XXII Congresso Brasileiro de Mecânica dos Solos e Engenharia Geotécnica
XI Simpósio Brasileiro de Engenheiros Geotécnicos Jovens
24 a 28 de agosto de 2026 – Brasília/DF

DOI: 10.47094/COBRAMSEG2026.2/8

Avaliação da performance de diferentes modelos de inteligência artificial generativa na resolução de problemas de cálculo na Mecânica dos Solos

Daniel de Abreu Henriques
Analista de Projetos, Tractebel, Belo Horizonte, Brasil, danielabreuh@gmail.com

Lucas Martins Correa
Engenheiro Geotécnico, Tractebel, Belo Horizonte, Brasil

William da Fonseca Simões
Assistente Técnico, Tractebel, Belo Horizonte, Brasil

RESUMO: A inteligência artificial (IA) tem se tornado um dos instrumentos mais utilizados para tarefas do cotidiano profissional, desde a realização de tarefas mais simples até mais complexas. Nesse contexto, as IA's generativas têm aparecido em posição de destaque, pela sua facilidade de utilização em plataformas consolidadas, a partir de simples prompts de comando. No campo da engenharia geotécnica, levanta-se a questão sobre a confiabilidade desses modelos no auxílio teórico e numérico. Com isso, este artigo visou a avaliar a performance de diferentes modelos de IA na resolução de questões associadas à geotecnia. A partir de um prompt inicial de configuração, foram definidas 15 questões de cálculo, retiradas da literatura pertinente da Mecânica dos Solos e divididas entre os níveis básico, intermediário e avançado. Notou-se que as IA's apresentam, no geral, uma boa performance na resolução destes problemas geotécnicos. Modelos mais robustos e pagos possuem maior taxa de acerto, mas tem maior prejuízo no custo computacional, enquanto modelos mais simples e gratuitos ainda apresentam um desempenho adequado, com menor tempo de processamento das respostas. Apesar de serem úteis e poderem servir de consulta, essas ferramentas não são infalíveis e apresentam erros conceituais e numéricos, mesmo em questões de livros didáticos.

PALAVRAS-CHAVE: Inteligência artificial, performance, Mecânica dos Solos, questões de cálculo.

ABSTRACT: Artificial intelligence (AI) has become one of the most widely used tools for everyday professional tasks, ranging from the simplest to the most complex activities. In this context, generative AIs have gained prominence due to their easy use on established platforms, through simple command prompts. In the field of geotechnical engineering, questions arise regarding the reliability of these models in providing theoretical and numerical support. Therefore, this article aimed to evaluate the performance of different AI models in solving issues related to geotechnics. Starting from an initial configuration prompt, 15 calculation problems were defined, taken from relevant soil mechanics literature and divided into basic, intermediate, and advanced levels. It was observed that AIs generally perform adequately in solving these geotechnical problems. More robust and paid models have higher accuracy rates but come at a greater computational cost, while simpler and free models still deliver adequate performance with shorter response processing times. Although useful and suitable for consultation, these tools are not infallible and exhibit conceptual and numerical errors, even in textbook problems.

KEYWORDS: Artificial intelligence, performance, soil mechanics, calculation problems.

1 INTRODUÇÃO

A inteligência artificial (IA) tem se consolidado como uma ferramenta transformadora no cotidiano profissional de diversas áreas do conhecimento. Na engenharia geotécnica, há aplicações recentes que já exploram essas tecnologias, conforme discutido por Simões et al. (2025). No entanto, à medida que ferramentas generativas são incorporadas para o auxílio em tarefas de projeto ou dimensionamento, torna-se imperativo verificar a confiabilidade desses modelos em aplicações de cálculo desta natureza.

Embora o potencial das IA's seja amplamente discutido na literatura técnica, existem poucos estudos que quantifiquem a precisão desses modelos na resolução de problemas específicos da Mecânica dos Solos. Em áreas correlatas, como a Hidráulica, Simões et al. (2024) evidenciaram que, apesar de úteis, essas ferramentas apresentam erros conceituais e recomendam cautela em sua aplicação técnica. Essa incerteza reside no fato de que tais ferramentas estão suscetíveis a estes erros mesmo ao tratar de questões clássicas da área do conhecimento.

