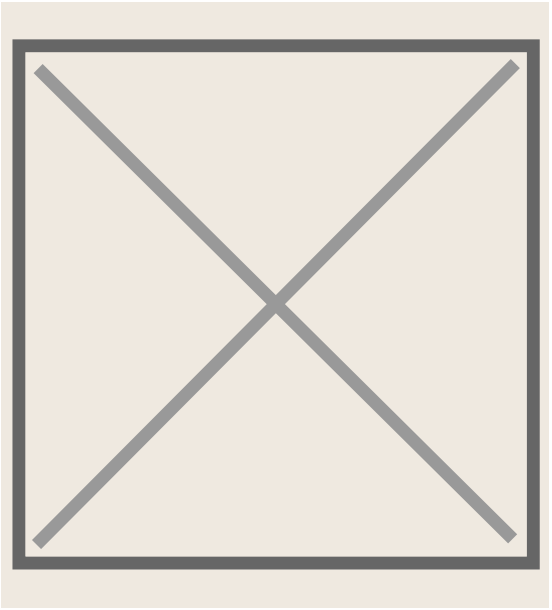
Primeira IA capaz de ser aprovada no Enam é lançada no Brasil

08/08/2024

No último dia 1º de agosto, a empresa americana Anthropic — uma das principais competidoras da OpenAI, criadora do ChatGPT — liberou acesso a seus modelos de linguagem ao público brasileiro. De acordo com diversas métricas, seu melhor modelo, o Claude 3.5 Sonnet, seria superior a todos os seus principais concorrentes, como o GPT-4o (a versão atual do ChatGPT), o Gemini 1.5 Pro (melhor modelo da Google) e o Llama 3.1 405b (melhor modelo da Meta, dona do Facebook, Instagram e WhatsApp).



Tendo acesso ao Claude 3.5 Sonnet, fui testar as suas habilidades jurídicas. Como há mais de um ano e meio há notícia de que o ChatGPT é capaz de passar na prova da OAB, optei por um desafio maior: as duas provas já aplicadas do Exame Nacional de Magistratura (Enam). [1] Por ser uma prova voltada ao concurso público da magistratura, as questões do Enam são bem mais difíceis que as da OAB.

O exame é formado por 80 questões e, para serem aprovados no exame, candidatos de ampla concorrência precisam acertar 70% das questões, o que equivale a 56 acertos. Candidatos pretos, pardos e indígenas têm de acertar 50% da prova — 40 questões. Entre os 40 mil inscritos, apenas 6.761 foram aprovados — uma taxa de aprovação de cerca de 17%.

Eu já havia testado outros modelos de linguagem nessa prova e alguns chegavam perto, mas nenhum era capaz de ser aprovado. Pensei que aconteceria o mesmo com o Claude 3.5 Sonnet. Para a minha surpresa, o modelo pontuou 59 na primeira prova e 56 na segunda, o suficiente para ser aprovado no Enam em ambas as oportunidades.

Além do Claude 3.5 Sonnet, testei outros 29 modelos de linguagem diferentes: [2] 13 não são capazes de acertar metade de nenhuma das duas provas [3]; nove acertam metade da primeira prova, mas não da segunda. [4] Na tabela abaixo, listo, em ordem de acertos, os oito modelos capazes de acertar metade de ambas as provas. [5]

Modelo	Desenvolvedora	Data de lançamento	Nota na 1ª prova	Nota na reaplicação	% de acertos
Claude 3.5 Sonnet	Anthropic (Estados Unidos)	20/06/2024	59	56	71,9
GPT 4o	OpenAI (Estados Unidos)	13/05/2024	55	54	68,1
Claude 3 Opus	Anthropic (Estados Unidos)	04/03/2024	53	53	66,3
Sabiá 3	Maritaca AI (Brasil)	05/07/2024	51	54	65,6
Llama 3.1 405b	Meta (Estados Unidos)	23/07/2024	52	46	61,3
Gemini 1.5 Pro	Google (Estados Unidos)	24/05/2024	48	42	56,2
Qwen 2 72b	Alibaba (China)	06/06/2024	47	43	56,2
Llama 3.1 70b	Meta (Estados Unidos)	23/07/2024	45	43	55,0

