Lisbon School of Economics & Management
Universidade de Lisboa

LISBOA
UNIVERSIDADE DE LISBOA

Generative AI, Ethics and Policy

Carlos J. Costa

- Defining GenAI
- Prompts and Prompt engineering
- Ethics and AI

Traditional Machine Learning

- Traditional ML:

$$X \rightarrow f(X) \rightarrow Y$$

- Example:
- **Input** *(X)*
  - income
  - age
  - employment
  - debt
- **Model** *(f)*
- **Output** *(Y)*
  - Probability of loan default = 0.17
- The objective is usually **prediction or classification**.



Generative AI

- Class of AI algorithms and models that are designed to generate new, original content.
- Gen AI learn the underlying patterns and structures in the data and can generate novel outputs.
- *Instead of being trained on specific examples and then making predictions or classifications*
- These models are particularly good at creating content that resembles or is similar to the data they were trained on.

Generative AI

- Generative AI:

$$Context \rightarrow Model \rightarrow Content$$

- Example:

  "Explain why interest rates affect investment."

  LLM

  Natural-language explanation

- The output is **generated content**, not simply a predicted class.

What can Generative AI generate?

- **Text**
  - reports
  - summaries
  - explanations
  - recommendations
- **Code**
  - Python
  - SQL
  - R
- **Images**
- **Audio**
- **Video**
- **Structured data**

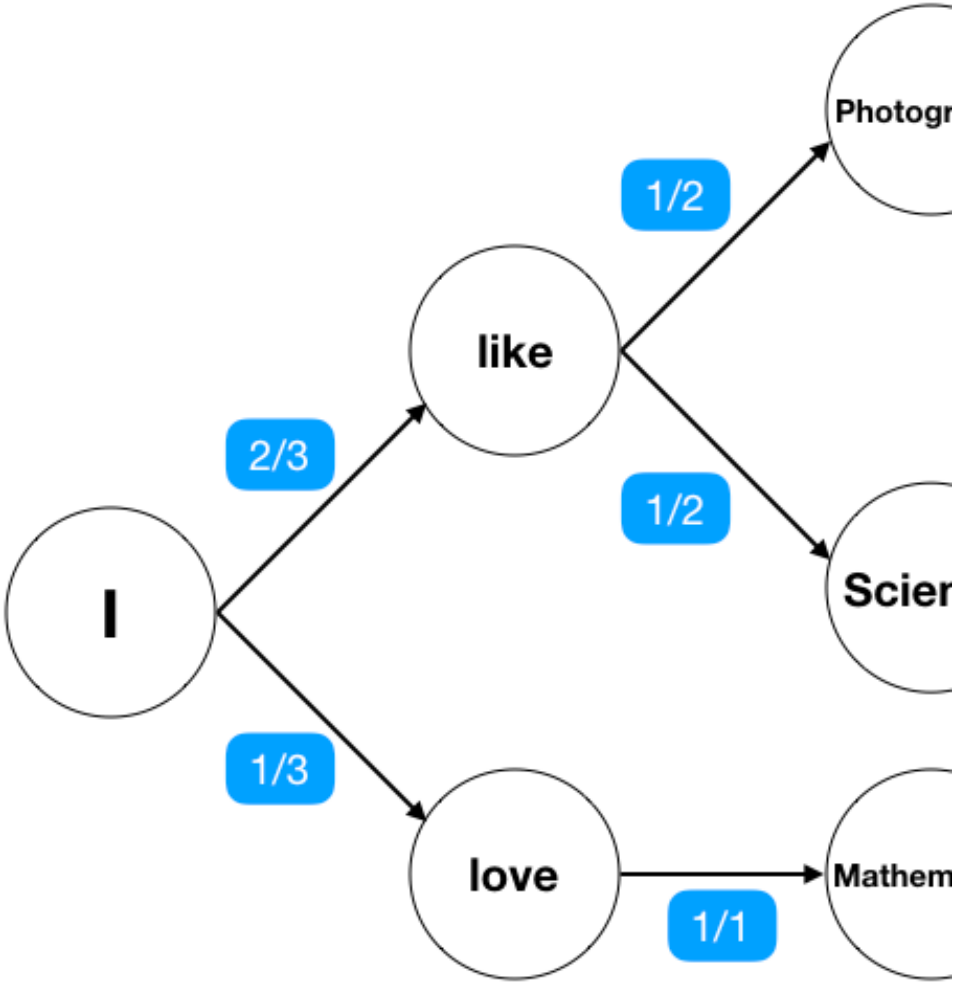
- **Generate new content based on learned patterns and the provided context.**

