



**INSTITUTO
FEDERAL**
Sergipe

TÓPICOS ESPECIAIS I

AULA 00

**DA MÁQUINA QUE
CALCULA À MÁQUINA
QUE CONVERSA**

2026

Prof Ângelo Santana

Clique nos tópicos para navegar

01	O QUE É INTELIGÊNCIA ARTIFICIAL?	4
02	UMA BREVE HISTÓRIA DA INTELIGÊNCIA ARTIFICIAL	4
03	POR QUE A IA AVANÇOU TANTO NAS ÚLTIMAS DÉCADAS?	6
04	O QUE ACONTECE QUANDO VOCÊ CONVERSA COM UM LLM?	7
05	EXPERIMENTO: VOCÊ TAMBÉM PREVÊ LINGUAGEM	8
06	O MAPA DA DISCIPLINA	8
07	ONDE QUEREMOS CHEGAR: O PROJETO INTEGRADOR	9

08	O QUE UMA IA GENERATIVA SABE, E O QUE ELA NÃO GARANTE?	9
09	DEBATE: SE A IA GERA CÓDIGO, POR QUE APRENDER PROGRAMAÇÃO?	10
10	RESUMO DA AULA 00	10
11	EXERCÍCIOS DE FIXAÇÃO	10
12	GLOSSÁRIO ESSENCIAL	12
13	PARA A PRÓXIMA AULA	13

Como usar esta apostila

Esta apostila acompanha a Aula 00 da disciplina Tópicos Especiais. Ela foi pensada para funcionar de três maneiras: como leitura prévia, como material de acompanhamento durante a aula e como referência para revisão posterior.

Ideia-guia da aula

Antes de aprender a construir aplicações com IA, precisamos enxergar o território: de onde a área veio, por que os modelos generativos ganharam tanta importância e quais peças compõem um sistema moderno baseado em LLMs.

Objetivos de aprendizagem

- Distinguir Inteligência Artificial, Machine Learning, Deep Learning e IA Generativa.
- Reconhecer os principais marcos históricos que levaram aos modelos atuais.
- Compreender, em nível introdutório, por que dados, hardware, algoritmos e escala impulsionaram a IA moderna.
- Explicar intuitivamente como um modelo de linguagem gera texto token a token.
- Visualizar a estrutura da disciplina: prompts, APIs, embeddings, RAG, ferramentas, agentes, avaliação, segurança e LLMOps.
- Desenvolver uma postura crítica: utilizar IA sem confundir fluência linguística com garantia de verdade ou compreensão humana.

Pergunta de abertura

Você já utilizou Inteligência Artificial hoje?

Antes de pensar apenas em chatbots, considere também recomendação de vídeos, filtros de spam, reconhecimento facial, tradução automática, rotas, busca, correção de texto, recomendações de música, detecção de fraude e autocomplete de código.

Para discutir em sala

Se já convivíamos com IA havia anos, por que a sensação de “explosão da IA” se tornou tão forte a partir da popularização dos modelos generativos?

Inteligência Artificial (IA) é uma área da Computação dedicada à criação de sistemas capazes de executar tarefas que associamos a capacidades inteligentes, como perceber padrões, classificar, prever, planejar, recomendar, compreender linguagem, gerar conteúdo e escolher ações.

A expressão é ampla. Por isso, é útil enxergar a IA como um grande guarda-chuva:

Área	Ideia central
Inteligência Artificial	Campo mais amplo: sistemas que executam tarefas associadas à inteligência.
Machine Learning	Sistemas aprendem padrões a partir de dados em vez de depender apenas de regras escritas manualmente.
Deep Learning	Subárea do Machine Learning baseada em redes neurais profundas.
IA Generativa	Modelos capazes de gerar novo conteúdo, como texto, imagem, áudio, vídeo e código.

Uma consequência importante: IA Generativa não é sinônimo de Inteligência Artificial. Ela é uma das vertentes atuais de um campo muito maior.

1.1 Duas formas de construir comportamento inteligente

Durante décadas, duas ideias apareceram repetidamente na história da IA: programar regras explicitamente ou fazer a máquina aprender padrões a partir de dados.

Abordagem baseada em regras	Aprendizado a partir de dados
Humano define regras → sistema executa.	Dados → algoritmo ajusta um modelo → modelo faz previsões.
Excelente quando as regras são claras e estáveis.	Útil quando o padrão é complexo demais para ser descrito regra por regra.

A história da IA não é uma linha reta. Ela alterna períodos de enorme entusiasmo com fases de frustração. Ideias antigas reaparecem quando novos algoritmos, dados e hardware tornam possível aquilo que antes era impraticável.

