

GenAI Learning Links

What is Generative AI?

- **Start Here**
 - Stanford UIT Gen AI Resources:
 - [GenAI Overview Page](#)
 - [Responsible AI Page](#)
 - [GenAI Topics and Services List](#) (Table of Contents for AI at Stanford)
- **Stanford News and Training Resources**
 - [AI News at Stanford](#)
 - [UIT Tech Training AI Demystified site](#)

How AI Works

- [What Is ChatGPT Doing ... and Why Does It Work? - Stephen Wolfram](#)
- [Generative AI for Everyone | Coursera.org](#)
- [Generative AI with Large Language Models | Coursera](#)
 - A lengthy and in-depth course (self-paced, up to 16 hours) that covers, among other things, LLM training and fine-tuning. It also provides an AI lifecycle and application model, which is a valuable framework for thinking about building on LLMs.
- [Amplify GenAI - Vanderbilt University's Enterprise Open Source Generative AI Platform](#) (section begins at 8:06)
-  [Generative AI in a Nutshell - how to survive and thrive in the age of AI](#)
 - very engaging overview of LLMs and the basic concepts of RAG
-  [Deep Dive into LLMs like ChatGPT](#) by Andrej Karpathy
 - EXCELLENT explainer on how LLMs are trained and how they function. Reserve some time for this one.
- [LLM Temperature: How It Works and When You Should Use It](#)
- [LangChain for Beginners](#) - LangChain for LLM Application Development

Prompting Basics

- [Prompt Engineering Guide](#) (Kaggle)
- Ethan Mollick's prompt library <https://www.moreusefulthings.com/>
- Stanford GSB External Relations team [Prompt Guide](#)

Stanford Online Courses

The short courses on deeplearning.ai are free, each one lasting about an hour. They are presented with side-by-side video and Jupyter notebooks, complete with embedded free API access.

- [Stanford CS25: V2 | Emergent Abilities and Scaling in LLMs](#)
- [Stanford CS25: V3 | Recipe for Training Helpful Chatbots](#)
- [Stanford CS25: V4 | Jason Wei & Hyung Won Chung of OpenAI](#)

- [Stanford CS25: V3 I Beyond LLMs: Agents, Emergent Abilities, Intermediate-Guided Reasoning, BabyLM](#)
- [Stanford CS25: V3 I Retrieval Augmented Language Models](#)

Executive Learning Resources

-  Executive Training Courses (Generative AI)
- MIT Management Executive Education: [ARTIFICIAL INTELLIGENCE: IMPLICATIONS FOR BUSINESS STRATEGY](#)
-  Generative AI in a Nutshell - how to survive and thrive in the age of AI

DeepLearning.ai Courses

The short courses on deeplearning.ai are free, each one lasting about an hour. They are presented with side-by-side video and Jupyter notebooks, complete with embedded free API access.

Here are a few recommended learning paths:

- Learn RAG / LangChain
 - [LangChain for LLM Application Development - DeepLearning.AI](#)
 - [LangChain: Chat with Your Data - DeepLearning.AI](#)
 - [Build LLM Apps with LangChain.js - DeepLearning.AI](#)
 - [Building and Evaluating Advanced RAG Applications - DeepLearning.AI](#)
- Finetuning Models
 - [Finetuning Large Language Models - DeepLearning.AI](#) (see more below)
- Semantic Kernel
 - [How Business Thinkers Can Start Building AI Plugins With Semantic Kernel - DeepLearning.AI](#)
- Other
 - [Prompt Engineering with Llama 2 - DeepLearning.AI](#)
 - [Serverless LLM apps with Amazon Bedrock - DeepLearning.AI](#)

AI Agents

- [Introduction to LLM Agents | NVIDIA Technical Blog](#)
- [Building Your First LLM Agent Application | NVIDIA Technical Blog](#)
- [OpenAI Build Hour: Function Calling](#)
- DeepLearning.ai
 - [AI Agentic Design Patterns with AutoGen - DeepLearning.AI](#)
 - [Multi AI Agent Systems with crewAI - DeepLearning.AI](#)
- [OpenAI's Agents SDK and Anthropic's Model Context Protocol \(MCP\)](#)
 - Commentary on OpenAI Agents API

Model Fine Tuning

In-depth tutorials on how to fine-tune an LLM using a custom dataset.