Generative AI

Guessing next word

Markov Chain

Training model



What is an LLM?

- A Large Language Model is a model trained on very large collections of text to model relationships between tokens.
- At a simplified level:
- Given the previous tokens, predict the next token.
- Example:

"The Portuguese economy grew by"

[next token]

- Possible continuation:

2.1%

Tokens

- LLMs do not process language exactly as humans do.
- Text is converted into **tokens**.

"Generative AI is changing business“ -> ["Generative", " AI", " is", " changing", " business"]

- Tokens may correspond to:
- words;
- parts of words;
- punctuation;
- symbols.
- **Tokenisation affects cost, context and model processing.**

Embeddings

- **An embedding is a numerical representation of a piece of information that captures aspects of its meaning.**
- Instead of representing text as words alone, an AI model represents it as a **vector of numbers**:

"Interest rates increased"

↓

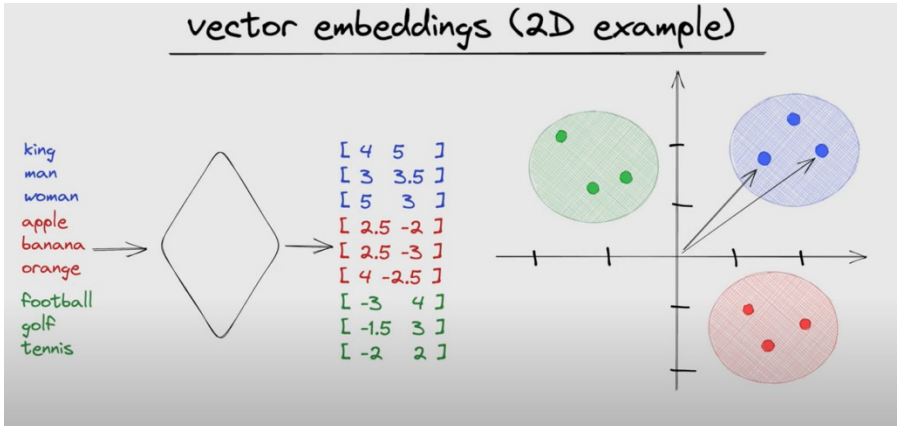
Embedding model

↓

[0.21, -0.47, 0.83, 0.15, ...]

↑

vector representation



- **Key idea:** texts with similar meanings tend to have embeddings that are **close together in vector space**.

Context

- An LLM does not simply receive the last sentence.
- It receives a **context**.

System instructions

+

User prompt

+

Previous messages

+

Retrieved information

+

Tools/results

↓

LLM

- The context strongly influences the output

Transformer architecture

- At a conceptual level:

Text
↓
Tokens
↓
Embeddings
↓
Transformer
↓
Attention
↓
Probability distribution
↓
Next token
↓
...
↓
Generated response

- The **Transformer architecture** is fundamental to modern LLMs

Attention

- Attention allows the model to represent relationships between tokens.
- Example:

"The company increased investment because **it** expected higher demand."

What does **it** refer to?

