memória virtual

A IA generativa e os modelos de linguagem em larga escala (LLMs) estão mudando a forma como o software empresarial é desenvolvido e entregue. O que começou como chatbots e ferramentas simples de automação agora está evoluindo para sistemas poderosos profundamente integrados às arquiteturas de software, afetando tudo, desde processos de back-end até interfaces de usuário.

O que você vai ler:

memória virtual

Chatbots são uma tendência de curto prazo

Atualmente, as empresas estão focadas no desenvolvimento de chatbots e GPTs customizados para solucionar diversos problemas. Estas ferramentas baseadas em IA são particularmente úteis em duas áreas: melhorar o acesso ao conhecimento interno e automatizar o atendimento ao cliente. Os chatbots podem ser usados para construir sistemas de resposta que dão aos funcionários acesso rápido a uma vasta base de conhecimento interna, quebrando silos de informação.

Estas ferramentas são úteis, mas o seu valor está a diminuir devido à falta de inovação ou diferenciação. Eles também podem fornecer uma interface de usuário inadequada devido à falta de compreensão das melhores alternativas para resolver um problema específico.

O futuro provavelmente contará com recursos de IA perfeitamente integrados aos produtos de software, sem serem visíveis para o usuário final.

IA generativa como tecnologia cotidiana

Nos próximos anos, a IA evoluirá de uma ferramenta opaca que requer interação direta do usuário para um componente totalmente integrado em seu conjunto de recursos. A IA generativa pode permitir recursos como criação de conteúdo dinâmico, tomada de decisão inteligente e personalização em tempo real, sem a necessidade de interação direta do usuário. Isso muda fundamentalmente o design da UI e a forma como o software é usado. Em vez de inserir manualmente determinados parâmetros, os usuários poderão cada vez mais descrever suas necessidades em linguagem natural.

Exemplos disso já podem ser vistos em ferramentas como Adobe Photoshop. O recurso ‘Preenchimento Generativo’ elimina a necessidade dos usuários ajustarem manualmente vários parâmetros. Em vez disso, simplesmente descreva o que deseja preencher na área selecionada da imagem. A entrada em linguagem natural pode se espalhar pelos aplicativos para tornar a experiência do usuário mais intuitiva e livre das restrições dos elementos tradicionais da UI.

O desafio futuro para a indústria não é a escassez, mas a abundância. Em outras palavras, é fundamental encontrar e priorizar as oportunidades mais promissoras.

O efeito da generalização do LLM em comparação com modelos especiais de ML

Uma mudança notável que a IA generativa trouxe para a indústria de TI foi a democratização das funções da IA. Antes do advento do LLM e dos modelos de difusão, as organizações tinham que investir tempo, esforço e recursos significativos no desenvolvimento de modelos especializados de ML para resolver problemas complexos. Isso

exigia funções e equipes especializadas para gerenciar a coleta de dados específicos do domínio, a preparação de recursos, a rotulagem de dados, o retreinamento e todo o ciclo de vida do modelo.

Os LLMs estão agora mudando a forma como as empresas abordam problemas que são difíceis ou impossíveis de resolver com algoritmos. Embora o termo “linguagem” em LLM, ou modelo de linguagem em larga escala, possa ser enganoso, este modelo autorregressivo pode processar imagens, vídeos, sons e até proteínas que podem ser facilmente decompostas em tokens. As empresas podem usar a arquitetura de geração aumentada de pesquisa (RAG) para potencializar essas ferramentas versáteis com seus próprios dados. Isto torna possível utilizar uma ampla gama de funções.

Não há necessidade de equipes especializadas, rotulagem extensa de dados ou pipelines de ML complexos. O amplo conhecimento prévio de aprendizagem do LLM permite processar e interpretar com eficácia até mesmo dados não estruturados.

Um aspecto importante desta democratização é que os LLMs são acessíveis através de uma API fácil de usar. Como quase todos os desenvolvedores hoje em dia sabem como trabalhar com serviços baseados em API, esses modelos podem ser perfeitamente integrados aos ecossistemas de software existentes. Isto permite que as empresas beneficiem de um modelo poderoso sem terem de se preocupar com a infraestrutura subjacente. Além disso, alguns modelos podem ser operados localmente se você tiver requisitos específicos de segurança ou proteção de dados. No entanto, seguir esse caminho significa sacrificar alguns dos benefícios que os modelos mais recentes oferecem.

Por exemplo, considere um aplicativo que registra e gerencia despesas de viagem. Tradicionalmente, esses aplicativos usavam modelos de ML especialmente treinados para classificar os recibos carregados em categorias contábeis, como DATEV. Isso exigia infraestrutura dedicada para gerenciar a coleta de dados, o treinamento e as atualizações de modelos e, idealmente, um pipeline MLOps completo (para treinamento, implantação e monitoramento de modelos).

