Se 2023 foi tudo sobre a ascensão da inteligência artificial generativa (IA) e a sua entrada nas principais conversas sobre tecnologia, 2024 tornou-se o ano em que a IA começou a demonstrar as suas capacidades transformadoras. O que começou como uma moda de chatbot baseado em texto que poderia responder aos usuários de uma forma humana, está hoje alimentando muitos dos principais produtos e plataformas de tecnologia que oferecem casos de uso práticos. Novos casos de uso da tecnologia também foram vistos na geração de música e vídeo, bem como em capacidades de agência. E, ao contrário das opiniões dos pessimistas, a bolha da IA não estourou este ano.

O ano de 2024 marcou a entrada de grandes modelos de linguagem (LLMs) focados em raciocínio avançado, o início da era dos PCs de IA (Copilot + PCs, se você acreditar na palavra da Microsoft) e o crescimento acelerado do espaço de IA de código aberto. No entanto, estes são apenas alguns dos principais eventos que dominaram as manchetes deste ano. Vamos dar uma olhada nos melhores e maiores momentos que moldaram o espaço da IA em 2024.

O que você vai ler:

Ano dos modelos de IA de alto desempenho da OpenAI

A OpenAI pode ter iniciado a tendência de IA generativa com sua arquitetura Generative Pre-trained Transformer (GPT) no final de 2022, mas no final de 2023, estava claro que os gigantes da tecnologia não iriam ficar fora da corrida por muito tempo. Google, Microsoft, Meta e até Amazon lançaram vários modelos de IA, tentando levar a coroa nas pontuações de benchmark.

A OpenAI começou o ano em grande com o lançamento do modelo GPT-4o AI com foco em raciocínio avançado em maio, que foi seguido pelo GPT-4o Mini em julho. A empresa de IA também encerrou o ano em alta com o lançamento da versão completa do modelo o1 e o tão aguardado lançamento de seu modelo de texto para vídeo Sora.

Além disso, a empresa também introduziu seu Modo de Voz Avançado com Visão no aplicativo ChatGPT, oferecendo novas maneiras de interagir com o chatbot. A OpenAI também lançou seu próprio mecanismo de busca denominado ChatGPT Search, que foi

integrado à plataforma chatbot.

Mas o maior golpe para a empresa de IA veio na forma de uma parceria com a Apple, que viu o ChatGPT ser integrado às ferramentas de inteligência da Apple. Após a parceria, a OpenAI também lançou um aplicativo independente para macOS e Windows para ChatGPT.

Conjunto diversificado de ofertas de IA do Google

O Google também enlouqueceu com seu grande número de lançamentos de modelos. Em fevereiro, a empresa lançou a série Gemini 1.5 de modelos de IA, incluindo o Gemini 1.5 Pro com um trilhão de parâmetros. Em dezembro, fechou o ano lançando a série Gemini 2.0, com o modelo Flash disponível para todos em pré-visualização, e um modelo maior reservado para assinantes pagantes.

Mas isso não foi tudo que a gigante da tecnologia com sede em Mountain View fez. Google DeepMind, a ala de IA da empresa, lançou o modelo de geração de imagem Imagen 3 e o modelo de geração de vídeo Veo 2, e apresentou uma prévia do modelo de IA de geração de música MusicLM. Além disso, a gigante da tecnologia também lançou o NotebookLM, uma ferramenta de IA para processar documentos grandes que também pode criar podcasts envolventes com dois hosts de IA.

A empresa também introduziu novos recursos no Gemini. Ele adicionou um recurso de comunicação de voz bidirecional chamado Gemini Live e integrou o assistente Gemini AI na maioria dos aplicativos do Google Workspace, incluindo Gmail, Documentos, Apresentações e Planilhas.

A Meta pode ter sido conhecida por suas plataformas de mídia social antes de 2024, mas este ano, a empresa mostrou suas capacidades desenvolvendo e lançando vários modelos de linguagem pequena (SLMs), muitos dos quais foram lançados em código aberto.

A gigante da tecnologia apresentou vários de seus modelos da série Large Language Model Meta AI (Llama), incluindo modelos focados em codificação 70B e 30B, o maior modelo de código aberto Llama 3.1 405B, bem como vários modelos de instrução. No entanto, o maior anúncio da empresa veio com a expansão global do seu chatbot nativo Meta AI.

Meta AI foi adicionado ao Messenger, Instagram e WhatsApp do Facebook e foi expandido para várias regiões, incluindo a Índia, em abril de 2024, antes de ser disponibilizado globalmente em setembro. O chatbot com tecnologia de IA também foi adicionado aos óculos Ray-Ban Meta com recursos de processamento de visão em tempo real.

