

IA em Cenários Reais

Assistentes Pessoais

Siri, Alexa, Google Assistant e Cortana são assistentes pessoais de quatro gigantes de tecnologias: Apple, Amazon, Google e Microsoft.

A função destas soluções é simples: receber comandos por voz, executar comandos e providenciar um feedback por meio de áudio para o usuário.

Embora seja simples, existem diferentes desafios tecnológicos na criação de um assistente pessoal virtual:

Reconhecimento de fala: O software deve ser capaz de receber o sinal captado pelo microfone e transformar o mesmo em texto

Processamento de comandos: O assistente deve com base nos comandos recebidos decidir quais ações executar

Gerar respostas: Uma vez que o comando tenha sido executado, é necessário dar uma resposta ao usuário por meio de duas etapas:

1. tomar a decisão de qual resposta deve ser dada

2. gerar o áudio a ser reproduzido.

Diferentemente do GPT (que é executado na nuvem, onde o dispositivo apenas recebe o input, envia, e recebe a resposta para ser reproduzida), os assistentes são executados no dispositivo com limitações, sendo um paradigma diferente e que não é possível determinar a exatidão com o mesmo nível do GPT.

Watsonx e IA Generativa

Lançado em maio de 2023 como plataforma de IA generativa e dados em nuvem.

Integra diferentes Modelos de Linguagem Grande (MLL), incluindo o GPT.

Aplicações em serviços empresariais e otimização de processos automatizados.

Expansão para Aplicações Comerciais

Saúde: suporte a diagnósticos médicos em oncologia e radiologia.

Chatbots: utilizados para atendimento automatizado e suporte empresarial.

Análise de dados: previsão do tempo, assistência a professores, e otimização fiscal.

Automação: auxílio na criação de código para programas.

Origens e Desenvolvimento

Criado em 2011 pela IBM para competir no programa Jeopardy!.

Watson venceu três partidas contra outros concorrentes humanos e IA.

Capacidade de responder perguntas com análise avançada de linguagem.

Hardware avançado: 2880 threads de processadores POWER7, 16 TB de RAM, 80 TeraFLOPs de processamento.

Algoritmos desenvolvidos em Java, C++ e Prolog.

Composto por dois módulos principais: uma parte responsável por responder as perguntas e um framework Apache responsável pela interface de entrada e saída

IBM Watson