- Tutorials
 - [How To Fine-Tune the Alpaca Model For Any Language | ChatGPT Alternative | by Martin Thissen | Medium](#)
 - [Fine-tuning Alpaca and LLaMA: Training on a Custom Dataset | MLExpert - Crush Your Machine Learning interview](#)
 - [Fine-Tuning LLaMA 2: A Step-by-Step Guide to Customizing the Large Language Model | DataCamp](#)
 - [Finetuning Large Language Models - DeepLearning.AI](#)
 - https://mlabonne.github.io/blog/posts/Fine_Tune_Your_Own_Llama_2_Model_in_a_Colab_Notebook.html
 - [How to fine-tune GPT-3.5 Turbo | Complete Crash Course Uses OpenAI libraries.](#)
- Reference
 - [LoRA](#) – Hugging Face Concept Guide

OpenAI

- [Pydantic](#) – useful to know for many python frameworks including OpenAI Agents
- [Create a Custom AI Assistant + API in 10 Mins](#)

Anthropic

- [Anthropic Cookbook](#)

Compound Systems Development

Read [The Shift from Models to Compound AI Systems](#)

(<https://bair.berkeley.edu/blog/2024/02/18/compound-ai-systems/>)

- **Language Model Programming** - frameworks and tools that allow developers to build applications by composing calls to AI models and other components in a programmatic way
 - **Output Controls** - techniques and tools to make the outputs of large language models more trustworthy, reliable, secure and aligned with specific goals
 - [Guardrails AI](https://www.guardrailsai.com/) (<https://www.guardrailsai.com/>) an open source platform that aims to make large language models more trustworthy, reliable and secure through crowdsourced mitigations
 - [Outlines](https://outlines-dev.github.io/outlines/) (<https://outlines-dev.github.io/outlines/>) “a Python library that allows you to use Large Language Model in a simple and robust way (with structured generation). It is built by .txt, and is already used in production by many companies”
 - [LMQL](https://lmql.ai/) (<https://lmql.ai/>) - a concept that allows mixing prompts in natural language with code to control and constrain language model outputs

- [SGLang](https://arxiv.org/pdf/2312.07104.pdf) (<https://arxiv.org/pdf/2312.07104.pdf>) an efficient programming language for large language models that simplifies LLM programming and optimizes execution through parallelism, caching, and compilation techniques.
- **Agent Frameworks** - Platforms and tools that allow users to create autonomous AI agents that leverage large language models to perform tasks based on specific goals
 - [AgentGPT](https://autogpt.net/) (<https://autogpt.net/>) a platform that allows users to create and deploy autonomous AI agents tailored to specific goals, leveraging large language models like GPT-3.5 and GPT-4 to perform tasks autonomously
 - [babyagi](https://github.com/yoheinakajima/babyagi) (<https://github.com/yoheinakajima/babyagi>) a simple, pared-down implementation of an autonomous AI agent, designed to be easy to understand and build upon, using OpenAI's language models
 - [AutoGPT](https://autogpt.net/) (<https://autogpt.net/>) an open-source application that creates fully autonomous AI agents using GPT-3.5 or GPT-4, capable of self-prompting, reviewing their work, and accessing external resources to accomplish user-defined goals and objectives
- **Inference Strategies** - techniques used during the inference (prompting) stage to generate better outputs by leveraging multiple calls to models and tools
 - [Chain-of-Thought](https://arxiv.org/pdf/2201.11903.pdf) (<https://arxiv.org/pdf/2201.11903.pdf>) elicits step-by-step reasoning from large language models, enabling them to solve challenging math, commonsense, and symbolic reasoning problems more effectively as model scale increases
 - [Self-Consistency](https://arxiv.org/pdf/2203.11171.pdf) (<https://arxiv.org/pdf/2203.11171.pdf>) decoding strategy that improves chain-of-thought reasoning in language models by sampling diverse reasoning paths and selecting the most consistent final answer, leveraging the intuition that correct reasoning processes tend to converge on the same answer
 - [WikiChat](https://arxiv.org/pdf/2305.14292.pdf) (<https://arxiv.org/pdf/2305.14292.pdf>) a multi-component system that combines a large language model with information retrieval to produce factual and up-to-date responses in open-domain conversations, addressing the hallucination issues of language model
 - [Retrieval Augmented Generation \[RAG\]](https://arxiv.org/pdf/2005.11401.pdf) (<https://arxiv.org/pdf/2005.11401.pdf>) an AI framework that retrieves relevant information from external knowledge sources to augment and improve the responses generated by large language models
- **Component Libraries** - libraries that provide reusable components for building compound AI systems
 - [LangChain](https://www.langchain.com/) (<https://www.langchain.com/>) a Python library that allows developers to build applications by composing different language

