**DIA 1**

**Explorando o Raciocínio em Grandes Modelos de Linguagem (LLMs): Uma Aula com Denny Zhou da Google DeepMind**

No dia 9 de setembro, tive a oportunidade de participar de uma aula enriquecedora sobre o tema **Raciocínio em Grandes Modelos de Linguagem (LLMs)**, ministrada por **Denny Zhou**, pesquisador da Google DeepMind. A aula foi transmitida ao vivo e abordou de maneira técnica e didática os desafios e avanços no uso de LLMs para tarefas complexas de raciocínio. Além disso, foram apresentados materiais introdutórios e discutidos artigos fundamentais que lançam luz sobre as limitações e potencialidades desses modelos.

**O que são LLMs e por que o raciocínio é um desafio?**

Grandes Modelos de Linguagem, como GPT e similares, são projetados para gerar textos coerentes e contextualizados, baseando-se em padrões estatísticos extraídos de grandes conjuntos de dados. Apesar de sua habilidade impressionante de processar e gerar linguagem natural, ainda enfrentam dificuldades significativas ao realizar tarefas que requerem raciocínio lógico ou correção de seus próprios erros.

Essa aula abordou questões fundamentais sobre as limitações desses modelos, com base em pesquisas de ponta e experimentos que revelam os desafios no desenvolvimento de LLMs capazes de lidar com problemas complexos.

DIA 2

**Agentes LLM: História Breve e Panorama Geral**

Esta aula apresentou uma visão técnica e detalhada sobre agentes baseados em modelos de linguagem grande (LLMs), abordando desde sua evolução até os principais conceitos que fundamentam seu funcionamento. Além disso, foram discutidos estudos relevantes que ilustram como esses agentes estão sendo aplicados em cenários do mundo real.

**O que são agentes LLM?**

Agentes LLM são sistemas baseados em modelos de linguagem que utilizam a capacidade de processar, interpretar e gerar linguagem natural para executar tarefas e interagir com ambientes. A grande diferença em relação a outros sistemas é sua versatilidade, permitindo que sejam usados em contextos variados, como assistentes virtuais, automação de processos e controle de sistemas.

Os principais conceitos discutidos na aula incluem:

* **Memória de longo e curto prazo**: A memória de curto prazo refere-se às informações armazenadas temporariamente durante uma tarefa ou sessão, sendo descartadas ao final. Já a memória de longo prazo permite que o modelo retenha informações entre sessões, armazenando dados em sistemas externos para continuidade e maior personalização.
* **Ciclo entre raciocínio e ação**: A interação entre raciocínio e ação foi destacada como fundamental para a eficácia de agentes LLM em ambientes dinâmicos, como a web. Nesse contexto, o raciocínio orienta ações iniciais, enquanto o feedback recebido dessas ações informa e refina raciocínios futuros, criando um ciclo adaptativo.
* **Geração Aumentada por Recuperação (RAG)**: Essa abordagem híbrida integra informações de bases externas para aumentar a precisão e o contexto das respostas.

**Exemplos Práticos**

* **Chatbots com memória longa**: Um chatbot com memória de longo prazo pode reter preferências do usuário ou informações de interações anteriores, melhorando a experiência do usuário ao longo do tempo.
* **Interação com a web**: Em cenários como navegação em sites ou consulta de bases de dados dinâmicas, o uso de feedback em tempo real foi apontado como essencial para manter a relevância e precisão das respostas.

DIA 3

**Explorando Frameworks de IA Agentic e AutoGen: Insights de uma Aula Técnica**

Essa sessão técnica, ministrada por Chi Wang, da AutoGen-AI, e Jerry Liu, da LlamaIndex, foi repleta de conceitos avançados sobre como frameworks de inteligência artificial podem ser utilizados para aprimorar a autonomia de sistemas baseados em modelos de linguagem (LLMs), além de como agentes colaborativos e assistentes multimodais estão moldando a próxima geração de aplicações de IA.