Período	Marco	Por que importa
1943	McCulloch e Pitts	Propõem um modelo matemático simplificado de neurônio, uma das sementes conceituais das redes neurais artificiais.

Tópicos Especiais I

Período	Marco	Por que importa
1950	Alan Turing	Publica "Computing Machinery and Intelligence" e populariza a pergunta: máquinas podem pensar? O chamado Teste de Turing passa a ser uma referência no debate.
1956	Dartmouth	O termo "Artificial Intelligence" ganha identidade formal como campo de pesquisa no encontro de Dartmouth.
1957–1958	Perceptron	Frank Rosenblatt apresenta o perceptron, modelo simples de aprendizagem inspirado em neurônios.
1960–1970	Primeiro entusiasmo e limitações	Expectativas crescem, mas poder computacional, dados e algoritmos ainda são insuficientes para muitas promessas.
1970–1980	AI Winter	Financiamento e entusiasmo diminuem após resultados aquém das expectativas.
1980	Sistemas especialistas	Sistemas baseados em regras e conhecimento especializado têm grande uso corporativo, mas são difíceis de manter e frágeis fora do domínio previsto.
1990	Machine Learning	Métodos estatísticos e aprendizagem a partir de dados ganham importância crescente.
1997	Deep Blue	O sistema da IBM derrota Garry Kasparov em uma partida de xadrez de alto nível e se torna símbolo da capacidade computacional aplicada a tarefas complexas.
2012	AlexNet e Deep Learning	Redes neurais profundas impulsionadas por GPUs obtêm um salto de desempenho em visão computacional e aceleram a adoção do Deep Learning.
2016	AlphaGo	O sistema da DeepMind derrota Lee Sedol no Go, combinando redes neurais e aprendizagem por reforço.
2017	Transformer	O artigo "Attention Is All You Need" apresenta a arquitetura Transformer, base de grande parte dos modelos de linguagem modernos.
2018–2020	Modelos fundacionais	BERT, GPT-2, GPT-3 e outros modelos ampliam a ideia de treinar modelos gerais que podem ser adaptados a muitas tarefas.

Período	Marco	Por que importa
2022	ChatGPT	A interface conversacional populariza o acesso aos LLMs e transforma a IA generativa em fenômeno de massa.
2023–2026	Sistemas com IA	O foco passa de apenas “ter um modelo” para integrar LLMs com RAG, ferramentas, memória, APIs, agentes, avaliação, segurança e observabilidade.

2.1 Os “invernos” da IA

A expressão AI Winter descreve períodos em que expectativas exageradas não se confirmaram, resultados práticos ficaram aquém do prometido e o financiamento diminuiu. Esse padrão histórico ensina uma lição valiosa: avanços reais podem coexistir com exagero, marketing e previsões imprecisas.

Parada para pensar

A IA atual está imune a novos ciclos de frustração? Ou algumas promessas de hoje também podem estar superestimadas? Diferencie avanço tecnológico comprovado de expectativa especulativa.

03

POR QUE A IA AVANÇOU TANTO NAS ÚLTIMAS DÉCADAS?

Nenhum fator sozinho explica a evolução recente. O salto foi resultado da combinação de vários elementos que passaram a se reforçar mutuamente.

1	Dados A internet, dispositivos digitais e sistemas corporativos produziram enormes volumes de texto, imagens, áudio, vídeo e registros de uso.
2	Hardware GPUs e aceleradores tornaram viável executar muitas operações matemáticas em paralelo, reduzindo drasticamente o tempo de treinamento.
3	Algoritmos Deep Learning, mecanismos de atenção, Transformer e novas técnicas de treinamento melhoraram o aproveitamento dos dados.
4	Escala Modelos passaram a ser treinados com mais parâmetros, mais dados e mais capacidade computacional, produzindo capacidades emergentes e maior generalidade.

3.1 A mudança de paradigma

Em muitos sistemas tradicionais, especialistas escreviam regras. No Machine Learning, o sistema aprende padrões nos dados. Nos grandes modelos generativos, esse aprendizado é realizado em escala muito maior, permitindo que um mesmo modelo seja reutilizado em diferentes tarefas por meio de instruções, contexto e ferramentas.

04

O QUE ACONTECE QUANDO VOCÊ CONVERSA COM UM LLM?

Um Large Language Model (LLM) é um modelo treinado para lidar com linguagem em grande escala. Em uma descrição introdutória, seu funcionamento pode ser entendido como um processo de previsão sequencial de tokens.

Texto do usuário
Tokenização
Representações numéricas (embeddings)
Processamento pelo Transformer
Distribuição de probabilidades
Escolha/geração do próximo token
Repetição do processo

4.1 Tokens

Modelos não trabalham diretamente com “palavras” da mesma maneira que uma pessoa. O texto é convertido em unidades chamadas tokens. Um token pode corresponder a uma palavra, parte de uma palavra, pontuação ou outro fragmento, dependendo do tokenizer utilizado.

