Resumo executivo

"A inteligência artificial não é um substituto da inteligência humana;
É uma ferramenta para ampliar a criatividade e a engenhosidade humanas".
Fei-Fei Li[22]

LLM: contexto, definição e regulamentação

1. A Inteligência Artificial Generativa (GenAI) e, dentro dela, os modelos de linguagem de grande escala (LLMs) representam um avanço significativo no campo da IA, definindo uma nova geração de interface homem-máquina em que a comunicação é feita por meio de linguagem natural e com aplicações revolucionárias em todos os setores, incluindo educação, saúde, finanças e comércio. No entanto, seu desenvolvimento e uso também trazem riscos e desafios significativos que precisam ser abordados.

2. Os LLMs são modelos de IA treinados para reconhecer, gerar, traduzir e resumir grandes quantidades de texto. Eles usam arquiteturas como transformers e são treinados em vastos conjuntos de dados para aprender padrões e estruturas linguísticas. Sua eficácia depende do tamanho em termos de número de parâmetros, da estrutura, da diversidade dos dados de treinamento e da sofisticação de seus algoritmos.

3. A evolução dos LLMs tem sido muito rápida, desde os primeiros modelos baseados em regras até os modelos atuais baseados em transformers. Os marcos importantes incluem a introdução da arquitetura do transformer e dos mecanismos de autocorreção, além dos primeiros LLMs comerciais, como o GPT. O ano de 2023 foi fundamental, com maior acessibilidade, contribuições globais e a proliferação de LLMs de código aberto.

4. Os LLMs têm inúmeras aplicações, como criação e aprimoramento de conteúdo, análise e organização de informações, interação e automação de tarefas. Com o surgimento de LLMs multimodais, novas possibilidades estão se abrindo na geração de conteúdo audiovisual e experiências interativas ricas.

5. Os órgãos reguladores estão tomando medidas para lidar com os riscos e as oportunidades da IA, com iniciativas como o AI Act da UE, o Bill of Rights dos EUA e a Declaração de Bletchley. Alguns dos principais requisitos incluem transparência, privacidade, imparcialidade, segurança, responsabilidade e supervisão humana.

Desenvolvimento e implantação de LLMs

6. O desenvolvimento de LLMs envolve vários componentes e decisões essenciais, como seleção e pré-processamento de dados, tokenização e embeddings, pré-treinamento, quantização e fine-tuning. Em particular, o alto custo do treinamento geralmente leva à opção de usar um modelo pré-treinado ou um modelo de código aberto e simplesmente fazer o fine-tuning com dados relativos ao aplicativo a ser desenvolvido. A implementação requer considerações sobre integração, monitoramento e questões éticas e legais.

7. O treinamento de modelos é um aspecto crucial que influencia sua eficácia. Fatores como a quantidade e a qualidade dos dados de treinamento, a arquitetura do modelo e os algoritmos de aprendizado usados podem ter um impacto significativo sobre o desempenho e a generalização de um LLM.

8. A arquitetura mais comum para os LLMs são os transformers, que usam mecanismos de autoatenção que permitem que o modelo encontre relações entre diferentes partes do texto, processe-o e gere um novo texto. Eles demonstraram um desempenho excepcional em uma variedade de tarefas de processamento de linguagem natural. Variantes e extensões buscam melhorar sua eficiência e escalabilidade.

[22] Fei-Fei Li (nascido em 1976). Co-diretora do Stanford Institute for Human-Centered Artificial Intelligence e IT Professor na Graduate School of Business, conhecida por criar a ImageNet e a AI4ALL, uma organização sem fins lucrativos que trabalha para aumentar a diversidade e a inclusão no campo da inteligência artificial.

9. O LLMOps é uma metodologia para gerenciar o ciclo de vida completo dos LLMs, abordando desafios como a gestão de grandes volumes de dados, a escalação de recursos computacionais[23], o monitoramento e a manutenção, o versionamento e a reproducibilidade.

10. Os principais desafios dos LLMs incluem vieses e alucinações, falta de explicabilidade e transparência, qualidade e acessibilidade dos dados, problemas de privacidade e segurança e alto consumo de recursos. Há também desafios de dependência, riscos de uso malicioso, problemas de propriedade intelectual e escalabilidade.

Estrutura de validação de LLMs

11. A validação dos LLMs é fundamental para garantir seu uso seguro e responsável, e uma perspectiva ampla deve ser adotada para abranger os vários riscos associados a eles. Uma estrutura de validação multidimensional deve abranger aspectos como risco de modelo, gestão de dados, segurança cibernética, riscos legais e operacionais, ética e reputação.

12. A validação dos LLMs deve ser articulada por meio de uma combinação de métricas quantitativas e técnicas de avaliação humana. A seleção das técnicas dependerá das características do caso de uso, como o nível de risco, a exposição pública, o processamento de dados pessoais e a linha de negócios.

13. As tendências emergentes na validação de LLMs incluem explicabilidade[24], o uso de LLMs para explicar outros LLMs, pontuação por atribuição, validação contínua, abordagens colaborativas, engenharia de prompts, alinhamento ético e regulatório e técnicas de desaprendizagem de máquina (machine unlearning).

Estudo de caso

14. O estudo de caso apresentado ilustra a aplicação de um framework de validação personalizado de um chatbot de política interna de uma empresa. O processo envolveu a definição do caso, o projeto da abordagem de validação, a execução de testes quantitativos e qualitativos e a interpretação dos resultados.

15. Os resultados da validação do chatbot mostraram um desempenho geral satisfatório, com pontos fortes em precisão, consistência, adaptabilidade e escalabilidade. Foram identificadas áreas de aprimoramento em explicabilidade, mitigação de viés e segurança. Foi recomendado prosseguir com a implementação, aplicando as melhorias sugeridas e estabelecendo um plano para monitoramento e melhoria contínuos.

Conclusão

16. Em conclusão, os LLMs têm um potencial significativo para transformar vários setores, mas seu desenvolvimento e implantação também trazem desafios significativos em áreas como transparência, imparcialidade, privacidade e segurança. Para aproveitar os benefícios dos LLMs de forma responsável, é fundamental estabelecer uma estrutura robusta de governança de IA que aborde esses desafios de forma abrangente, incluindo uma abordagem rigorosa e multidimensional de validação que cubra todo o ciclo de vida dos modelos. Essa é a única maneira de garantir que os LLMs sejam confiáveis, éticos e alinhados com os valores e objetivos das organizações e da sociedade em geral.

[23] Management Solutions (2022).). AutoML, rumo à automação dos modelos.
[24] Management Solutions (2023).). Explainable Artificial Intelligence (XAI): desafios na interpretabilidade de modelos.