Todos esses modelos são bastante recentes. Com exceção do Claude 3 Opus, lançado em março (mas também disponibilizado no Brasil apenas ontem), todos os outros modelos foram lançados nos últimos três meses. Esse fato reforça a constatação de que os modelos de linguagem seguem melhorando e devem melhorar ainda mais. Se hoje apenas o Claude 3.5 Sonnet é capaz de ser aprovado no Enam, as próximas versões do ChatGPT, do Sabiá, do Llama e do Qwen — além de outros modelos não listados — muito provavelmente também terão essa habilidade.

Além disso, chama atenção nesse ranking o fato de que a Maritaca AI, empresa brasileira fundada por pesquisadores da Unicamp, tem sido capaz de bater de frente com as Big Techs americanas — ao menos no que diz respeito à avaliação do conhecimento jurídico brasileiro dos modelos de linguagem. Esse resultado demonstra o potencial de se treinar os modelos de linguagem especificamente em língua portuguesa, se o objetivo for obter melhores resultados em nossa língua. Esse foco, junto de um treinamento especializado em conhecimentos jurídicos, pode fornecer resultados bastante superiores aos obtidos pelas empresas estadunidenses [6], e muito provavelmente com um custo bem menor de treinamento.

Com maior investimento na área — como o do Plano IA para o Bem de Todos, anunciado pelo Ministério de Ciência, Tecnologia e Inovação essa semana —, o desenvolvimento de modelos de linguagem jurídicos brasileiros pode enfim

deslanchar. Hoje, isso significa apenas maiores notas nas avaliações desses modelos, mas, com algum tempo, essas ferramentas podem acelerar e qualificar bastante o trabalho de operadores do direito, além de contribuir para tornar informações jurídicas mais acessíveis à população.

---

[1] Além da primeira edição, ocorrida em 14 de abril desse ano, houve também em 19 de maio uma reaplicação da primeira edição em Manaus, devido à falta de energia elétrica na cidade na data da primeira prova.

[2] GPT 3.5 Turbo, GPT 4o, GPT 4o mini, Gemini 1.0 Pro, Gemini 1.5 Flash, Gemini 1.5 Pro, Gemma 7b, Gemma 2 9b, Gemma 2 27b, Llama 3 8b, Llama 3 70b, Llama 3.1 8b, Llama 3.1 70b, Llama 3.1 405b, Mistral 7b v0.3, Mixtral 8x7b, Mixtral 8x22b v0.1, Mistral Nemo, Mistral Large 2, Sabiá 2 Medium, Sabiá 2 Small, Sabiá 3, Qwen 2 7b, Qwen 2 72b, Claude 3 Haiku, Claude 3 Sonnet, Claude 3 Opus, Claude 3.5 Sonnet, Yi Large e DeepSeek V2.

[3] GPT 3.5 Turbo, Gemini 1.0 Pro, Gemini 1.5 Flash, Gemma 7b, Gemma 2 9b, Gemma 2 27b, Llama 3 8b, Llama 3.1 8b, Mixtral 8x7b, Mistral 7b, Mistral Nemo, Sabiá 2 Small e Qwen 2 7b.

[4] GPT 4o Mini, Llama 3 70b, Mixtral 8x22b v0.1, Mistral 2 Large, Sabiá 2 Medium, Claude 3 Haiku, Claude 3 Sonnet, Yi Large e DeepSeek V2.

[5] Os resultados completos podem ser encontrados em: <https://github.com/bdcdo/enam-exams>

[6] A própria Maritaca AI já demonstrou isso com seu modelo de linguagem Juru, cujo desempenho supera o de modelos de linguagem não especificamente treinados em textos jurídicos. Para saber mais, acesse: <https://arxiv.org/html/2403.18140v1>.

Fonte: https://conjur.jumps.com.br/2024-ago-08/primeira-ia-capaz-de-ser-aprovada-no-enam-e-lancada-no-brasil/