4.2 Embeddings

Os tokens são representados numericamente. Embeddings são vetores que permitem representar propriedades e relações semânticas em um espaço matemático. Mais adiante na disciplina, utilizaremos embeddings para busca semântica e sistemas RAG.

4.3 Atenção e Transformer

O mecanismo de atenção permite ao modelo atribuir pesos diferentes às partes do contexto ao processar uma sequência. A arquitetura Transformer organiza esse mecanismo de forma altamente paralelizável e tornou-se central para os modelos de linguagem modernos.

4.4 Geração token a token

Considere o início de uma frase:

“O Brasil é um ...”

Possível próximo token	Probabilidade ilustrativa
país	72%
dos	11%
grande	4%
lugar	2%
outros	11%

Os números acima são apenas ilustrativos. O ponto importante é o mecanismo: o modelo calcula uma distribuição para possibilidades de continuação, gera um token e repete o processo utilizando o contexto atualizado.

Importante

Dizer que um LLM prevê tokens não significa que ele seja “apenas autocomplete simples”. A rede aprendeu estruturas extremamente complexas da linguagem e dos dados. Ao mesmo tempo, fluência não garante verdade: o modelo pode produzir uma resposta plausível e incorreta.

05

EXPERIMENTO: VOCÊ TAMBÉM PREVÊ LINGUAGEM

Complete rapidamente as frases abaixo, sem pesquisar:

Era uma vez ...

GABARITO _____

O céu está ...

GABARITO _____

Para abrir uma porta, normalmente usamos ...

GABARITO _____

2 + 2 é ...

GABARITO _____

Pergunta para discussão: como você soube quais continuações eram prováveis? Pessoas aprendem padrões de linguagem e do mundo por experiência. Um LLM aprende regularidades estatísticas a partir de grandes conjuntos de dados, embora o processo e a natureza desse aprendizado sejam diferentes da cognição humana.

06

O MAPA DA DISCIPLINA

A disciplina será organizada como uma construção progressiva. Cada assunto resolve um problema que surge quando tentamos criar aplicações reais com modelos de IA.

Etapa	Pergunta	Tecnologia/conceito
1	Como orientar o modelo?	Engenharia de Prompt
2	Como chamar um modelo a partir de software?	APIs de LLMs
3	Como representar significado matematicamente?	Embeddings
4	Como fazer a IA consultar nossos documentos?	RAG + banco vetorial
5	Como conectar diferentes componentes?	LangChain / LlamaIndex e orquestração
6	Como permitir que a IA execute ações?	Tools / Tool Calling
7	Como fazer o sistema escolher ações?	Agentes inteligentes
8	Como saber se é confiável, seguro e sustentável?	Evals, LLMOps, governança e segurança

6.1 A evolução de uma aplicação ao longo do semestre

Começo: Usuário → LLM

Depois: Usuário → Aplicação → API → LLM

Com conhecimento privado: Usuário → Aplicação → RAG → LLM

Com ações: Usuário → Agente → Ferramentas + RAG + APIs + LLM

Sistema completo: Aplicação + Agente + Avaliação + Segurança + Monitoramento

07 ONDE QUEREMOS CHEGAR: O PROJETO INTEGRADOR

Ao final da disciplina, o objetivo é que os estudantes consigam compreender e implementar uma aplicação de IA que combine várias das peças estudadas. Um exemplo de arquitetura é:

Entrada		Processamento
Usuário	→	Agente
		↓
Documentos / APIs / Calculadora	←	Ferramentas + RAG
		↓ LLM → Resposta

O sistema final pode assumir diferentes formas: assistente acadêmico, tutor inteligente, agente de suporte técnico, buscador semântico, analisador de documentos, assistente de programação ou outra aplicação proposta pela turma.

08 O QUE UMA IA GENERATIVA SABE, E O QUE ELA NÃO GARANTE?

Modelos de linguagem podem produzir respostas impressionantemente úteis, mas algumas propriedades precisam acompanhar qualquer uso responsável:

- **Fluência ≠ verdade:** Uma resposta bem escrita pode conter fatos incorretos.
- **Confiança aparente ≠ certeza:** O modelo pode apresentar informação incorreta de forma convincente.
- **Contexto importa:** A qualidade da instrução, dos dados fornecidos e das fontes recuperadas influencia fortemente a saída.
- **Ferramentas ampliam poder e risco:** Um agente capaz de executar ações precisa de permissões, validação e limites.
- **Avaliar é parte da engenharia:** Não basta “parecer bom”. Sistemas de IA precisam de testes, métricas e monitoramento.
- **Segurança começa no projeto:** Prompt injection, vazamento de dados e uso indevido de ferramentas devem ser tratados desde a arquitetura.