**Frameworks de IA Agentic: Planejamento, Recuperação e Execução Autônoma**

Chi Wang apresentou os fundamentos dos frameworks de IA agentic, que têm como objetivo principal capacitar sistemas de IA a agirem de forma autônoma. Essa abordagem combina planejamento estratégico, recuperação de dados e execução de tarefas em fluxos iterativos, tornando os sistemas mais adaptáveis e dinâmicos.

Entre os principais benefícios desse tipo de framework, destacam-se:

1. **Planejamento autônomo:** Permite que os sistemas definam e priorizem metas de forma independente, dividindo-as em tarefas menores.
2. **Recuperação eficiente de informações:** Utiliza fontes internas e externas para coletar dados relevantes que fundamentem as decisões.
3. **Tomada de decisões dinâmica:** O sistema refina continuamente suas ações com base no feedback do ambiente.
4. **Flexibilidade:** Os frameworks são especialmente úteis para lidar com tarefas complexas e de múltiplas etapas, reduzindo a necessidade de supervisão constante do usuário.

Essa autonomia é essencial para aplicações que exigem interações contínuas com ambientes dinâmicos, como pesquisa, automação industrial e resolução de problemas complexos.

**AutoGen: Facilitando Colaboração e Comunicação entre Agentes LLM**

A segunda parte da aula foi dedicada ao *AutoGen*, um framework inovador apresentado por Chi Wang, que habilita múltiplos agentes baseados em LLMs a trabalharem de forma colaborativa e autônoma. Esse modelo utiliza fluxos de conversação entre agentes para resolver tarefas complexas que vão além das capacidades de um único modelo.

Entre os destaques do AutoGen, estão:

1. **Colaboração eficaz:** Cada agente pode se especializar em uma subtarefa específica, contribuindo de maneira coordenada para o resultado final.
2. **Comunicação dinâmica:** O framework facilita a troca de informações estruturadas entre agentes, alinhando estratégias e aprimorando o raciocínio coletivo.
3. **Execução autônoma de tarefas:** Ao orquestrar interações entre agentes, o AutoGen viabiliza a solução de problemas que exigem múltiplas etapas e raciocínio integrado.

**Assistentes Multimodais e o Papel do LlamaIndex**

Jerry Liu complementou a aula com insights sobre assistentes multimodais, construídos utilizando o framework *LlamaIndex*. Esse framework permite que LLMs interajam com dados externos de maneira estruturada, utilizando índices como árvores, grafos e vetores para melhorar a recuperação de informações durante a execução de tarefas.

O *LlamaIndex* é essencial para:

1. **Estruturar dados externos:** Organiza informações de diferentes fontes, como documentos, bancos de dados e APIs, em formatos acessíveis aos LLMs.
2. **Realizar consultas eficientes:** Facilita a busca e recuperação de informações relevantes em tempo real.
3. **Integração com múltiplos tipos de dados:** Assistentes multimodais utilizam texto, imagens, áudio e vídeo, ampliando o escopo de interações e criando experiências mais ricas e dinâmicas para os usuários.

DIA 4

**Tendências Empresariais para IA Generativa e Componentes-Chave na Construção de Agentes e Aplicações de Sucesso**

A sessão foi conduzida por **Burak Gokturk**, especialista do Google, e abordou de forma aprofundada como a IA generativa está moldando o mercado, além de estratégias práticas para maximizar o potencial dessa tecnologia.

**Principais tendências em IA generativa**

Uma das tendências mais marcantes na IA generativa é o desenvolvimento de **modelos multimodais** que integram diferentes tipos de dados, como texto, imagens, áudio e vídeo, para oferecer respostas ricas e contextualizadas. Durante a aula, foi destacado que modelos como o **GPT-4** da OpenAI e o **Gemini** da Google DeepMind já possuem essa capacidade, permitindo aplicações em tarefas complexas, como descrição de imagens, sumarização de vídeos e conversação interativa.

