Aplicação de modelos de linguagem na estratificação do risco cardiovascular: estudo de concordância com o cálculo manual pelo Escore de Framingham

Application of language models in cardiovascular risk stratification: a concordance study with manual calculation using the Framingham Risk Score.

(kaykyqueiroz18@gmail.com)

KAYKY QUEIROZ
Acadêmico de medicina pela Universidade Estadual do Mato Grosso (UNEMAT)

GIULIA MARTINAZZO ZIGIOTTO
Acadêmico de medicina pela Faculdade Assis Gurgacz (FAG)

RICARDO RODRIGUES DE OLIVEIRA JUNIOR
Acadêmico de medicina pela Universidade Estadual do Mato Grosso (UNEMAT)

ANTONIO MATHEUS MOURA SANT'ANNA SOUZA
Acadêmico de medicina pela Universidade Estadual do Mato Grosso (UNEMAT)

HEMILY PAOLA DE OLIVEIRA SCANDIANI
Acadêmico de medicina pela Universidade Estadual do Mato Grosso (UNEMAT)

LUANA ARAÚJO IURCKEVICZ
Acadêmico de medicina pela Universidade Estadual do Mato Grosso (UNEMAT)

MARIA EDUARDA D CASTRO LIMA
Acadêmico de medicina pela Faculdade Assis Gurgacz (FAG)

LETICIA DE GODOY DUARTE
Acadêmico de medicina pela Faculdade Assis Gurgacz (FAG)

FERNANDA KAORI YOKOYAMA
Acadêmico de medicina pela Faculdade Assis Gurgacz (FAG)

RESUMO

Objetivo: comparar modelos de linguagem com o cálculo manual do Escore de Framingham do SUS. **Metodologia:** estudo de concordância com 80 casos simulados. **Resultados e Discussão:** o ChatGPT 5.4 Thinking teve melhor desempenho (91,25% de acerto; Kappa quadrático 0,934), enquanto Instant, Gemini e Claude mostraram menor concordância e maior subestimação do risco, sobretudo em alto risco. **Considerações finais:** o uso deve ser apenas auxiliar e supervisionado.

Palavras-chave: Escore de Framingham; Pontuação de risco cardiovascular; Inteligência artificial

ABSTRACT

Objective: to compare language models with manual calculation of the SUS-adopted Framingham score. **Methodology:** agreement study with 80 simulated cases. **Results and Discussion:** ChatGPT 5.4 Thinking showed the best performance (91.25% accuracy; quadratic Kappa 0.934), whereas Instant, Gemini and Claude showed lower agreement and greater risk underestimation, especially in high-risk cases. **Final considerations:** use should remain auxiliary and supervised.

Keywords: Framingham Risk Score; Cardiovascular risk; Artificial intelligence.

INTRODUÇÃO

As doenças cardiovasculares permanecem como a principal causa de morte no mundo, o que torna a prevenção e a estratificação precoce do risco etapas fundamentais da prática clínica. Nesse contexto, instrumentos de estimativa de risco em 10 anos ajudam a identificar indivíduos com maior probabilidade de eventos futuros e orientam condutas preventivas mais adequadas, especialmente na atenção primária. No Brasil, o Escore Global de Risco de Framingham é utilizado em linhas de cuidado do Sistema Único de Saúde, com base em variáveis objetivas como idade, sexo, pressão arterial sistólica, colesterol total, colesterol HDL, tabagismo, diabetes e uso de anti-hipertensivo.

Paralelamente, a evolução recente da inteligência artificial, especialmente dos modelos de linguagem, ampliou seu uso em atividades acadêmicas e profissionais da área da saúde. Essas ferramentas vêm sendo aplicadas para organização de informações, síntese de conteúdo, apoio educacional e, cada vez mais, para tarefas relacionadas ao raciocínio clínico e à tomada de decisão. No entanto, organismos internacionais têm destacado que a incorporação dessas tecnologias em saúde deve ser acompanhada de avaliação crítica, validação e análise de segurança, uma vez que respostas fluentemente construídas não necessariamente correspondem a resultados clinicamente corretos ou reprodutíveis.