Com vistas a preencher tal lacuna, este artigo objetiva avaliar a performance de diferentes modelos de IA generativa na resolução de problemas de cálculo aplicados à geotecnia. O estudo compara o desempenho de plataformas amplamente difundidas, como o Gemini (Google), Copilot (Microsoft) e ChatGPT (OpenAI), ao submetê-las a 15 questões extraídas da literatura de referência, categorizadas em níveis de dificuldade básico, intermediário e avançado.

De forma complementar à taxa de acerto, a pesquisa analisa o custo computacional, mensurado pelo tempo de processamento das respostas. Busca-se, assim, fornecer um panorama sobre quais ferramentas oferecem o melhor equilíbrio entre precisão e eficiência para estudantes e profissionais da área, confrontando desde modelos gratuitos e simplificados até versões pagas e mais robustas.

2 MATERIAIS E MÉTODOS

2.1 Inteligência Artificial Generativa

Segundo Liu et al. (2023), um prompt pode ser definido como uma instrução ou conjunto de entradas textuais que reformulam uma tarefa para que ela possa ser solucionada por um modelo de linguagem pré-treinado. A qualidade da resposta dos modelos está diretamente relacionada com a formulação do *prompt* utilizado, que, com requisitos bem definidos, assegura que a IA generativa receba um direcionamento claro sobre as necessidades do usuário. Além disso, essa padronização garante que, neste estudo, todas as IA's partam da mesma base de comando, o que viabiliza a comparação direta entre os resultados obtidos.

Utilizou-se o seguinte comando neste estudo: *"Você será avaliado quanto à sua performance em questões relacionadas à engenharia geotécnica e Mecânica dos Solos. Responda as perguntas a seguir, relacionadas a tópicos simples a avançados, como um profissional geotécnico de ampla expertise em uma avaliação acadêmica de conhecimentos. Utilize como base conceitos traçados pela literatura pertinente e com base teórica relevante e demonstre conhecimento profundo sobre a geotecnia. As questões terão cálculos, apresente seu raciocínio e resuma suas respostas e os valores obtidos ao final."*

Para avaliar e comparar diferentes modelos, foram selecionadas três plataformas amplamente utilizadas por profissionais e estudantes: Gemini (Google), Copilot (Microsoft) e ChatGPT (OpenAI). Cada ferramenta foi submetida às mesmas questões sob o mesmo prompt. Adicionalmente, testou-se mais de um modelo dentro das plataformas Gemini e Copilot, para verificar o impacto da utilização de arquiteturas mais simples em comparação às mais robustas.

A Tabela 1 reúne as plataformas e os modelos utilizados neste estudo. Nota-se que o Copilot possui função integrada com o GPT-5 da OpenAI, que permite sua comparação com o GPT-5.2 nativo do ChatGPT.Essa comparação também pode ser feita nos modelos 2.5 Flash e 3 Pro da plataforma do Gemini.

Tabela 1. IAs utilizadas neste estudo.

Plataforma	**Gemini**		**Copilot**		**ChatGPT**
Modelo	2.5 Flash	3 Pro	Padrão	GPT-5	GPT-5.2

2.2 Problemas de cálculo da Mecânica dos Solos

O estudo abrange sete eixos temáticos fundamentais da Mecânica dos Solos: caracterização física, permeabilidade, tensões, adensamento, resistência ao cisalhamento, estados críticos e estabilidade de taludes. Tais tópicos constituem a base conceitual desta disciplina, sendo obrigatórios nos currículos de graduação e pós-graduação.

Para avaliar a eficácia das IA's, definiu-se uma hierarquia de complexidade em três níveis: básico, intermediário e avançado. A classificação e atribuição de pontos máximos para as questões se baseou na experiência dos autores e na sequência didática das disciplinas de Mecânica dos Solos I e II da Universidade Federal de Minas Gerais (UFMG), de forma a permitir a analisar variações de desempenho mediante o nível de complexidade das questões, uma vez que questões avançadas pressupõem o domínio de conceitos fundamentais.