Microsoft e a era dos PCs Copilot+

Mesmo usando modelos de IA do OpenAI, a Microsoft teve sucesso em criar um nicho de IA no espaço do PC. A gigante da tecnologia com sede em Redmond rapidamente manifestou

seus desejos quando fez parceria com a Snapdragon (e mais tarde com a Intel e AMD) para introduzir a classificação AI PC, que tinha um requisito obrigatório – a adição de um botão físico do Copilot no teclado. Surgiu assim a era do Copilot+ PC, onde o chatbot nativo da empresa era integrado a desktops e laptops por meio do sistema operacional Windows.

Expandir seu chatbot de IA para milhões de usuários seria considerado um sucesso em todos os manuais de negócios, no entanto, a gigante da tecnologia estava longe de terminar. Em 2024, também integrou ferramentas Copilot em produtos Microsoft 365 e adicionou capacidade de voz e visão ao chatbot. Além disso, também lançou o recurso Recall baseado em IA (em beta) que permite aos usuários de PC fazer perguntas à IA sobre atividades anteriores do dispositivo.

O papel da Amazon como agregador de IA

Muitos analistas do setor disseram que a Amazon estava atrasada para entrar no espaço da IA e, embora isso possa ser verdade, a empresa seguiu um caminho único em 2024 para ainda permanecer relevante no espaço da IA. Em termos de lançamentos baseados em IA, a empresa não teve muitos momentos de destaque. Ela lançou a ferramenta Rufus AI no aplicativo Amazon que atua como assistente de compras. Também lançou a série Titan de modelos de IA e um modelo de geração de vídeo para empresas.

No entanto, a empresa também assumiu discretamente o papel de agregadora e começou a integrar modelos de IA de um grande número de terceiros à sua plataforma Amazon Web Services (AWS). Também investiu no lançamento de ferramentas de IA que melhoram a eficiência das respostas e reduzem as alucinações. A Amazon também reforçou seus servidores para permitir que eles lidassem com grandes volumes de processamento de IA.

Outros anúncios notáveis de IA

Embora os holofotes estivessem voltados para os principais players de IA em 2024, as empresas menores de IA também não deixaram de impressionar. A Anthropic continuou seu sucesso com Claude ao lançar a série Claude 3 no início do ano e a série Claude 3.5 no final. A empresa também lançou um aplicativo de desktop para Mac e Windows em versão beta, bem como aplicativos independentes para Android e iOS. Além disso, seus recursos de uso de ferramentas e compreensão de PDF tornaram Claude um chatbot mais capaz em 2024.

Perplexity, o mecanismo de pesquisa baseado em IA, lançou um modo Pro que mostra respostas detalhadas para consultas complexas. Ela também lançou um aplicativo independente para Mac este ano. No entanto, embora houvesse aspectos positivos, a decisão da empresa de incorporar anúncios até mesmo para os assinantes premium atraiu algumas críticas.

Mistral continuou seu lançamento consistente de modelos de IA totalmente de código aberto mesmo em 2024. Tudo começou com o lançamento dos modelos de IA 8x22B Mixture of

Experts (MoE) e seguiu com Mixtral Open 2 LLM. A empresa também surpreendeu os desenvolvedores com o lançamento do modelo Pixtral 12B AI, que vem com recursos de visão computacional.

IA em 2025: uma breve perspectiva

Embora tenhamos tentado capturar todos os principais anúncios no espaço de IA em 2024, é praticamente impossível mencionar todos os lançamentos notáveis, dada a febre da IA que está correndo solta na indústria de tecnologia. Mas agora que o ano está a terminar, esperamos que 2025 seja um ano igualmente repleto de ação para esta tecnologia.

No próximo ano, esperamos ver o aumento da IA de agência e a sua integração em plataformas e dispositivos. Imagine pedir ao seu chatbot para reservar um ingresso de cinema ou comprar um produto pelo menor preço possível e ele completar a ação sem necessidade de qualquer intervenção. É isso que os agentes de IA podem oferecer.

Além disso, também acreditamos que no próximo ano veremos uma melhor implementação da função de memória em chatbots, abandonando a estrutura rudimentar de geração aumentada de recuperação (RAG). Isso fará com que os chatbots se tornem melhores assistentes e companheiros para os usuários. O processamento de vídeo em tempo real também poderá se tornar mais acessível no próximo ano. E, finalmente, acreditamos que a Índia dará grandes passos no sentido da adoção da IA em 2025.