Essa abordagem multimodal não apenas amplia a aplicabilidade desses modelos, mas também os torna mais escaláveis e adaptáveis a diferentes cenários empresariais. Além disso, a generalização entre tarefas é um avanço técnico significativo, pois permite que esses modelos sejam utilizados em uma variedade de domínios sem a necessidade de retrabalho ou configurações complexas.

**Componentes-chave para o sucesso de aplicações de IA generativa**

Outro ponto crucial abordado foi a personalização de modelos base utilizando dados específicos de cada organização. Esse processo, conhecido como fine-tuning, garante que os sistemas de IA gerem resultados mais precisos e alinhados aos objetivos de negócios.

A personalização não apenas aumenta a relevância dos modelos, mas também permite que as empresas adaptem suas soluções às mudanças do mercado, integrando novas fontes de dados ou respondendo a desafios emergentes. Burak destacou que investir em plataformas robustas, como o **Vertex AI** da Google Cloud, é essencial para garantir essa flexibilidade e obter vantagem competitiva.

**Grounding e confiabilidade**

A aula também enfatizou a importância do **grounding** — um processo que alinha as saídas dos modelos de IA ao conhecimento factual do mundo real. Essa técnica é fundamental para evitar “alucinações” (respostas plausíveis, mas incorretas) e garantir que as aplicações de IA sejam confiáveis, especialmente em setores críticos como saúde, finanças e direito.

O uso de sistemas de recuperação de informações, como o **RAG (Retrieval-Augmented Generation)**, foi destacado como uma solução eficaz para reforçar o grounding. Burak apresentou exemplos práticos e discutiu como a abordagem do **Gemini 1.5 Pro** está ajudando a superar desafios no teste "Needle in a Haystack", um framework desenvolvido para avaliar a precisão de sistemas RAG.

DIA 5

**Curso sobre Sistemas de IA Compostos e o Framework DSPy**

A aula, ministrada por Omar Khattab, da Databricks, abordou de forma técnica e didática o conceito de sistemas de IA compostos, as vantagens de sua aplicação e a relevância do framework DSPy para o desenvolvimento de soluções inteligentes. Também foram discutidos tópicos como a otimização de instruções para programas baseados em modelos de linguagem e estratégias de ajuste fino e otimização de prompts.

**O que são sistemas de IA compostos?**

Sistemas de IA compostos representam uma abordagem inovadora que combina diversos modelos e componentes de inteligência artificial para resolver tarefas complexas e multifacetadas. Ao contrário dos modelos monolíticos tradicionais, esses sistemas são modulares e colaborativos, integrando modelos especializados para trabalhar de forma conjunta em diferentes etapas de um problema.

Essa integração oferece diversas vantagens:

* **Composição confiável de capacidades:** Cada modelo se especializa em uma tarefa, permitindo maior precisão nos resultados.
* **Transparência:** A modularidade facilita a compreensão e depuração de cada componente, melhorando a interpretabilidade.
* **Eficiência:** Com a utilização de modelos menores e específicos para cada tarefa, reduz-se o custo computacional e aumenta-se a flexibilidade para futuras adaptações.

Exemplos de aplicações incluem assistentes virtuais que integram reconhecimento de fala, processamento de linguagem natural e execução de ações, além de sistemas de IA na saúde que combinam análise de imagem com recuperação de conhecimento.

**O Framework DSPy**

O DSPy (Declarative Systems for AI) é uma estrutura projetada para facilitar a criação e personalização de sistemas de IA compostos. Ele adota uma abordagem declarativa, permitindo aos desenvolvedores especificar workflows de forma estruturada, integrando componentes especializados de maneira eficiente e transparente.

Principais benefícios do DSPy:

* **Customização:** Permite adaptar sistemas de IA para atender necessidades específicas.
* **Integração modular:** Combina modelos e ferramentas de IA de maneira coesa.
* **Eficiência:** Maximiza o uso de recursos, priorizando modelos menores e otimizados.