A Tabela 2 detalha as questões por nível, tema e fonte bibliográfica. Foram utilizadas obras de referência clássica como Maranha das Neves (2016), Souza Pinto (2006), Craig (2005) e Das e Sobhan (2014), com um total de cinco questões por nível.

Devido às restrições de extensão do artigo, os enunciados integrais foram suprimidos, porém se preservou a descrição dos procedimentos de cálculo exigidos e a pontuação atribuída. É importante ressaltar que foram excluídos problemas dependentes de interpretação gráfica que continham figuras ou diagramas, o que limitou o escopo a enunciados estritamente textuais ou tabulares. Esses enunciados foram apresentados às IA's para testar o raciocínio lógico sem a interferência da interpretação de desenhos técnicos.

Tabela 2. Questões selecionadas.

Nível	ID	Pontos	Fonte	Tema	Descrição
Básico	5.5.E	10	Maranha das Neves (2016)	Caracterização física	Cálculo de índice de vazios, peso específico, índice de liquidez e plasticidade
	8.3.E	1		Permeabilidade	Cálculo da condutividade hidráulica
	6.1.E	1		Tensões	Cálculo da tensão vertical nos solos
	6.3.E	3		Tensões	Cálculo da tensão vertical nos solos
	12.6	5	Souza Pinto (2006)	Tensões	Cálculo do coeficiente de empuxo, tensões totais e efetivas
Intermediário	10.1 e 10.2	2		Adensamento	Cálculo do coeficiente de adensamento e tempo de recalque
	8.1.E	18	Maranha das Neves (2016)	Tensões	Cálculo de acréscimo de tensões e poropressão de acordo com o tempo de carregamento
	10.1.E	2		Resistência ao cisalhamento	Cálculo do ângulo de atrito de ruptura e de pico de *direct simple shear*
	10.3.E	1		Resistência ao cisalhamento	Cálculo do ângulo de atrito de triaxial CU
	10.6.E	5		Resistência ao cisalhamento	Cálculo do ângulo de atrito e tensões na ruptura de *direct simple shear*
Avançado	12.1.E	11		Estados críticos	Cálculo de parâmetros do estado crítico, Linha dos Estados Críticos (LEC), resistência crítica
	12.3.E	3		Estados críticos	Cálculo de parâmetros do estado crítico e resistência crítica
	Ex. 4.4	1		Estados críticos	Cálculo da resistência crítica

Nível	ID	Pontos	Fonte	Tema	Descrição
	Ex. 9.3	4	Craig (2004)	Estabilidade de taludes	Cálculo de fator de segurança e estabilidade por resistência residual e do estado crítico
	Ex. 14.1	5	Das e Sobhan (2014)	Empuxo	Cálculo de empuxo passivo em um muro de arrimo por diferentes teorias

É evidente que o espaço amostral não é extenso, de forma a ter um resultado absoluto sobre a superioridade ou inferioridade de um determinado modelo. As 15 questões selecionadas cobrem um *range* de temas mais relevantes e usuais na prática contemporânea de Mecânica dos Solos e os resultados devem ser entendidos como exploratórios, sendo apenas indicações do desempenho identificado de cada IA generativa.

2.3 Métricas de avaliação

A avaliação de performance dos modelos de IA generativa se baseou em duas métricas principais: a pontuação obtida e o tempo de processamento. A primeira quantifica diretamente a eficácia de cada modelo pela mensuração da taxa de acerto em cada questão proposta. Esse indicador permite inferir a confiabilidade técnica das respostas geradas frente aos gabaritos da literatura.

A segunda métrica se refere ao intervalo de tempo demandado para o processamento e a geração completa da resposta. Essa variável atua como um indicador indireto do custo computacional envolvido na interpretação do prompt, na execução dos cálculos e na redação da solução. Dessa forma, é possível estabelecer um comparativo entre a robustez dos modelos e a agilidade de entrega, o que permite identificar quais ferramentas demandam maior processamento.