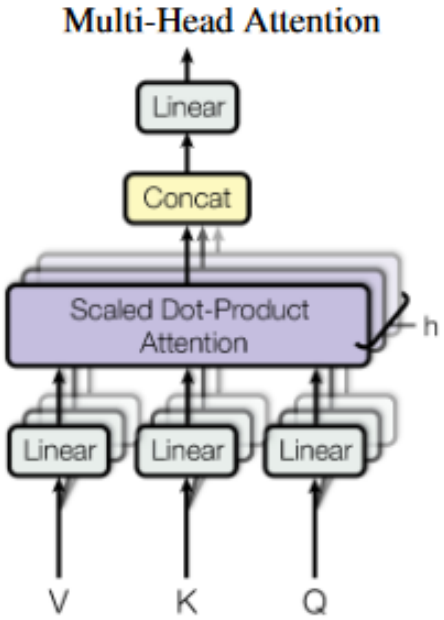
The model uses contextual relationships between tokens.

- The important idea:

Meaning is strongly dependent on context.

Transformer

- Deep learning architecture based on the multi-head attention mechanism





Attention Is All You Need

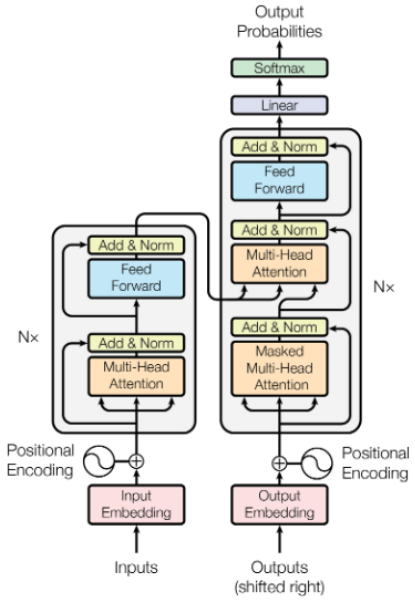
Ashish Vaswani* Google Brain avaswani@google.com
Noam Shazeer* Google Brain noam@google.com
Niki Parmar* Google Research nikip@google.com
Jakob Uszkoreit* Google Research usz@google.com
Llion Jones* Google Research llion@google.com
Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu
Łukasz Kaiser* Google Brain lukaszkaiser@google.com
Illia Polosukhin* ‡ illia.polosukhin@gmail.com

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.0 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature.

1 Introduction

Recurrent neural networks, long short-term memory [12] and gated recurrent [7] neural networks in particular, have been firmly established as state of the art approaches in sequence modeling and transduction problems such as language modeling and machine translation [29, 2, 5]. Numerous efforts have since continued to push the boundaries of recurrent language models and encoder-decoder architectures [31, 21, 13].



Vaswani, et al. (2017)

Why are LLMs not databases?

- Consider:
- "What was Portugal's GDP in 2025?"
- An LLM may generate a plausible number.
- But generation is not the same as retrieval from an authoritative database.
- Therefore:
  - **LLM knowledge ≠ verified data source**
  - This becomes particularly important in Economics and Finance

LLM The basic interaction



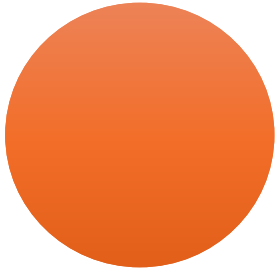
- But the output depends on:
  - model;
  - prompt;
  - context;
  - parameters;
  - available information.

Prompts

- “Prompts are your input into the AI system to obtain specific results.
- In other words, prompts are conversation starters: what and how you tell something to the AI for it to respond in a way that generates useful responses for you.
- After that, you can build a continuing prompt, and the AI will produce another response accordingly.
- It’s like having a conversation with another person, only in this case the conversation is text-based, and your interlocutor is AI.”

- https://mitsloanedtech.mit.edu/ai/basics/effective-prompts/

Prompt Engineering



- is the practice of designing, formulating, and refining input prompts to guide the behavior of large language models (LLMs) or other generative AI systems toward producing desired outputs.

