Large Language Models and AI in Finance

Athanasios Sakkas, AUEB

Learning objectives

By the end of this lecture, you should be able to explain:

1. What a large language model (LLM) is in plain language.
2. Why LLMs work surprisingly well despite predicting the next word.
3. How LLMs can be used in finance: news, trading, banking, risk, compliance.
4. Why AI output can be useful but also dangerous.
5. Why better prediction does not automatically mean profitable trading.

Takeaway

AI in finance is not mainly about replacing people. It is about lowering the cost of processing information.

The big idea: finance is an information industry

Finance is built on information:

- Investors read news, filings, earnings calls and macro announcements.
- Banks process customer documents, credit histories and transaction data.
- Regulators monitor reports, suspicious activities and disclosures.

Information → Beliefs → Decisions → Prices

LLMs matter because they make the first step - processing information - faster and cheaper.

What is artificial intelligence?

Artificial intelligence means computer systems that perform tasks normally associated with human intelligence.

Examples:

- Recognising patterns
- Understanding language
- Making predictions
- Solving problems
- Generating text, images or code

Traditional AI vs modern AI

Older AI often used manually written rules. Modern AI usually learns patterns from data.

In finance, this shift is important because financial data are too complex for simple hand-written rules.

From rule-based AI to machine learning

Rule-based system

Example: "Reject loan if income is below £20,000."

- Easy to explain
- Brittle
- Cannot easily learn new patterns

Machine learning system

Example: Estimate default risk from thousands of past borrowers.

- Learns from data
- Adapts to complex patterns
- May be harder to explain

A simple machine-learning view is:

$$\hat{y} = f(X),$$

where X is information and $\hat{y}$ is the predicted outcome.

In finance, $\hat{y}$ might be default risk, fraud risk, expected return or sentiment.

What makes generative AI different?

Traditional machine learning usually predicts a label or a value:

- Fraud or not fraud
- Default or no default
- Positive or negative sentiment
- stock returns

Generative AI creates new content:

- Summaries
- Emails
- Code
- Reports
- Explanations

Takeaway

A generative AI model is not just a classifier or a prediction model. It can produce a written answer that looks like something a human analyst might write.

What is a large language model?

A **large language model** is a model trained on huge amounts of text to predict and generate language.

Examples of tasks:

- Summarise an earnings call
- Translate a document
- Explain a financial concept
- Draft a market commentary
- Write Python code for data analysis

Simple definition

An LLM is a very large statistical model that learns patterns in language and uses those patterns to generate useful text.

Why do LLMs work? The simple version

The core task is surprisingly simple:

Predict the next token.

Example:

The central bank raised interest rates because inflation was ________

Possible completions:

- high
- falling
- irrelevant
- delicious

A well-trained model assigns high probability to words that make sense in context.

Takeaway

By learning to predict words at massive scale, LLMs learn grammar, facts, style, associations and some reasoning patterns.

Tokens: the building blocks of LLMs

LLMs do not directly see words. They see **tokens**.

Example sentence:

```
Interest rates increased rapidly.
```

It may be broken into tokens such as:

```
[Interest] [rates] [increased] [rapidly] [.]
```

Each token is converted into numbers. The model then learns relationships among these numbers.

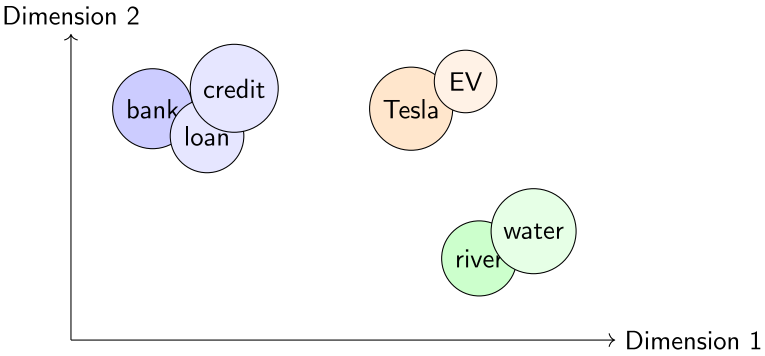
Finance intuition

Just as finance turns a firm into numbers such as revenues, margins and beta, an LLM turns language into numbers it can process.