Diante desse cenário, torna-se relevante investigar se modelos de linguagem apresentam concordância satisfatória com métodos consagrados de estratificação cardiovascular. No presente estudo, buscou-se avaliar o desempenho comparativo de diferentes modelos de linguagem na estimativa do risco cardiovascular em 10 anos, utilizando como referência o cálculo manual do Escore Global de Risco de Framingham segundo a tabela por pontos adotada nas diretrizes brasileiras. Ao comparar os resultados numéricos e as categorias de risco produzidas pelos modelos com a estratificação manual, pretende-se discutir não apenas a precisão dessas ferramentas, mas também sua estabilidade, sua tendência a subestimar ou superestimar o risco e suas possíveis implicações para o uso seguro da inteligência artificial como apoio à prática clínica.

METODOLOGIA

Trata-se de um estudo metodológico, exploratório, de concordância, desenvolvido a partir de cenários clínicos simulados, com o objetivo de comparar o desempenho de diferentes modelos de linguagem na estratificação do risco cardiovascular em 10 anos. O padrão de referência adotado foi o cálculo manual do Escore Global de Risco de Framingham, utilizando exclusivamente a tabela por pontos disponibilizada na Linha de Cuidado do Sistema Único de Saúde (SUS). O estudo foi conduzido por comparação direta entre a classificação manual de referência e as respostas fornecidas pelos modelos de linguagem, sem utilização de dados de pacientes reais. O Escore Global de Risco de Framingham é disponibilizado pelo SUS como ferramenta de estimativa de risco cardiovascular em 10 anos em indivíduos sem doença cardiovascular estabelecida.

Foram elaborados 80 cenários clínicos fictícios e independentes, constituídos exclusivamente pelas variáveis necessárias ao cálculo do Escore Global de Risco de Framingham segundo a tabela por pontos adotada nas diretrizes brasileiras: idade, sexo, pressão arterial sistólica, colesterol total, colesterol HDL, tabagismo, diabetes e uso de anti-hipertensivo. Os cenários clínicos simulados foram construídos a partir de combinações aleatórias de valores clinicamente plausíveis das variáveis previstas no escore. A distribuição final das categorias de risco não foi previamente fixada por cotas, sendo definida pelo resultado da estratificação manual de cada caso segundo a tabela por pontos do Escore Global de Risco de Framingham. A validade dos cenários foi assegurada por construção, uma vez que todos os casos foram elaborados com base estrita nas variáveis oficialmente previstas pelo instrumento de referência, respeitando coerência clínica interna e compatibilidade com o domínio de

aplicação da tabela. Dessa forma, buscou-se garantir validade de conteúdo para a tarefa proposta, isto é, testar a capacidade dos modelos em reproduzir a estratificação derivada do escore e não realizar julgamento clínico amplo sobre pacientes reais. A distribuição manual final foi de 20 casos classificados como baixo risco, 22 como risco intermediário e 38 como alto risco, evidenciando predomínio de cenários de alto risco na amostra.

O padrão de referência foi o cálculo manual do risco cardiovascular em 10 anos pela tabela por pontos do Escore Global de Risco de Framingham adotada na Linha de Cuidado do SUS. Cada cenário foi pontuado manualmente antes da interação com os modelos de linguagem, obtendo-se o valor final do risco global e sua respectiva categoria: baixo risco, risco intermediário ou alto risco. A categoria "muito alto risco" não foi incluída no presente estudo, pois, na linha de cuidado brasileira, ela depende de critérios clínicos adicionais, como presença de doença aterosclerótica estabelecida e outros estratificadores clínicos, não sendo derivada exclusivamente do valor numérico do risco global em 10 anos. Como o objetivo do estudo foi avaliar a concordância dos modelos com o **cálculo do escore**, e não com julgamento clínico ampliado, fatores clínicos adicionais foram deliberadamente excluídos da análise. A própria Linha de Cuidado do SUS diferencia o cálculo do risco global dos estratificadores clínicos adicionais, incluindo condições que extrapolam a pontuação da tabela.

Nos cenários cujo resultado manual ultrapassou o limite superior da tabela oficial, expresso como ">30%", os valores foram truncados em 30% para fins de análise numérica, mantendo-se, contudo, a categoria "alto risco" na análise categórica. Essa decisão metodológica foi adotada para preservar comparabilidade entre o padrão de referência e os resultados fornecidos pelos modelos, uma vez que algumas inteligências artificiais retornaram valores contínuos acima de 30%.