**Optimizing Instructions and Demonstrations for Multi-Stage Language Model Programs**

Um dos artigos discutidos na aula destacou estratégias para a otimização de instruções e demonstrações em programas baseados em modelos de linguagem de múltiplas etapas. Foram exploradas metodologias como o **OPRO (Coordinate-Ascent Optimization Proposal)** e o **MIPRO (Multi-prompt Instruction Proposal Optimizer)**, destacando suas diferenças:

* O OPRO realiza refinamentos iterativos, ajustando cada elemento da instrução passo a passo.
* O MIPRO gera e avalia múltiplos prompts simultaneamente, oferecendo maior velocidade e eficiência na identificação de soluções ótimas.

Essa comparação evidenciou a importância de escolher a metodologia certa para diferentes cenários, equilibrando precisão e tempo de processamento.

**Fine-Tuning and Prompt Optimization**

Outro tópico relevante foi a discussão sobre a combinação de ajuste fino e otimização de prompts. Essa estratégia dupla mostrou-se altamente eficaz para aprimorar o desempenho de sistemas de IA, aproveitando o ajuste fino para adaptar modelos a tarefas específicas e a otimização de prompts para melhorar as interações com modelos de linguagem.

DIA 6

**Agentes para Desenvolvimento de Software: Explorando Novas Fronteiras com Graham Neubig**

A apresentação, ministrada pelo professor Graham Neubig, da Carnegie Mellon University , trouxe insights valiosos sobre o uso de agentes baseados em inteligência artificial (IA) para automação e suporte no desenvolvimento de software. Nesta aula, abordamos conceitos técnicos e práticos, explorando como agentes podem transformar práticas de engenharia de software, e discutimos artigos recentes que ilustram avanços na área.

**Automatização no Desenvolvimento de Software**

O desenvolvimento de software tradicional enfrenta desafios como alta complexidade, consumo de tempo e vulnerabilidade a erros humanos. Nesse contexto, agentes inteligentes estão emergindo como soluções para reduzir essas limitações. Esses agentes são projetados para interagir de maneira eficiente com humanos e sistemas computacionais, oferecendo suporte em tarefas como geração de código, depuração e otimização de processos.

Durante a aula, aprofundamo-nos em questões técnicas por meio de exemplos práticos e perguntas de quiz, destacando métodos eficazes e ineficazes no uso de agentes. Por exemplo, uma das discussões enfatizou que métodos manuais de localização de arquivos — como escanear diretórios inteiros em busca de arquivos relevantes — são ineficientes e propensos a erros, especialmente em projetos de grande escala. Além disso, analisamos métricas como o *Pass@K*, que mede a probabilidade de pelo menos uma das saídas geradas por modelos de IA atender aos requisitos especificados, tornando-se uma métrica crucial na avaliação da confiabilidade de modelos de geração de código.

**Questões de Segurança e Vazamento de Dados**

Outro ponto crítico abordado foi a importância de mitigar preocupações de segurança ao utilizar agentes de codificação. Por exemplo, enquanto ferramentas de formatação de código ajudam na legibilidade, elas não mitigam problemas como permissões inadequadas ou execução de ações potencialmente perigosas. A solução reside em medidas como restrição de permissões, revisão de ações realizadas e isolamento dos agentes em ambientes controlados, garantindo a segurança e a integridade dos sistemas.

Além disso, discutimos como o vazamento de dados pode comprometer a avaliação de agentes. Vazamento ocorre quando informações de conjuntos de teste influenciam o treinamento, levando a métricas inflacionadas e uma falsa impressão de desempenho. Isso destaca a importância de práticas rigorosas para evitar a contaminação entre dados de treinamento e teste.

**Code Infilling: Preenchendo Lacunas no Desenvolvimento**

Outro conceito fascinante explorado foi o *code infilling*, uma técnica onde agentes preenchem automaticamente trechos faltantes de código com base no contexto. Essa funcionalidade é particularmente útil para acelerar o processo de desenvolvimento, preenchendo lacunas em códigos parciais e reduzindo a probabilidade de erros.

