

• Llama Installer

Download the latest installer for for Llama, available for Windows, Windows, macOS, and Linux platforms.

Llama Support

Access the Llama support resources, including forums, community discussions, and direct direct support channels.