Esses modelos de ML agora podem ser substituídos por LLMs que aproveitam amplo conhecimento com instruções apropriadas para classificar documentos. Os recursos multimodais do LLM simplificam a pilha de tecnologia, eliminando a necessidade de reconhecimento óptico de caracteres (OCR). Os dados de recebimento também devem incluir o preço líquido e o preço bruto ou a taxa de imposto? Um LLM pode fazer isso.

Recursos baseados em IA que não eram possíveis no passado

A IA generativa permite uma variedade de recursos que não são facilmente acessíveis às organizações porque exigem grandes investimentos em soluções personalizadas de ML ou algoritmos complexos. Vejamos um exemplo específico.

memória virtual

Pesquisa baseada em sentimento e contexto além da pesquisa por palavra-chave

A pesquisa baseada em sentimentos é uma melhoria significativa em relação aos sistemas tradicionais de pesquisa baseados em palavras-chave.

Quando os utilizadores expressam as suas intenções em linguagem natural, a IA generativa pode capturar todo o contexto e "sentimento". Um exemplo é o seguinte:

Pesquise palavras-chave existentes: "Melhor restaurante de Berlim"
Pesquisa baseada em sentimento e contexto: "Sou um gourmet exigente e gosto de wine bars que também servem comida, de preferência aqueles que utilizam ingredientes locais. Recomendo restaurantes em Berlim Mitte e Kreuzberg. No entanto, exclua bares de vinho naturais intrometidos.

Na pesquisa baseada em sentimento e contexto, o LLM pode compreender e abordar:

- Autodescrição como um 'gourmet exigente'
- Preferência por wine bars que também servem comida
- Preferência por ingredientes locais
- Preferências regionais específicas (Mitte e Kreuzberg)
- Distinção entre bares de vinho regulares e "bares de vinho naturais intrometidos"

Este nível de nuance e contexto permite que a pesquisa forneça resultados altamente personalizados e relevantes, em vez de simplesmente combinar palavras-chave.

A implementação de pesquisa baseada em sentimento e contexto pode melhorar significativamente a experiência do usuário em muitos aplicativos.

- Base de conhecimento interna: Os funcionários podem usar consultas em linguagem natural para encontrar informações que descrevam sua situação ou necessidades específicas.
- Plataforma de comércio eletrônico: os clientes podem descrever produtos mesmo que não conheçam a terminologia exata.
- Sistema de atendimento ao cliente: os usuários podem descrever seus problemas em detalhes. O sistema fornece uma solução mais precisa ou conecta você à pessoa de suporte apropriada.
- Sistema de gerenciamento de conteúdo: os editores de conteúdo podem descobrir ativos ou conteúdo por descrição, sem depender de inúmeras tags ou metadados.

Análise inteligente de dados e conteúdo

Análise de sentimento

Um exemplo seria um sistema interno onde os funcionários podem postar mensagens curtas de status sobre seu trabalho. Os gerentes desejam avaliar o humor geral de sua equipe

durante uma semana específica. No passado, era difícil implementar a análise de sentimento de postagens com um modelo de ML personalizado. Com o LLM, essa complexidade é reduzida a uma simples chamada de API.

Os resultados não precisam necessariamente ser produzidos em linguagem legível. Você pode fornecê-lo como JSON estruturado para que o sistema exiba um ícone ou gráfico correspondente. Alternativamente, o LLM pode simplesmente imprimir emojis para indicar o clima. É claro que a implementação destas funções requer o consentimento dos funcionários.

Obtenha insights de dados complexos

Outro exemplo da capacidade do LLM de analisar dados complexos é o gerenciamento inteligente de alarmes para sistemas de refrigeração. As áreas nas quais esses sistemas tradicionalmente se concentram são:

- Painel de alerta gráfico com dados e alertas em tempo real
- Formato tabular complexo e filtrável para dados de série temporal

Embora esses recursos sejam úteis, eles exigem uma interpretação humana significativa para produzir insights significativos. Aqui, o LLM pode transformar dados brutos em insights imediatamente acionáveis e expandir a funcionalidade do sistema de forma imediata, sem a necessidade de um modelo de ML dedicado.