DIA 7

**Explorando Agentes de IA para Fluxos de Trabalho Corporativos: Aprendizados da Aula Ministrada por Nicolas Chapados**

A aula, ministrada por Nicolas Chapados, da ServiceNow, foi uma imersão técnica e didática em como agentes de inteligência artificial estão sendo desenvolvidos e avaliados para otimizar fluxos de trabalho no contexto corporativo. Além disso, trouxe reflexões sobre frameworks e benchmarks inovadores que ampliam nossa compreensão das capacidades desses agentes.

**Conceitos Chave: Agentes de API vs. Agentes da Web**

Um dos pontos mais didáticos da aula foi a explicação sobre as diferenças entre agentes de API e agentes da web. Enquanto agentes de API interagem diretamente com sistemas por meio de interfaces programáticas predefinidas, agentes da web navegam e interagem com páginas da web, mimetizando ações humanas como cliques e preenchimento de formulários. Essa distinção reflete a diversidade de abordagens para resolver problemas em diferentes tipos de fluxos de trabalho.

**Reflexões sobre Avaliação de Agentes**

Outro aspecto importante abordado foi como avaliamos a eficácia de agentes da web. Isso inclui testes em cenários reais, análise das ações realizadas pelos agentes e validação dos resultados obtidos. A flexibilidade das configurações de avaliação — locais ou remotas — foi destacada como um fator crucial para simular condições reais e identificar desafios práticos.

**Avanços e Desafios**

A aula destacou que benchmarks como WorkArena++ e frameworks como TapeAgents representam avanços significativos na área, mas também levantam desafios. Avaliar habilidades complexas, como planejamento composicional e raciocínio, requer não apenas testes robustos, mas também uma compreensão profunda das aplicações práticas desses agentes em contextos corporativos.

DIA 8

**Rumo a um Framework Unificado de Tomada de Decisão Neural e Simbólica**

No dia 28 de outubro, participei de uma aula foi ministrada por Yuandong Tian, pesquisador do Meta AI (FAIR). Essa aula abordou, de forma técnica e didática, os avanços na integração de modelos neurais e métodos simbólicos para resolução de problemas complexos de planejamento e tomada de decisão. A seguir, compartilho os principais insights e conceitos apresentados, bem como os artigos de destaque que fundamentaram a discussão.

**A Estruturação da Tomada de Decisão Neural e Simbólica**

A busca por um framework unificado que combine redes neurais e métodos simbólicos é motivada pelos desafios de tarefas que envolvem planejamento complexo e raciocínio lógico. Enquanto modelos baseados em redes neurais, como Transformadores, se destacam em generalização e reconhecimento de padrões, os métodos simbólicos são ferramentas eficazes para lidar com restrições estruturais e resolver problemas bem definidos, como otimização combinatória.

Durante a aula, Yuandong Tian enfatizou a importância de sistemas híbridos que aproveitam o melhor desses dois mundos: a capacidade de aprendizado e inferência das redes neurais e a precisão dos algoritmos simbólicos. Essa abordagem é particularmente poderosa em cenários onde as tarefas possuem tanto aspectos dinâmicos e não lineares quanto exigências de robustez lógica.

**Principais Conceitos e Aplicações**

* **Modos Rápido e Lento em Modelos Neurais:**  
  A integração de modos rápido e lento, como proposto no *Dualformer*, é essencial para modelos que precisam alternar entre soluções eficientes e análises detalhadas, dependendo das necessidades da tarefa.
* **Feedback Contínuo em Sistemas Híbridos:**  
  Um sistema híbrido que combina redes neurais e solucionadores simbólicos pode se otimizar continuamente. Essa interação dinâmica é fundamental para resolver tarefas de planejamento com restrições não lineares e adaptáveis.
* **Generalização e Estrutura:**  
  Os modelos neurais fornecem a capacidade de generalizar padrões em grandes volumes de dados, enquanto os algoritmos simbólicos trazem estrutura e precisão para problemas com soluções bem definidas.