Ressalta-se que, devido à ausência de logs nativos de tempo nas interfaces das plataformas avaliadas, a aferição foi realizada de forma independente pelos autores. Para garantir a isonomia do estudo, cronometrou-se, em segundos, o intervalo exato entre o envio do comando (envio da questão) e a conclusão da geração do texto pela inteligência artificial.

3 RESULTADOS E DISCUSSÃO

Com base na metodologia supracitada, todas as questões foram processadas pelos modelos de IA generativa, cujos resultados constam na Tabela 3. É incluída na tabela a discretização da pontuação individual para cada questão, bem como a totalização dos acertos sob a forma de mapa de calor.

Tabela 3. Pontuação obtida para cada modelo.

Nível	ID	Pontuação máxima	Pontuação				
			Gemini 2.5 Flash	Gemini 3 Pro	Copilot	Copilot GPT-5	ChatGPT 5.2
Básico	5.5.E	10	9	9	7	9	9
	8.3.E	1	1	1	1	1	1
	6.1.E	1	1	1	1	1	0
	6.3.E	3	3	3	3	3	3
	12.6	5	5	5	3	5	3
Interme-diário	10.1 e 10.2	2	2	2	2	2	2
	8.1.E	18	16	17	18	18	16
	10.1.E	2	2	2	2	2	2
	10.3.E	1	1	1	0	1	1
	10.6.E	5	2	5	0	2	2

Nível	ID	Pontuação máxima	Pontuação				
			Gemini 2.5 Flash	Gemini 3 Pro	Copilot	Copilot GPT-5	ChatGPT 5.2
Avançado	12.1.E	11	5	9	7	8	5
	12.3.E	3	2	3	1	1	1
	Ex. 4.4	1	1	1	1	1	1
	Ex. 9.3	4	4	4	3	2	0
Total		**72**	**55**	**64**	**49**	**56**	**46**

O melhor desempenho foi obtido pelo Gemini 3 Pro, com uma total de 64 pontos, correspondente a 89% da pontuação máxima. O Copilot GPT-5 foi o segundo melhor, com 56 acertos, o que equivale a 78%, seguido pelo Gemini 2.5 Flash, com 55 acertos ou 76%. O Copilot e o ChatGPT 5.2 apresentaram desempenhos inferiores, com acertos de 49 (68%) e 46 (64%), respectivamente.

Os menores índices de assertividade foram registrados nas questões 10.6.E de nível intermediário, 12.1.E de nível avançado e 12.3.E de nível avançado. A primeira abordou o tema de resistência ao cisalhamento, sendo exigido o cálculo do ângulo de atrito e das tensões na ruptura a partir de dados de um ensaio de *direct simple shear*. Já as questões avançadas trataram da Mecânica dos Estados Críticos e demandaram a determinação de parâmetros de estado crítico, da Linha de Estados Críticos (LEC) e da resistência crítica, sendo um tema reconhecidamente complexo da área da Mecânica dos Solos.

A Figura 1 apresenta a evolução da taxa de acerto por nível de dificuldade para cada modelo.