- **Geração automática de relatórios**: O LLM pode analisar dados de séries temporais e gerar relatórios detalhados em linguagem natural. Isso pode destacar tendências, anomalias e indicadores-chave de desempenho que são valiosos tanto para engenheiros quanto para gerentes. Por exemplo, você pode criar um relatório que resuma os alertas da semana anterior, identifique problemas recorrentes e sugira áreas de melhoria.
- **Mergulho profundo**: O LLM pode ir além da simples representação de dados e identificar e explicar padrões complexos em dados. Por exemplo, você pode identificar sequências de alarmes que indicam problemas graves do sistema. Este é um insight que pode ser esquecido em tabelas e gráficos tradicionais.
- **Insights Preditivos**: Ao analisar dados passados, o LLM pode prever o status futuro do sistema. Isso permite a manutenção proativa e ajuda a prevenir possíveis falhas.
- **saída estruturada**: além de relatórios em linguagem natural, o LLM pode gerar dados estruturados como JSON. Isso permite criar interfaces gráficas de usuário dinâmicas que apresentam visualmente informações complexas.
- **consultas em linguagem natural**: O engenheiro pergunta: “Quais dispositivos provavelmente entrarão em modo de failover nas próximas semanas?” Você pode fazer a mesma pergunta em linguagem natural e receber imediatamente respostas e visualizações relevantes. Este recurso agora também está disponível como API em tempo real por meio do OpenAI.

Caixa preta multimodal que suporta escrita, fala, visualização e audição

Multimodal expande muito as capacidades do LLM. Modelos que podem processar texto, imagens, som e voz permitem combinações complexas de funções. Um exemplo seria um aplicativo que ajuda os usuários a processar conteúdo visual complexo e prepará-lo como texto ou voz.

A gama de possíveis casos de uso é muito ampla. Ele pode ser aplicado a uma variedade de situações, como preencher um banco de dados com títulos de livros reconhecidos no vídeo de alguém folheando uma estante, identificar animais desconhecidos em imagens de vigilância de um galinheiro ou uma mulher escocesa dizendo nomes de estradas para seu carro alugado alemão. sistema de navegação.

Restrições e soluções tecnológicas

LLM tem algumas limitações técnicas. Entre eles, a janela de contexto, que é a quantidade de texto (mais precisamente, o número de tokens) que o modelo de linguagem pode processar de uma só vez, é um elemento muito importante.

A maioria dos LLMs possui janelas de contexto limitadas a milhares ou dezenas de milhares. Por exemplo, a janela de contexto do GPT-4 pode lidar com 128.000 tokens e o Gemini 1.5 Pro pode lidar com até 2 milhões de tokens. Este pode parecer um número bastante alto, mas pode rapidamente se tornar um gargalo ao processar conjuntos de entradas como livros ou vídeos longos.

Felizmente, existem diversas estratégias para superar essas limitações.

- **Agrupando e resumindo**: divida um documento grande em pequenos segmentos que caibam na janela de contexto. Processe cada segmento individualmente e depois mescle os resultados.
- **Geração de aumento de pesquisa (RAG)**: em vez de confiar apenas no amplo conhecimento do modelo, as informações relevantes são recuperadas de fontes de dados separadas e incorporadas ao prompt.
- **Ajuste de Domínio**: Ao combinar engenharia imediata e base de conhecimento específica de domínio, o conhecimento especializado pode ser utilizado sem limitar a versatilidade do modelo.
- **janelas de correr**: use uma janela deslizante para analisar sequências de texto longas, como dados de série temporal ou documentos longos. O modelo mantém algum contexto ao mover o documento inteiro.
- **inferência em várias etapas**: Divida um problema complexo em uma série de pequenas etapas. Cada etapa usa o LLM dentro das restrições da janela de contexto e os resultados das etapas anteriores influenciam as etapas subsequentes.
- **Abordagem híbrida**: Métodos tradicionais de recuperação de informações, como TF-IDF e BM25, podem ser usados para pré-filtrar passagens de texto relevantes. Isto

reduz significativamente a quantidade de dados para análise LLM subsequente, aumentando a eficiência do sistema geral.

IA generativa como componente padrão de software empresarial

As empresas devem reconhecer a IA generativa como uma tecnologia de uso geral que terá impacto em tudo. A IA generativa não só se tornará parte da pilha de desenvolvimento de software, mas também se tornará um elemento essencial na implementação de recursos novos ou existentes. Para garantir a competitividade futura do desenvolvimento de software, devemos ir além da simples introdução de ferramentas de IA para o desenvolvimento de software e preparar infraestruturas, padrões de design e operações para a crescente influência da IA.

À medida que essas mudanças progridem, as funções dos arquitetos de software, desenvolvedores e designers de produtos provavelmente evoluirão. À medida que as capacidades técnicas puras se tornam mais baratas e mais automatizadas, novas tecnologias e estratégias terão de ser desenvolvidas para conceber funções de IA, processar resultados não determinísticos e integrar-se perfeitamente com uma variedade de sistemas empresariais. Espera-se que as competências interpessoais e a colaboração entre trabalhadores técnicos e não técnicos se tornem mais importantes do que nunca.

* *Robert Glaser é chefe de dados e IA do INNOQ. Em relação ao uso prático da IA generativa, estamos hospedando o podcast 'AI und Jetzt' e discutindo o potencial da IA em todos os setores.*
dl-ciokorea@foundryco.com