A poor prompt

- Analyse Portugal.
- What does this mean?
  - GDP?
  - inflation?
  - unemployment?
  - investment?
  - public debt?
  - financial markets?
  - 2025?
  - 2026?
  - comparison with Europe?
- **The task is underspecified**

Effective Prompts

- Give **context** (set the role, scenario, or background)
- Be **specific** (clear instructions)
- Show **format**/structure
- Provide examples/**guidance**
- **Iterate**/refine
- Verify/**evaluate**

Prompt engineering

- A prompt can specify:
- **Task**
  - What should the model do?
- **Context**
  - What does it need to know?
- **Constraints**
  - What should it or should it not do?
- **Output**
  - What should the answer look like?
- **Evaluation**
  - What constitutes a good answer?

Example

- **Weak**
  - Analyse Portugal's economy.
- **Engineered**
  - Analyse Portugal's economic performance for an undergraduate economics report. Discuss GDP growth, inflation and unemployment. Use only the data provided. Do not invent values. Present the results in a table and then provide a 300-word interpretation.
- The second prompt defines the **problem more precisely**.

Prompt strategies

- zero-shot;
- few-shot;
- role/context prompting;
- explicit constraints;
- structured output;
- examples;
- decomposition;
- iterative prompting.

- The objective is not:
  - **"Write a very long prompt."**
- The objective is:
  - **"Provide the information and constraints necessary for the task."**

PROGRAMMING WITH LLM

LLM using

Prepare to download
LLM from Hugging face

!pip install -q
transformers accelerate



Hugging Face

LLM

```
from transformers import pipeline

class LLM:
    def __init__(self, model):
        self.model = model
        self.generator = pipeline(
            "text-generation",
            model=model
        )
    def generate(self, prompt):
        response = self.generator(
            prompt,
            max_new_tokens=150
        )
        return response[0]["generated_text"]
```

Using LLM

```
llm = LLM("Qwen/Qwen2.5-0.5B-Instruct")
answer = llm.generate(
    "Explain inflation to an economics student."
)
print(answer)
```