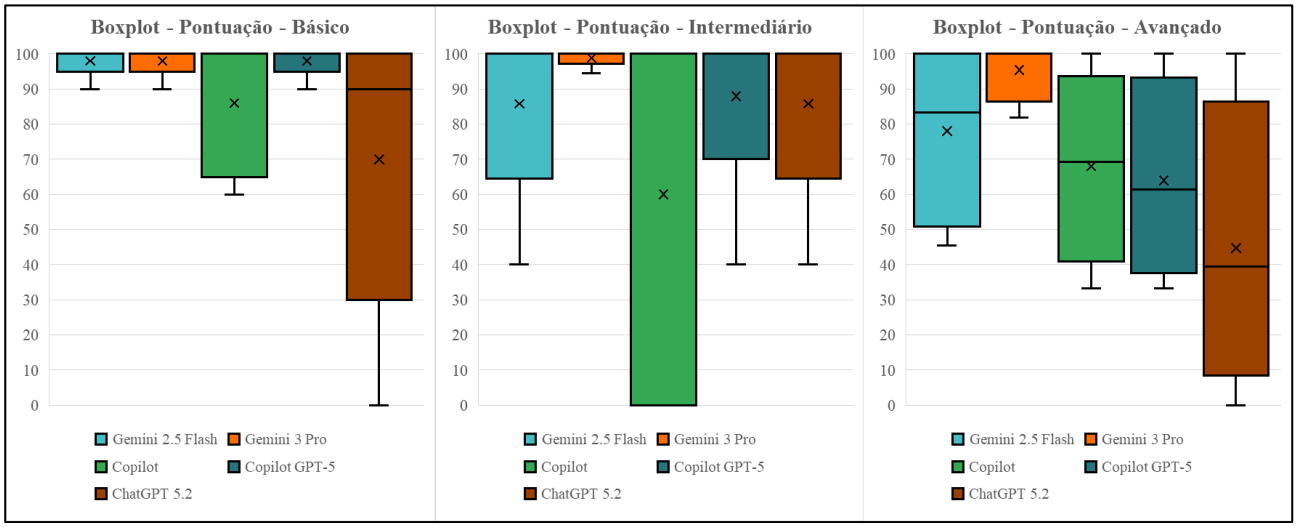


Figura 1. Evolução da taxa de acerto em percentual de cada modelo.

Confirma-se a tendência de acertos em questões de nível fácil e intermediário, ao passo que se reduzem nas questões mais difíceis. Ao comparar ChatGPT 5.2 e Copilot GPT-5, percebe-se que arquiteturas mais recentes não necessariamente implicam em melhores resultados, visto que a versão 5.2 do GPT, mais recente, apresentou desempenho inferior à versão 5, mais antiga. É evidente que o ChatGPT 5.2 teve o pior desempenho, com base na Figura 1, diante de erros mesmo em questões básicas. Em contrapartida, confirma-se que o Gemini 3 Pro obteve melhor desempenho quando comparado aos demais modelos.

Em um panorama geral, as IA's apresentaram coerência quanto à teoria dos exercícios avaliados e da Mecânica dos Solos clássica e dos estados críticos. Alguns erros, especialmente no Ex. 14.1 sobre empuxo passivo, consistem em problemas de formulação. Entretanto, o padrão de erro consiste em problemas matemáticos, seja por conversão de unidades ou na aplicação em fórmulas. Outra questão persistente é a geração de raciocínios equivocados, de forma a desviarem do gabarito das referências bibliográficas.

Vale ressaltar que a qualidade do *prompt* de entrada é importante para o resultado apresentado pela IA. O parágrafo utilizado para o presente trabalho foi eficaz em induzir o processamento das IA's de forma que

fossem estabelecidas linhas de raciocínio para cada questão, o que impacta significativamente o resultado. Assim, as IA's apresentaram, no geral, robustez teórica.

Além do número absoluto de acertos, é interessante verificar o tempo de processamento de cada modelo de IA. Esse resultado está explicitado na Tabela 4, que apresenta o tempo para cada questão, além do somatório total. Também é apresentada a eficiência geral de cada IA em termos de pontos obtidos por segundo de processamento, que consiste na média global da eficiência de cada questão.

Tabela 4. Tempo de processamento.

Nível	ID	Tempo (s)				
		Gemini 2.5 Flash	Gemini 3 Pro	Copilot	Copilot GPT-5	ChatGPT 5.2
Básico	5.5.E	15	59	30	25	48
	8.3.E	13	70	16	19	73
	6.1.E	19	55	21	20	91
	6.3.E	9	38	15	14	36
	12.6	22	47	20	17	54
Intermediário	10.1 e 10.2	15	56	23	22	50
	8.1.E	15	57	31	25	74
	10.1.E	15	70	22	19	67
	10.3.E	13	40	30	22	58
	10.6.E	15	96	28	55	72
Avançado	12.1.E	23	73	33	44	90
	12.3.E	17	70	32	19	57
	Ex. 4.4	16	58	33	35	69
	Ex. 9.3	66	66	33	32	56
	Ex. 14.1	30	88	28	29	65
Total (segundos e pontos/segundo)		303	943	395	397	960
		0.222	0.071	0.121	0.153	0.051