Embeddings: meaning as geometry

An **embedding** represents a word, sentence or document as a vector of numbers.

Words with related meanings are placed closer together.



Takeaway

Embeddings allow computers to search by meaning, not only by exact words.

Why context matters: the word "bank"

The same word can mean different things.

Finance meaning

"The bank approved the mortgage."

Here, bank means a financial institution.

Geography meaning

"The fisherman sat on the river bank."

Here, bank means land beside a river.

Modern LLMs use surrounding words to infer the correct meaning.

This is why they are much better than old dictionary methods for financial language.

Attention: how the model decides what matters

A transformer model uses **attention**: it looks across the sentence and gives more weight to important words.

Example:

> "Apple reported strong earnings, but investors worried about weak guidance."

A simple keyword model might see "strong earnings" and call this positive.

An LLM can understand that "weak guidance" may matter more for the stock price.

Takeaway

Attention helps LLMs interpret language in context. This is essential in finance, where words such as "liability", "risk" and "guidance" depend on context.

Why the simple next-word task becomes powerful

Language contains much more than grammar. It contains:

- facts about the world
- economic relationships
- cause-and-effect explanations
- reasoning patterns
- professional writing styles

When an LLM is trained on a very large text corpus, it learns many of these patterns indirectly.

Analogy

A student learns finance by reading textbooks, cases, annual reports and news. An LLM learns patterns from a much larger amount of text, but without human understanding in the same sense.

Important distinction: plausible is not always true

An LLM does not directly optimise truth. It optimises likely language.

LLM objective: choose likely next tokens

Finance objective: make correct decisions under uncertainty

This gap explains why LLMs can:

- sound confident but be wrong
- invent citations or facts
- miss recent information
- misunderstand specialist context

Takeaway

Treat an LLM as an intelligent assistant, not as a source of guaranteed truth.

Hallucination: the main risk for finance students

A **hallucination** is a confident answer that is false or unsupported.

In finance, hallucinations are dangerous because decisions have real costs.

Examples:

- A fake accounting rule
- A non-existent regulation
- A wrong number from an annual report
- A fabricated market event

Question for Finance Students

Would you let an LLM approve a loan, place a trade, or file a regulatory report without human verification? Why or why not?

Prompt engineering: asking better questions

A prompt is the instruction given to the model.

Better prompts usually provide:

1. Role: "Act as an equity analyst."
2. Context: company, sector, time period, documents.
3. Task: what exactly should be produced.
4. Format: table, bullets, memo, score.
5. Evidence: cite the source or quote the relevant sentence.

Bad prompt

"Is this company good?"

Better prompt

"Using the attached earnings report, list three positive and three negative signals for equity investors. Separate facts from interpretation."

Zero-shot and few-shot prompting

Zero-shot

The model receives the task only.

Example: "Summarise this earnings report."

Few-shot

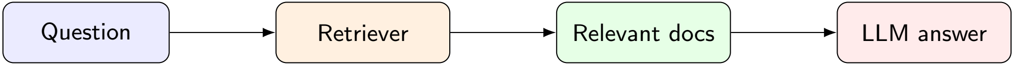
The model receives examples of the desired answer.

Example: "Here is one good earnings summary. Now summarise this new report in the same style."

Few-shot prompting is useful in finance because firms want consistent report formats, risk labels and investment memos.

Retrieval-augmented generation (RAG)

LLMs have limited knowledge of private and recent information. RAG adds external documents at the time of the query.



Finance examples:

- Search SEC filings before answering an investment question.
- Search internal bank policy before answering a compliance question.
- Search loan documents before answering a covenant question.

RAG workflow: retrieval plus generation

Step 1: Query representation

- The user submits a question.
- The query is converted into an embedding vector.
- Similarity search identifies economically relevant documents.

Step 2: Retrieval

- The retriever searches and relevant passages are ranked using vector similarity.

Step 3: Prompt augmentation

- Retrieved text is appended to the prompt.
- The LLM now conditions on external information rather than relying only on training data.

Step 4: Generation

- The final response becomes more grounded and up-to-date.