DIA 9

**Project GR00T: Um Roteiro para Robótica Generalista**

No dia 4 de novembro, a palestra foi ministrada por Jim Fan, pesquisador da NVIDIA. Essa aula foi parte de um curso abrangente sobre robótica e inteligência artificial, explorando o estado da arte e os desafios do desenvolvimento de agentes generalistas. Além de uma apresentação técnica detalhada, o conteúdo incluiu estudos de caso e referências a artigos recentes que fundamentam o progresso da área.

**A aula e sua abordagem técnica**

Jim Fan apresentou uma visão estratégica e técnica sobre o Project GR00T, um projeto inovador da NVIDIA que busca desenvolver um agente robótico generalista. Diferentemente de abordagens tradicionais que treinam robôs para tarefas específicas, o GR00T se destaca por sua capacidade de operar de forma eficaz em diferentes corpos, conjuntos de habilidades e ambientes, tanto simulados quanto reais.

Um dos pontos centrais da apresentação foi o pipeline de dados necessário para robôs generalizáveis. Aprendi que, ao contrário de grandes modelos de linguagem, que utilizam dados amplos e não filtrados da internet, robôs precisam de fontes de dados altamente específicas e estruturadas. Por exemplo, datasets curados e simuladores de alta fidelidade desempenham um papel crítico no treinamento de robôs, enquanto dados da internet, por sua falta de relevância e ruído, não fazem parte desse processo.

Outro aspecto importante discutido foi o uso intensivo de simulações. Segundo Fan, os robôs passarão a maior parte de suas vidas treinando em ambientes simulados. Motores físicos avançados permitem movimentos e interações realistas, enquanto simuladores possibilitam escalabilidade e eficiência, reduzindo custos e acelerando o aprendizado.

DIA 10

**A Ciência Aberta e os Modelos Fundacionais: Reflexões da Aula de Percy Liang**

No dia 18 de novembro, Percy Liang, professor da Stanford University, ministrou a aula *"Open-Source and Science in the Era of Foundation Models"*. A apresentação abordou a relevância do código aberto no avanço da ciência e no desenvolvimento de modelos fundacionais, oferecendo uma visão técnica e estratégica sobre como a transparência e a acessibilidade podem impulsionar a inovação.

**Fundamentos dos Modelos Fundacionais e Acessos Existentes**

Durante a aula, discutiu-se os diferentes níveis de acesso disponíveis para modelos fundacionais, destacando práticas comuns, como:

1. **Consulta via API sem visibilidade interna** – Amplamente utilizada em plataformas como a API GPT da OpenAI, essa abordagem permite interagir com o modelo sem expor os dados de treinamento ou a arquitetura.
2. **Acesso ao código-fonte e dados de treinamento** – Comum em projetos de código aberto, esta prática permite um estudo detalhado e replicável.
3. **Uso de pesos pré-treinados para personalização** – Disponibilizados por frameworks como Hugging Face, esses pesos permitem transferência de aprendizado e ajustes finos para tarefas específicas.

Por outro lado, Percy deixou claro que ajustar diretamente os parâmetros neurais do modelo por meio de hardware não é um nível de acesso prático ou existente. O ajuste dos parâmetros ocorre exclusivamente no nível do software.

**Código Aberto: Potencial e Limitações**

Liang também explorou argumentos a favor do acesso de código aberto, destacando que ele:

* **Facilita inovação e colaboração** – Ao permitir a modificação e redistribuição sem restrições, o código aberto estimula avanços tecnológicos.
* **Promove a reprodutibilidade na pesquisa** – A estabilidade oferecida pelos modelos abertos elimina dependências de APIs, muitas vezes sujeitas à obsolescência.
* **Amplia a transparência** – Pesquisadores podem analisar criticamente todos os aspectos de um modelo, desde a arquitetura até os dados utilizados.