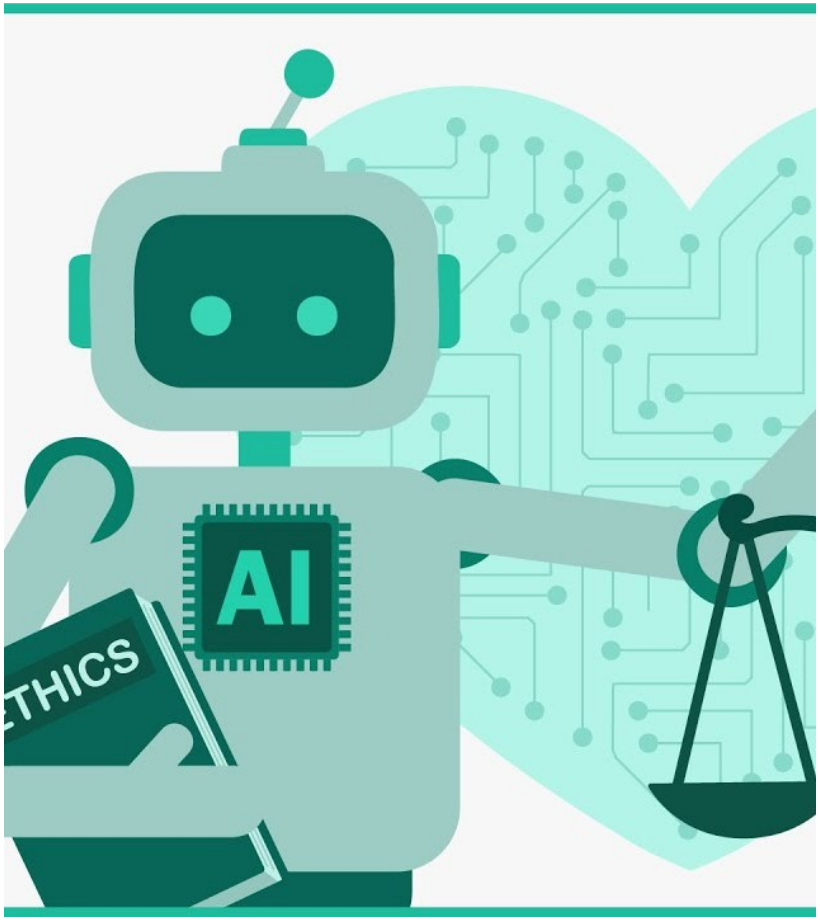
LLM parameters

- Important parameters include:
- **Temperature**
  - Controls randomness in generation.
- **Maximum tokens**
  - Controls output length.
- **Model**
  - Different models have different capabilities and costs.
- **System instructions**
  - Define behaviour/context.

Temperature

- Ask the same question several times.
  - **Temperature ≈ 0**
- More deterministic.
  - **Higher temperature**
- More variation.
- Question: Does higher variation mean better answers?
  - **No.**
  - It depends on the task.
- Creative writing and financial reporting have different requirements.

Ethics and AI



- Legal/Ethical
  - Hidden / Indirect Prompt Injection
  - Lawyers Citing Fake Cases Made up by AI
  - Deepfake Imagery Using AI
  - …

Why ethics?

- GenAI can affect:
  - people;
  - organisations;
  - markets;
  - employment;
  - information;
  - decisions.
- Therefore technical performance is not enough.

- We also ask:
- **What should the system be allowed to do?**

Main ethical dimensions

- **Fairness -** Does the system treat groups differently?
- **Privacy -** What personal information is processed?
- **Transparency -** Can users understand how AI is being used?
- **Accountability -** Who is responsible?
- **Human autonomy -** Can people meaningfully control important decisions?
- **Safety -** What happens when the system fails?

Hallucination

- An LLM may generate information that is:
  - plausible;
  - fluent;
  - specific;
  - completely wrong.
- For Economics and Finance this is particularly important.
- Example:
  - fabricated economic statistics
  - fabricated academic references
  - incorrect financial ratios
- **Technical response**
- **Retrieval + verification + evaluation**

Bias

- Bias can enter through:

Data - > Model -> Prompt – Application -> Decision

- Therefore:
  - AI bias is not necessarily a single "model problem".
  - It can emerge from the entire system.

Human agency

- **AI-assisted decision -** AI provides information; human decides.
- **AI-mediated decision -** AI substantially shapes the decision.
- **AI-autonomous action -** AI decides and acts with limited human intervention.
- The ethical implications change with the level of autonomy.