Foram avaliados quatro modelos de linguagem, selecionados de forma intencional por representarem ecossistemas tecnológicos distintos e por estarem amplamente acessíveis ao pesquisador e ao usuário final em ambiente real de uso. Os modelos analisados foram: ChatGPT Plus 5.3 Instant, ChatGPT Plus 5.4 Thinking (modo de pensamento estendido), Gemini PRO 3 raciocínio e Claude Sonnet 4.6 versão gratuita. A inclusão de dois modos distintos do ChatGPT permitiu comparar o desempenho de um modelo orientado à resposta rápida com outro configurado para raciocínio mais deliberado, enquanto Gemini e Claude ampliaram a comparação para fornecedores diferentes, conferindo maior validade externa ao estudo. A escolha desses modelos não teve caráter exaustivo, mas pragmático: buscou-se comparar ferramentas de uso conversacional amplo, contemporaneamente disponíveis e potencialmente utilizadas em contextos acadêmicos e clínicos.

Cada cenário clínico foi submetido individualmente a cada modelo em sessão independente, sem continuidade de contexto entre os casos, a fim de reduzir viés de memória conversacional. Em todas as interações foi utilizado exatamente o mesmo prompt padronizado, elaborado para restringir a resposta à metodologia da tabela por pontos do Escore Global de Risco de Framingham adotada pelo SUS, proibindo o uso de outras calculadoras ou versões do escore e excluindo critérios clínicos adicionais não previstos no cálculo do risco global.

O prompt empregado foi o seguinte: "Você atuará exclusivamente como ferramenta de cálculo estatístico. Calcule o risco cardiovascular em 10 anos utilizando exclusivamente a metodologia da tabela por pontos do Escore de Risco Global de Framingham adotada nas Diretrizes Brasileiras (Linha de Cuidado do SUS). Utilize apenas os dados fornecidos abaixo. Não utilize outras versões do Framingham, não utilize ASCVD, não utilize Pooled Cohort Equations e não considere critérios clínicos adicionais como doença aterosclerótica estabelecida. Considere que este paciente NÃO possui diagnóstico prévio de doença cardiovascular. Forneça exclusivamente: o valor percentual final do risco em 10 anos (com uma casa decimal) e a classificação do risco como baixo risco, risco intermediário ou alto risco. Não inclua explicações, justificativas ou comentários adicionais. Idade: _. Sexo: _. Pressão arterial sistólica (mmHg): _. Colesterol total (mg/dL): _. HDL (mg/dL): _. Tabagismo: _. Diabetes: _. Uso de anti-hipertensivo: _."

A resposta de cada modelo foi registrada literalmente quanto ao valor percentual do risco em 10 anos e quanto à categoria atribuída. Não houve correção manual das respostas geradas pelos modelos, nem repetição dos casos para obtenção de valores "mais consistentes", preservando-se a saída originalmente fornecida por cada ferramenta sob condições padronizadas.

Os desfechos principais foram: concordância categórica entre cada modelo e o cálculo manual de referência; magnitude do erro numérico na estimativa do risco global; direção do erro, distinguindo subestimação e superestimação; e capacidade de identificação de pacientes classificados manualmente como alto risco.

Para a análise numérica, foi calculado o erro absoluto de cada caso. A partir dessa variável foram obtidos média, desvio-padrão, valor mínimo, valor máximo e distribuição por faixas de erro. Também foi calculada a diferença assinada entre os valores, permitindo avaliar a tendência de subestimação ou superestimação.

Para a análise categórica, cada resposta foi classificada em baixo risco, risco intermediário ou alto risco, sendo comparada à categoria manual correspondente. Foram calculados o percentual global de acerto, o percentual de erro, a frequência de subestimação e

superestimação. A concordância foi mensurada pelo coeficiente Kappa simples e pelos coeficientes Kappa ponderado linear e quadrático, considerando a natureza ordinal das categorias de risco. A interpretação dos valores de Kappa seguiu a classificação convencional: concordância muito fraca, fraca, moderada, substancial ou quase perfeita. Adicionalmente, foram calculadas sensibilidade e especificidade para identificação de alto risco, por se tratar da faixa clinicamente mais relevante para fins de prevenção cardiovascular.