Nota-se, pela Tabela 4, que o Gemini 3 Pro, apesar de ser o modelo que mais teve acertos, possui o segundo pior desempenho do conjunto avaliado, quando se insere a métrica temporal na avaliação. Apesar da sua alta taxa de acertos, demanda um elevado tempo de processamento, o que o torna uma opção menos atrativa. Enquanto isso, a versão 2.5 Flash, da mesma plataforma, apresentou a melhor eficiência da relação acertos por tempo. Já as duas versões do Copilot, tanto a padrão quanto com GPT-5, tiveram desempenho intermediário.

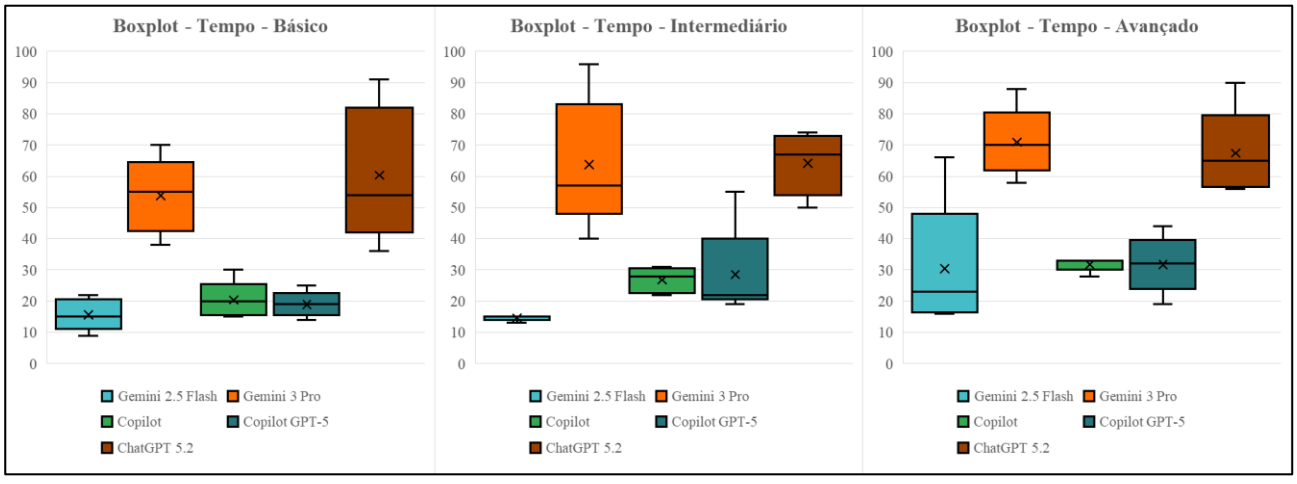


Figura 2. Evolução do tempo de processamento em segundos de cada modelo.

De forma similar ao desempenho em relação aos acertos absolutos, o ChatGPT 5.2 teve a pior performance em relação ao tempo de processamento, garantindo o pior desempenho geral neste artigo, como está evidente nas Figuras 1 e 2. Vale ressaltar que foi observado, durante a obtenção desses resultados, que o ChatGPT 5.2 foi o modelo que apresentou melhor detalhamento do raciocínio utilizado e a metodologia adotada nas questões, em conjunto do Gemini 3 Pro, que são os modelos mais robustos avaliados. Entretanto, no caso do GPT 5.2, o raciocínio apresentou erros fundamentais de cálculo, que inviabilizam sua utilização em comparação aos outros modelos.

É válido ressaltar que as plataformas avaliadas (Gemini, Copilot e ChatGPT), apesar de facilmente acessíveis, possuem diferenciação entre modelos gratuitos e pagos pelos usuários. Notou-se que modelos pagos (Gemini 3 Pro e Copilot GPT-5) apresentaram performance superior aos respectivos modelos gratuitos das mesmas plataformas (Gemini 2.5 Flash e Copilot). Apesar disso, a versão gratuita do Gemini ainda apresenta performance superior à versão paga do Copilot, tanto em termos de acertos quanto de tempo de processamento. O único modelo pago que foge dessa perspectiva é o ChatGPT 5.2, que apresentou o pior desempenho geral.