Por outro lado, a ideia de que o código aberto garantiria um alinhamento ético perfeito em todos os domínios foi refutada. Percy argumentou que questões éticas são complexas e dependem de variáveis culturais, contextuais e do propósito de aplicação.

**Memória em Simulações de Agentes**

Outro ponto discutido foi o uso de memórias em simulações de agentes. Percy detalhou critérios úteis para a recuperação de memórias, como:

* Proximidade temporal com o presente.
* Relevância de eventos marcantes.
* Alinhamento com o contexto atual.

Entretanto, ele descartou a seleção aleatória como critério válido, já que a efetividade da memória está diretamente ligada à relevância para os objetivos do agente.

DIA 11

**Medindo Capacidades de Agentes e a Política de Escalonamento Responsável da Anthropic**

No dia 25 de novembro, a aula foi ministrada por Ben Mann, da Anthropic. A apresentação abordou conceitos fundamentais sobre como medir as capacidades de agentes de IA e os princípios por trás da Política de Escalonamento Responsável (RSP) desenvolvida pela Anthropic.

**O que é a Política de Escalonamento Responsável (RSP)?**

A Política de Escalonamento Responsável da Anthropic é uma abordagem estruturada para gerenciar os riscos associados ao desenvolvimento de sistemas de IA cada vez mais avançados. Inspirada por protocolos de biossegurança, a RSP organiza os modelos de IA em Níveis de Segurança de IA (ASLs), categorizando-os com base em suas capacidades e nos riscos potenciais associados.

Esses ASLs vão desde o ASL-1, para modelos com capacidades básicas, até níveis mais elevados para sistemas mais avançados. A política inclui mecanismos de avaliação de capacidades, implementação de salvaguardas específicas para cada nível e resposta dinâmica a gatilhos que possam demandar ajustes nos níveis de segurança.

Entre os objetivos da RSP estão:

* **Aprimorar decisões seguras:** Garantir que os sistemas de IA sejam desenvolvidos e utilizados de maneira segura e ética.
* **Fornecer uma estrutura para decisões complexas:** Criar diretrizes claras para lidar com desafios de segurança.
* **Inspirar a indústria:** Compartilhar boas práticas para que outros profissionais e formuladores de políticas possam implementar protocolos semelhantes.

**Por que medir capacidades de agentes é essencial?**

Medir as capacidades de agentes de IA é fundamental para garantir que eles se comportem de forma confiável e segura. No entanto, a aula destacou um desafio importante: *“benchmarks não duram”*. Essa frase ressalta a natureza dinâmica do campo da IA, onde avanços rápidos tornam benchmarks obsoletos em pouco tempo.

Isso ocorre porque:

1. **Melhoria contínua dos modelos:** Modelos mais avançados superam rapidamente os benchmarks existentes.
2. **Novos desafios:** A pesquisa constantemente cria benchmarks mais complexos para refletir os avanços da tecnologia.
3. **Benchmarks como etapa:** Em vez de padrões permanentes, os benchmarks são ferramentas transitórias para medir o progresso.

**Riscos e Salvaguardas em Agentes de LLM**

A aula também abordou os riscos associados a agentes baseados em Modelos de Linguagem de Grande Escala (LLMs). Entre os riscos mais relevantes estão:

* **Melhoria autônoma:** Agentes que podem aprimorar suas capacidades sem supervisão humana, o que pode gerar consequências inesperadas.
* **Prompt injection:** Ataques que exploram vulnerabilidades nas instruções do modelo, induzindo comportamentos não intencionais.
* **Acesso a informações sensíveis:** Riscos de exploração do agente como vetor para atividades maliciosas.

Por outro lado, aspectos como a capacidade de realizar tarefas rapidamente foram reconhecidos como benefícios, e não como riscos de segurança.

