UnCodemy

Generative AI Notes PDF

Generative AI creates new content such as text, images or code by learning patterns from large amounts of data. These notes explain large language models, tokens, prompts, context windows, retrieval-augmented generation, fine-tuning, evaluation, common risks like hallucination and responsible use, in plain language for beginners.

Key takeaways

- Large language models predict the next token, which is why they can sound fluent and still be wrong.
- Prompts, context and retrieval decide most of the answer quality in applications.
- RAG adds your own documents to the prompt at question time instead of retraining the model.
- Evaluation with a fixed test set is needed to know if a change helped.
- Hallucination, privacy and bias are real risks that need design and review.

What is generative AI notes?

Generative AI is a branch of artificial intelligence in which models learn patterns from large datasets and then generate new content, such as text, images, audio or code, in response to a prompt.

How do large language models work?

A large language model (LLM) is trained on very large amounts of text to predict the next piece of text, called a **token**, from the tokens before it. A token is often part of a word. After training on general text, many models are further tuned to follow instructions and to be helpful and safe. The model does not look up facts in a database; it produces likely text from patterns it learned, which explains both its fluency and its errors.

What is a context window?

The context window is the amount of text, measured in tokens, the model can consider at once, including your prompt, any documents you add and its reply. If the input is longer than the window, parts must be left out or summarised. Longer prompts also cost more and take longer.

What makes a good prompt?

State the task, give context, say the format and add constraints. Examples help: showing two input and output pairs often improves consistency. Ask for step-by-step reasoning only when the task

needs it and always check results. Prompts are a design tool, and they cannot promise a correct answer.

What is retrieval-augmented generation?

RAG finds relevant passages from your own documents and adds them to the prompt so that the model answers using that material. The parts are: split documents into chunks, turn chunks into embeddings (numbers that capture meaning), store them in a vector database, retrieve the closest chunks for a question and ask the model to answer from them with citations. RAG is often cheaper and easier to update than retraining.

When should I fine-tune?

Fine-tuning trains a model further on your examples to change its style, format or behaviour on a narrow task. It does not reliably add new facts. Try prompting and RAG first; consider fine-tuning when you need consistent format or tone at scale and have good-quality examples.

What are agents?

An agent is a system where a model can decide steps, call tools such as a search or a calculator and loop until a task is done. Agents are powerful but harder to test, so add limits, logging and human approval for risky actions.

How do I evaluate a generative AI system?

Create a set of representative questions with expected answers or scoring rules. Measure accuracy, groundedness (is the answer supported by the sources), safety and cost. Rerun the set after every change to a prompt, model or retrieval setting. Human review is important because automatic scores can miss problems.

What are the main risks?

Risk	What it means	Mitigation
Hallucination	Confident but wrong output	Retrieval, citations, human checks
Privacy	Sensitive data in prompts	Data rules, redaction, approved tools
Bias	Unfair or skewed output	Testing across groups, review
Prompt injection	Hidden instructions in content	Input checks, limited tool rights
Copyright	Use of protected material	Policy, licensed data, review

What does responsible use look like?

Tell users when AI is used, keep humans accountable for decisions that affect people, avoid putting confidential data into public tools, test for harm and keep records of changes. Follow your organisation's policy and the law in your region.

What should I learn next?

Practise with a small project, such as a question-answering tool over a few documents, and write an evaluation set for it. Then explore agents and tool use.

What is the difference between traditional AI and generative AI?

Traditional machine learning models usually predict a label or a number from data, such as whether a transaction is fraud. Generative models produce new content, such as a paragraph or an image. Both learn from data, and many real systems use both: a classifier to route a request and a generative model to write the answer.

What are multimodal models?

Multimodal models accept or produce more than one kind of data, such as text and images together. They can describe a picture, read a chart or answer a question about a document that contains tables. The same cautions apply: check the output, because errors in reading images or numbers are possible.

How do I begin a first project?

Choose a small, safe task, for example answering questions about a short policy document. Write ten test questions with expected answers, build a simple retrieval step, add the retrieved text to the prompt and measure how many answers are correct and supported by the text. Improve one thing at a time, and keep a record of every change.

Frequently asked questions

What is generative AI?

It is AI that learns patterns from large datasets and generates new content such as text, images, audio or code from a prompt.

What is a token?

A token is a small piece of text, often part of a word, that a language model reads and predicts.

What is the difference between RAG and fine-tuning?

RAG adds relevant documents to the prompt at question time, while fine-tuning changes the model with extra training examples.

Why do language models hallucinate?

They generate likely text from patterns rather than looking up verified facts, so they can produce fluent but wrong statements.

Is generative AI safe to use at work?

It can be, with clear rules about data, human review of outputs and approved tools. Follow your organisation policy.

Sources and further reading

- OpenAI documentation
- Anthropic documentation
- Hugging Face documentation
- NIST AI Risk Management Framework

About this PDF

These PDFs are study material for learning and revision. They are not a certificate, a degree or a promise of a job, and technology changes, so check official documentation for the latest details. Spot an error? Write to info@uncodemy.com and the trainer team will review it.

More free resources and courses: uncodemy.com | info@uncodemy.com | +91 98183 66550