4 CONSIDERAÇÕES FINAIS

O objetivo deste trabalho foi verificar a acurácia da IA generativa na resolução de cálculos na Mecânica dos Solos, permitindo a verificação dos melhores modelos. Os resultados demonstram que as ferramentas de IA generativa possuem capacidade adequada para auxiliar na resolução de problemas da Mecânica dos Solos, desde que utilizadas com critério.

Observou-se uma clara correlação entre a robustez dos modelos e a precisão das respostas: versões mais sofisticadas, como o Gemini 3 Pro, alcançaram as maiores taxas de acerto. Entretanto, versões mais robustas, apesar de mais precisas, usualmente demandam maiores tempos de processamento, o que pode impactar a fluidez em fluxos de trabalho que exigem respostas rápidas. Para consultas rotineiras e com nível baixo de dificuldade, modelos intermediários e mais velozes se mostraram como alternativas adequadas.

O tempo de processamento, apesar de importante para comparação entre modelos de IA, não se compara ao tempo que levaria para um profissional ou um estudante realizar os mesmos cálculos de forma manual. Logo, esse tempo não é o gargalo para a avaliação, visto que qualquer modelo a ser utilizado é consideravelmente mais rápido que o processamento humano. Assim, deve-se dar preferência à métrica de acertos, sendo o tempo um fator secundário.

Com os resultados deste trabalho, é interessante levantar também a problemática e o impacto relacionados ao ensino da Mecânica dos Solos. Com modelos de IA muito precisos, como o Gemini 3 Pro com 89% de acertos, incluindo questões avançadas, as listas de exercícios tradicionais das disciplinas podem não ter a mesma relevância quanto antes da disseminação das IA's. No geral, entende-se que é essencial que alunos passem pelo aprendizado dos fundamentos e da lógica por trás da teoria na Mecânica dos Solos, mas a restrição de avaliações acadêmicas por cálculos pode estar se tornando obsoleta.

É fundamental destacar, todavia, que o uso dessas ferramentas não elimina a necessidade de verificação técnica. A identificação de erros em cálculos fundamentais demonstra que a IA deve atuar como um copiloto para o engenheiro. O julgamento crítico do profissional geotécnico permanece sendo o fator determinante para a aplicação segura dessas informações em projetos reais.

5 REFERÊNCIAS

Craig, R.F. (2004). *Craig's Soil Mechanics*. 7ª ed., Spon Press, Taylor & Francis Group, London, ISBN 0-415-32702-4), 458 p.

Das, B., Sobhan, K. (2014). *Fundamentos de engenharia geotécnica*. 8ª ed., Cengage Learning, São Paulo, ISBN 978-85-221-1824-3, 632 p.

Liu, P. et al. (2023). Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *ACM Computing Surveys*, v. 55, n. 9, p. 1-35, 2023. DOI:10.1145/3560815.

Maranha das Neves, E. (2016). *Mecânica dos Estados Críticos*: solos saturados e não saturados. 1ª ed, IST Press, Lisboa, ISBN 978-989-8481-48-1, 528 p.

Simões, A. L. A.; Fontes, D. C. H.; Anjos, I. S. M. (2024). Análise de erros conceituais produzidos por algoritmos de IA generativa quando consultados sobre hidráulica. In: *XVII Simpósio Recursos Hídricos do Nordeste*, João Pessoa.

Simões, W. F.; Costa, M.; Pereira, M. F.; Brasil, L. S. S. (2025). *Aplicação da inteligência artificial na gestão de segurança de barragens:* cenário atual e perspectivas para o futuro. In: *XXXV Seminário Nacional de Grandes Barragens*, Porto Alegre.

Souza Pinto, C. (2006). *Curso Básico de Mecânica dos Solos: em 16 aulas*. 3ª ed., Oficina de Textos, São Paulo, ISBN 978-85-86238-51-2, 363p.