O estudo utilizou exclusivamente cenários clínicos simulados, sem participação de seres humanos, sem análise de prontuários, sem utilização de dados secundários identificáveis e sem qualquer intervenção em animais. Por essa razão, não se caracterizou como pesquisa envolvendo seres humanos ou animais, não havendo necessidade de submissão ao Comitê de Ética em Pesquisa ou ao Comitê de Ética no Uso de Animais. A condução metodológica observou os princípios gerais de transparência, reprodutibilidade e integridade científica, com explicitação do padrão de referência, do prompt padronizado e dos critérios analíticos previamente definidos.

RESULTADOS E DISCUSSÃO

A comparação entre os modelos de linguagem demonstrou desempenho heterogêneo quanto à estimativa numérica do risco cardiovascular em 10 anos e à classificação categórica em baixo, intermediário e alto risco. De modo geral, o modelo **ChatGPT 5.4 Thinking** apresentou o melhor desempenho global, com menor erro absoluto médio, maior taxa de acerto categórico e maior concordância com o padrão de referência manual. Em contraste, o **ChatGPT Instant 5.3** apresentou o pior desempenho relativo, com menor concordância e importante tendência à subestimação do risco. O **Gemini 3 Raciocínio** apresentou desempenho intermediário, com concordância moderada a substancial, enquanto o **Claude Sonnet 4.6** mostrou acurácia categórica superior à do ChatGPT Instant, porém com padrão importante de reclassificação de casos manualmente altos para risco intermediário.

Tabela 1. Desempenho comparativo dos modelos de linguagem na estratificação do risco cardiovascular

Variável	ChatGPT Instant 5.3	ChatGPT 5.4 Thinking	Gemini 3 Raciocínio	Claude Sonnet 4.6
Média do erro absoluto	4,9	0,58	5	4,13
Desvio-padrão do erro absoluto	4,82	2,01	5,55	4
Faixa de erro ≤ 2 (%)	38,75	93,75	41,25	40
Faixa de erro > 2 e ≤ 5 (%)	26,25	3,75	22,5	28,75
Faixa de erro > 5 e ≤ 10 (%)	20	1,25	20	18,75
Faixa de erro > 10 (%)	15	1,25	16,25	12,5
Riscos acertados (%)	45	91,25	67,5	55
Riscos errados (%)	55	8,75	32,5	45
Erros por subestimação (%)	51,25	6,25	25	41,25
Erros por superestimação (%)	3,75	2,5	7,5	3,75
Sensibilidade para alto risco (%)	31,57	97,36	63,1	26,3
Kappa simples	0,217	0,864	0,509	0,363
Kappa ponderado linear	0,356	0,9	0,578	0,481
Kappa ponderado quadrático	0,481	0,934	0,642	0,623

Os resultados evidenciam maior proximidade entre o cálculo manual e o resultado gerado pelo **ChatGPT 5.4 Thinking**, cuja média do erro absoluto foi de 0,58 pontos percentuais, com 91,25% de acerto categórico e índices Kappa classificados como quase perfeitos. Além disso, esse modelo apresentou **sensibilidade de 97,36% para identificação de alto risco**, o maior valor entre os modelos analisados. Em contraste, o **ChatGPT Instant 5.3** apresentou erro absoluto médio de 4,9 pontos percentuais, com apenas 45,0% de concordância categórica e Kappa simples de 0,217, caracterizando concordância fraca. A sensibilidade do modelo para alto risco foi de **31,57%**, evidenciando baixa capacidade de reconhecer

corretamente essa categoria. O **Gemini 3 Raciocínio** apresentou desempenho intermediário, com erro absoluto médio de 5,0 pontos percentuais, acerto categórico de 67,5% e Kappa simples de 0,509, compatível com concordância moderada. A sensibilidade para identificação de alto risco foi de **63,1%**, indicando desempenho intermediário também nesse desfecho. Já o **Claude Sonnet 4.6** apresentou acerto categórico de 55,0%, Kappa simples de 0,363 e Kappa ponderado quadrático de 0,623, indicando concordância fraca a substancial a depender da ponderação utilizada. Embora o modelo tenha apresentado acerto global superior ao ChatGPT Instant, sua capacidade de identificar corretamente casos de alto risco foi baixa, com sensibilidade de **26,3%**.