From ethics to policy

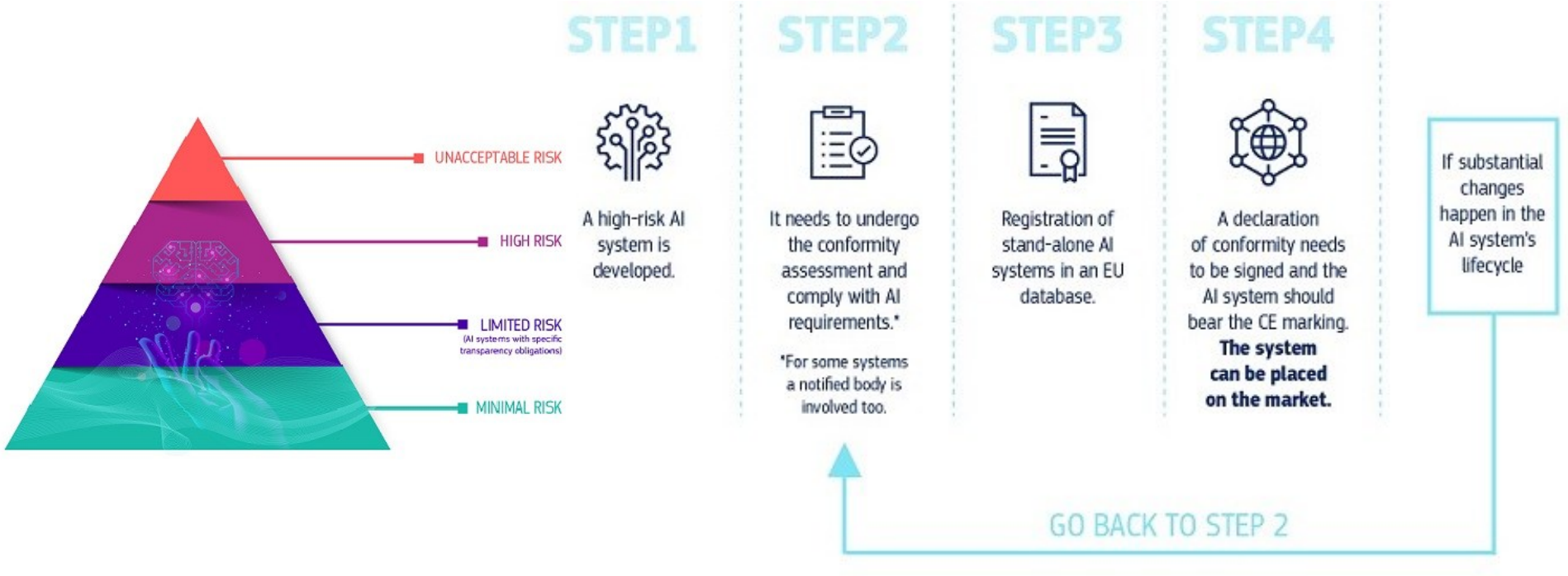
- **Ethics** - What should we do?
- **Governance** - How should organisations control AI?
- **Policy** - What rules should apply?

- These are related but different questions

EU AI Act

- AI Act

https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai



EU AI Act

- The EU AI Act uses a **risk-based regulatory framework**.
- Different AI applications face different obligations depending on their risk and characteristics.
- Relevant concepts include:
  - prohibited practices;
  - high-risk AI;
  - transparency obligations;
  - general-purpose AI;
  - systemic risk.
- The regulatory framework is being implemented progressively

General-purpose AI

- LLMs such as foundation models can be used in many downstream applications.
- This creates a distinction between:

Model provider

↓

AI application provider

↓

Organisation

↓

End user

- Questions of responsibility can therefore extend across the AI value chain

AI governance

- An organisation needs more than a model.
- A governance system can include:

AI Policy

↓

Risk Assessment

↓

Data Governance

↓

Testing

↓

Human Oversight

↓

Monitoring

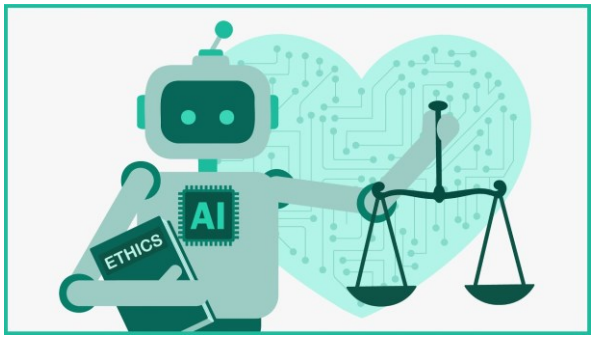
↓

Audit

Ethics and AI

Recommended usages

Management PhD

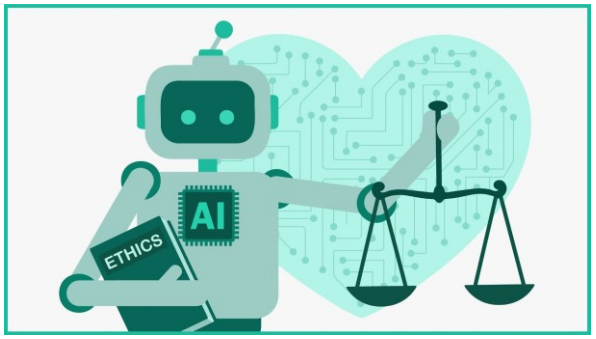


Situation	Recommended Use of GenAI
Academic writing	Improve style, textual coherence, and chapter organization. Suggest chapter structure and reformulate paragraphs.
Literature review	Support in structuring topics — with source validation.
Programming and data analysis	Suggest code, libraries, and tests — always with supervision and validation.
Idea exploration	Brainstorm hypotheses, methodological alternatives, counterarguments. Generate initial research questions.
Personalized tutoring	Explanations of statistical, epistemological, or computational concepts.
Bibliographic summary	Summarize articles, suggest authors or theories (with validation).
Language review	Translation or grammatical correction, especially in English.
Presentation preparation	Generate ideas for slides or communication structures.

Ethics and AI

No Recommended usages

Management PhD



Situation	Incorrect Use of GenAI
Submitting generated text without review	Violates authorship and integrity standards.
Creating non-existent results	Generating data that was not collected constitutes scientific fraud.
Citing fabricated sources	High risk of "hallucinations" — undermines credibility.
Avoiding reading articles	Replacing critical reading with AI reduces scientific rigor and the depth of knowledge developed.
Using AI as the dominant voice	Prevents the development of independent scientific thinking.
Omitting the use of AI	Violates the principle of academic transparency.
Assessments	Using GenAI to answer exams or assessments without authorization.
Omission of GenAI use	Failure to acknowledge the use of GenAI in documents where its contribution was significant.

References

- Belcic, I. (2024) *What is RAG (retrieval augmented generation)?* IBM https://www.ibm.com/think/topics/retrieval-augmented-generation
- Costa, C. J. (2024). Neural Networks: A Comprehensive Overview of Their History, Development, and Future in AI. *OAE – Organizational Architect and Engineer Journal*. https://doi.org/10.21428/b3658bca.13fccc0e
- Costa, C. J. (2025). Introduction to Machine Learning. *OAE – Organizational Architect and Engineer Journal*. https://doi.org/10.21428/b3658bca.35c216ce
- Costa, C. J. (2025). Generative AI Models: A Comprehensive Review. *OAE – Organizational Architect and Engineer Journal*. https://doi.org/10.21428/b3658bca.d5d1872f
- Costa, C. J. (2025). Prompt Engineering Strategies. *OAE – Organizational Architect and Engineer Journal*. https://doi.org/10.21428/b3658bca.210ec518
- Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., ... & Piao, Y. (2024). Deepseek-v3 technical report. *arXiv preprint https://doi.org/10.48550/arXiv.2412.19437*
- Marcondes, F.S., Gala, A., Magalhães, R., Perez de Britto, F., Durães, D., Novais, P. (2025). Using Ollama. In: Natural Language Analytics with Generative Large-Language Models. SpringerBriefs in Computer Science. Springer, Cham. https://doi.org/10.1007/978-3-031-76631-2_3
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Bibliography

Aparicio, J. T., de Sequeira, J. S., & Costa, C. J. (2021). Emotion analysis of portuguese political parties communication over the covid-19 pandemic. In 2021 16th Iberian Conference on Information Systems and Technologies (CISTI) (pp. 1-6). IEEE.

Aparicio, J. T., Romao, M., & Costa, C. J. (2022). Predicting Bitcoin prices: The effect of interest rate, search on the internet, and energy prices. In 2022 17th Iberian Conference on Information Systems and Technologies (CISTI) (pp. 1-5). IEEE.