**Como a hierarquia de instruções melhora a robustez?**

Outro ponto relevante discutido foi como a hierarquia de instruções nos prompts pode tornar os modelos mais robustos. Essa técnica ensina os modelos a priorizar instruções privilegiadas do sistema e ignorar comandos desalinhados ou maliciosos. Com isso, o modelo consegue distinguir melhor entre entradas legítimas e potenciais ataques, garantindo maior segurança e confiabilidade.

DIA 12

**Rumo a Agentes de IA Seguros e Confiáveis e um Caminho para Políticas de IA Baseadas em Ciência e Evidências**

No dia 2 de dezembro, a apresentação foi ministrada pela professora Dawn Song, da Universidade da Califórnia, Berkeley. A aula abordou questões cruciais no campo da Inteligência Artificial, destacando os desafios e avanços na construção de agentes de IA seguros e confiáveis, além de discutir caminhos para políticas de IA baseadas em ciência e evidências.

**Tópicos Abordados**

**1. Segurança e Confiança em Agentes de IA**  
A construção de agentes de IA seguros e confiáveis é um dos pilares para garantir que a tecnologia avance sem causar danos. A aula destacou a diferença entre *AI Safety* e *AI Security*:

* **AI Safety** foca em prevenir que o sistema cause danos ao ambiente ou à sociedade, abordando problemas como comportamento inesperado e tomada de decisões éticas.
* **AI Security**, por outro lado, concentra-se na proteção dos sistemas contra ameaças externas, como ataques cibernéticos e adulteração.

Um dos principais desafios discutidos foi a complexidade de projetar sistemas que equilibrem segurança com funcionalidade, garantindo que comportamentos indesejados ou ataques adversariais sejam mitigados.

**2. Exemplos Adversariais e Vulnerabilidades**  
Exemplos adversariais — entradas modificadas sutilmente para enganar modelos de aprendizado profundo — foram apresentados como um exemplo de como os sistemas de IA ainda são vulneráveis. Pequenas perturbações nas imagens podem levar a erros significativos na classificação de modelos, destacando a necessidade de desenvolver sistemas mais robustos e resilientes.

**3. Segurança em Agentes LLMs (Large Language Models)**  
A aula também explorou a segurança de agentes que utilizam modelos de linguagem (LLMs). A diferença entre segurança de agentes baseados em LLMs e segurança típica de LLMs foi discutida em profundidade:

* A segurança de agentes LLMs envolve proteger ações executadas no mundo real, como comandos e interações com sistemas externos.
* A segurança típica de LLMs concentra-se no conteúdo gerado pelo modelo, como evitar a disseminação de desinformação ou linguagem prejudicial.

**4. Injeção Indireta de Prompts e Desafios de Segurança**  
Outro ponto crítico abordado foi a injeção indireta de prompts, em que atacantes inserem instruções maliciosas em dados externos processados pelo modelo, resultando em comportamentos não intencionais. Esse tipo de ataque destaca a vulnerabilidade dos sistemas de IA a fontes externas não confiáveis e a importância de políticas de segurança rigorosas.

**5. Assimetria Entre Ataques e Defesas**  
Por fim, discutiu-se a assimetria entre ataques e defesas em segurança de IA: enquanto ataques são fáceis de escalar e têm alta tolerância ao fracasso, as defesas precisam ser extremamente confiáveis e são intensivas em recursos. Essa disparidade exige um investimento contínuo em pesquisa e inovação para fortalecer as defesas.

**Reflexão e Caminhos Futuros**

A aula foi um convite a refletir sobre o impacto da IA em nosso cotidiano e a necessidade de construir sistemas que sejam seguros, transparentes e confiáveis. O caminho para uma IA responsável passa não apenas pelo desenvolvimento tecnológico, mas também por políticas baseadas em evidências científicas que guiem a aplicação dessa tecnologia de forma ética e segura.

O desafio está posto: como equilibrar inovação e responsabilidade para garantir que a IA seja uma força para o bem na sociedade? A aula da professora Dawn Song trouxe insights valiosos que continuarão guiando minha jornada no estudo da segurança e confiabilidade em IA.