09

DEBATE: SE A IA GERA CÓDIGO, POR QUE APRENDER PROGRAMAÇÃO?

A capacidade de gerar código muda a forma de programar, mas não elimina a necessidade de compreender problemas, requisitos, dados, arquitetura, testes, segurança e integração entre sistemas.

Tese para a disciplina

Quanto mais poderosa fica a IA, maior é o valor de saber formular problemas, escolher ferramentas, verificar resultados e reconhecer quando o sistema está errado.

Questões para debate:

- Qual parte do trabalho de um programador tende a ser mais automatizada?
- Quais responsabilidades continuam humanas quando uma IA gera o código?
- Quem é responsável quando uma aplicação baseada em IA causa um erro?
- É possível utilizar IA de forma profissional sem saber avaliar tecnicamente suas respostas?

10

RESUMO DA AULA 00

- IA é um campo amplo; Machine Learning, Deep Learning e IA Generativa são conceitos relacionados, mas não equivalentes.
- A história da IA inclui ciclos de entusiasmo, limitações e retomadas.
- Dados, hardware, algoritmos e escala ajudam a explicar o avanço recente.
- O Transformer, apresentado em 2017, tornou-se uma arquitetura central para os LLMs modernos.
- LLMs processam tokens e geram continuações a partir de distribuições de probabilidade condicionadas ao contexto.
- Embeddings representam informações numericamente e serão fundamentais para busca semântica e RAG.
- Aplicações modernas combinam modelos com prompts, APIs, RAG, ferramentas e agentes.
- Avaliação, segurança, governança e LLMOps são partes da engenharia, não detalhes opcionais.

11

EXERCÍCIOS DE FIXAÇÃO

1. [Conceitual] Explique com suas palavras a diferença entre Inteligência Artificial, Machine Learning, Deep Learning e IA Generativa.

2. [História] Por que 1956 é frequentemente tratado como um marco na história da IA?

3. [História] O que os chamados "AI Winters" ensinam sobre expectativas tecnológicas?

4. [Evolução] Quais quatro fatores desta apostila ajudam a explicar o avanço recente da IA? Explique o papel de cada um.

5. [Transformer] Por que o ano de 2017 é especialmente importante para esta disciplina?

6. [LLM] Explique, em nível introdutório, como um LLM gera uma resposta.

7. [Pensamento crítico] Por que uma resposta linguisticamente convincente não deve ser automaticamente considerada verdadeira?

8. [Arquitetura] Qual é o papel do RAG em uma aplicação baseada em LLM?

9. [Agentes] Qual é a diferença básica entre um chatbot que apenas responde e um agente que utiliza ferramentas?

10. [Aplicação] Imagine um problema real da sua escola, trabalho ou cotidiano. Proponha uma aplicação de IA e indique pelo menos três componentes que ela poderia utilizar ao longo da disciplina.

12

GLOSSÁRIO ESSENCIAL

Termo	Definição
Agente de IA	Sistema que utiliza um modelo para decidir ações e interagir com ferramentas ou ambientes.
API	Interface que permite a comunicação entre programas e serviços.
Attention	Mecanismo que permite ponderar quais partes do contexto são mais relevantes durante o processamento.
Embedding	Representação vetorial de dados, utilizada para capturar relações e similaridades.
IA Generativa	Categoria de modelos capazes de produzir novo conteúdo.
LLM	Large Language Model: modelo de linguagem treinado em grande escala.
LLMOps	Práticas para implantação, operação, avaliação, monitoramento e controle de aplicações com LLMs.
Prompt	Instrução, contexto ou entrada fornecida ao modelo.
Prompt Injection	Ataque em que instruções maliciosas tentam alterar o comportamento esperado de uma aplicação com LLM.
RAG	Retrieval-Augmented Generation: arquitetura que recupera informações externas e as fornece ao modelo para fundamentar a geração.
Token	Unidade de texto utilizada pelo modelo durante o processamento.
Transformer	Arquitetura neural baseada em mecanismos de atenção, central em muitos modelos modernos.

Termo	Definição
Vector Database	Banco otimizado para armazenar e buscar representações vetoriais, como embeddings.

13

PARA A PRÓXIMA AULA

Na próxima etapa, vamos sair do panorama geral e investigar mais de perto o funcionamento dos modelos de linguagem: tokenização, embeddings, atenção, Transformer, janela de contexto e geração do próximo token.

Próxima parada: “Como um LLM funciona por dentro?”