Aparicio, S., Aparicio, J. T., & Costa, C. J. (2019,). Data Science and AI: trends analysis. In 2019 14th Iberian Conference on Information Systems and Technologies (CISTI) (pp. 1-6). IEEE.

Arriaga, A., & Costa, C. J. (2023, May). Modeling and Predicting Daily COVID-19 (SARS-CoV-2) Mortality in Portugal: The Impact of the Daily Cases, Vaccination, and Daily Temperatures. In Proceedings of International Conference on Information Technology and Applications: ICITA 2022 (pp. 275-285). Singapore: Springer Nature Singapore.

Costa, C. J. (2023). Artificial Intelligence in Management: Enhancing Planning, Organization, Direction, and Control. *OAE – Organizational Architect and Engineer Journal*. https://doi.org/10.21428/b3658bca.b55ee6ae

Costa, C. J. (2024). Neural Networks: A Comprehensive Overview of Their History, Development, and Future in AI. *OAE – Organizational Architect and Engineer Journal*. https://doi.org/10.21428/b3658bca.13fccc0e

Costa, C. J. (2025). Introduction to Machine Learning. *OAE – Organizational Architect and Engineer Journal*. https://doi.org/10.21428/b3658bca.35c216ce

Costa, C. J., & Aparicio, J. T. (2020). POST-DS: A methodology to boost data science. In 2020 15th Iberian Conference on Information Systems and Technologies (CISTI) (pp. 1-6). IEEE.

Costa, C. J., & Aparicio, M. (2023). Applications of Data Science and Artificial Intelligence. Appl. Sci, 13, 9015.

Costa, C., Aparicio, M., & Aparicio, J. (2021). Sentiment analysis of portuguese political parties communication. In Proceedings of the 39th ACM International Conference on Design of Communication (pp. 63-69).

Costa, C. J., Aparicio, J. T., & Aparicio, M. (2024). Socio-Economic Consequences of Generative AI: A Review of Methodological Approaches. *arXiv preprint arXiv:2411.09313*

Costa, C. J., Aparicio, M., Aparicio, S., & Aparicio, J. T. (2024). The Democratization of Artificial Intelligence: Theoretical Framework. *Applied Sciences*, *14*(18), 8236. https://doi.org/10.3390/app14188236

Custódio, J. P. G., Costa, C. J., & Carvalho, J. P. (2020). Success prediction of leads–A machine learning approach. In 2020 15th Iberian Conference on Information Systems and Technologies (CISTI) (pp. 1-6). IEEE.

Hajishirzi, R., & Costa, C. J. (2021). Artificial Intelligence as the core technology for the Digital Transformation process. In 2021 16th Iberian Conference on Information Systems and Technologies (CISTI) (pp. 1-6). IEEE.

Ortiz, F. C. M., & Costa, C. J. (2020, June). RPA in Finance: supporting portfolio management: Applying a software robot in a portfolio optimization problem. In 2020 15th Iberian Conference on Information Systems and Technologies (CISTI) (pp. 1-6). IEEE.

Piteira, M., Aparicio, M., & Costa, C. J. (2019, June). Ethics of artificial intelligence: Challenges. In 2019 14th Iberian Conference on Information Systems and Technologies (CISTI) (pp. 1-6). IEEE.

Samadani, S., & Costa, C. J. (2021). Forecasting real estate prices in Portugal: A data science approach. In 2021 16th Iberian Conference on Information Systems and Technologies (CISTI) (pp. 1-6). IEEE.

Vaswani, A.; Shazeer, N.; Parmar, N. Uszkoreit, J.; Jones, L.; Gomez, A; Kaiser, Ł; Polosukhin, I (2017). "Attention is All you Need" Advances in Neural Information Processing Systems. Curran Associates, Inc. 30.

Veiga, M., & Costa, C. J. (2024). Ethics and Artificial Intelligence Adoption. *arXiv preprint arXiv:2412.00330*.


Q&A