models, data sources, and other components into a single workflow or "chain"

- [LlamaIndex](https://www.llamaindex.ai/) (https://www.llamaindex.ai/) an open-source data structuring and indexing library that allows developers to create custom data indexes optimized for question-answering and other language model applications
- **AI Gateways** - Systems that optimize the performance and cost of AI applications by intelligently routing requests to different models and services based on the input and desired quality/cost tradeoff
 - [OpenRouter](https://openrouter.ai) - (https://openrouter.ai) a tool that helps you integrate multiple NLP APIs in your application. It provides a single API endpoint that you can use to call multiple NLP APIs
 - [Databricks Model Serving](https://docs.databricks.com/en/generative-ai/external-models/index.html) - (https://docs.databricks.com/en/generative-ai/external-models/index.html) enables organizations to securely serve and manage external large language models from providers like OpenAI and Anthropic through a unified interface
 - [FrugalGPT](https://arxiv.org/abs/2305.05176) - (https://arxiv.org/abs/2305.05176) a framework that reduces the cost of using large language models from providers like OpenAI and Anthropic by adaptively selecting prompts, approximating expensive models, and cascading multiple models based on query difficulty.
 - [Martian Model Router](https://withmartian.com/products/model-router) - (https://withmartian.com/products/model-router) dynamically selects the optimal language model for each request, enabling higher performance than any single model like GPT-4 while reducing costs

See also:

- ▶ What's next for AI agentic workflows ft. Andrew Ng of AI Fund
- ▶ Adding Agentic Layers to RAG

Machine Learning

- [Discover the latest machine learning / AI courses on YouTube](#) – this is a great compilation of free lectures from various sources, including MIT, Caltech, and Stanford
- <https://www.kaggle.com/learn> – an excellent collection of data science courses ranging from beginning Python to TensorFlow to NLP

Misc Blog Posts

- [Meta Llama2 prompting overview](#)
 - this is a great doc that a) gives an overview of llama and how to access it, b) shows how to interact with it via langchain, and c) dives into detailed prompting advice
- [Getting started with Llama 2 - AI at Meta](#)
 - a more in-depth look at how to work with Llama2. Very readable, very informative
- Useful Ollama / Langchain Blog post:
 - [Run Large Language Models with Ollama and AWS Lightsail for Research](#)

Podcasts

- [The Practical AI podcast |> Changelog](#)
- [The Artificial Intelligence Podcast](#)
- [The AI Podcast](#)
- [Data Skeptic](#)
- [Lex Fridman Podcast](#)

Newsletters

- Daily General AI news
 - [The Rundown AI](#)
 - [The Neuron](#)
- Deeper Dives
 - [Data Machina](#) – a weekly newsletter that showcases and summarizes the latest research in AI. Recent topics include 3D scene generation, advanced RAG techniques, and classifying and mitigating hallucinations.

Research Papers

- [\[1706.03762\] Attention Is All You Need](#)
- [\[2305.14314\] QLoRA: Efficient Finetuning of Quantized LLMs](#)
- [\[2307.03172\] Lost in the Middle: How Language Models Use Long Contexts](#)