Ao se analisar a distribuição do erro absoluto, verificou-se que o **ChatGPT 5.4 Thinking** concentrou a quase totalidade dos casos em erros de até 2 pontos percentuais, com 75 de 80 cenários (93,75%) nessa faixa. Nos demais modelos, a dispersão foi substancialmente maior. No **ChatGPT Instant 5.3**, apenas 38,75% dos casos apresentaram erro ≤ 2, enquanto 15,0% tiveram erro superior a 10 pontos percentuais. No **Gemini 3 Raciocínio**, 41,25% dos casos apresentaram erro ≤ 2, ao passo que 16,25% ultrapassaram 10 pontos percentuais. No **Claude Sonnet 4.6**, 40,0% ficaram na faixa ≤ 2, enquanto 12,5% apresentaram erro > 10. Esses achados sugerem que, além da concordância categórica, o modelo Thinking apresentou desempenho numérico mais estável e mais próximo do padrão manual.

Outro aspecto relevante foi a direção do erro. O **ChatGPT Instant 5.3** apresentou 41 subestimações e apenas 3 superestimações, o que significa que **93,2% dos seus erros ocorreram por subestimação do risco cardiovascular**. O **Gemini 3 Raciocínio** apresentou 20 subestimações e 6 superestimações, de modo que **76,9% dos erros** também se concentraram na subestimação. O **Claude Sonnet 4.6** apresentou 33 subestimações e 3 superestimações, mantendo o mesmo padrão de tendência a classificar o risco abaixo do observado no cálculo manual. O **ChatGPT 5.4 Thinking** apresentou menor frequência absoluta de erros e menor assimetria, com 5 subestimações e 2 superestimações. Em conjunto, esses resultados indicam que o principal problema metodológico dos modelos com pior desempenho não foi a superestimação do risco, mas sim a **subestimação**, fenômeno clinicamente mais sensível por potencialmente retardar intervenções preventivas.

No **ChatGPT Instant 5.3**, observou-se importante rebaixamento de casos manualmente classificados como risco intermediário para baixo risco, bem como de casos manualmente altos

para risco intermediário. Dos 38 cenários classificados manualmente como alto risco, apenas 12 foram corretamente identificados como altos, o que corresponde à sensibilidade de 31,57%.

No **ChatGPT 5.4 Thinking**, a matriz evidenciou forte concentração dos casos na diagonal principal, indicando elevada concordância com a classificação manual. Apenas 7 cenários foram classificados de forma divergente, e a maioria dos erros ocorreu por deslocamento de apenas uma categoria, sem padrão expressivo de rebaixamento severo. Dos 38 casos manualmente altos, 37 foram classificados corretamente, resultando em sensibilidade de 97,36%.

No **Gemini 3 Raciocínio**, a distribuição dos casos mostrou desempenho intermediário, com maior concentração na diagonal quando comparado ao ChatGPT Instant, porém ainda com reclassificação importante de cenários altos para intermediários e de intermediários para baixos. Dos 38 cenários classificados manualmente como alto risco, 24 foram reconhecidos como altos, correspondendo à sensibilidade de 63,1%.

No **Claude Sonnet 4.6**, observou-se padrão marcante de reclassificação de casos manualmente altos para risco intermediário, com 28 dos 38 cenários dessa categoria sendo deslocados para uma faixa inferior. Embora o modelo tenha apresentado acerto global superior ao ChatGPT Instant, sua capacidade de identificar corretamente alto risco foi baixa, com sensibilidade de 26,3%.

A análise conjunta dos resultados mostra que os modelos não se comportaram de maneira uniforme diante da mesma tarefa. O **ChatGPT 5.4 Thinking** destacou-se por apresentar alta concordância com o padrão de referência, pequena magnitude de erro, baixa frequência de reclassificações incorretas e **sensibilidade de 97,36% para alto risco**, o melhor desempenho entre os modelos avaliados. O **Gemini 3 Raciocínio** apresentou desempenho intermediário, com concordância superior à do Claude e do ChatGPT Instant, porém ainda com erro numérico e reclassificação clinicamente relevantes. O **Claude Sonnet 4.6** apresentou concordância categórica global de 55,0%, mas baixa capacidade de reconhecer alto risco, com sensibilidade de 26,3%, além de tendência importante a deslocar casos altos para a categoria intermediária. Já o **ChatGPT Instant 5.3** apresentou o pior desempenho global, com concordância fraca, erro numérico mais elevado, forte predomínio de subestimação e sensibilidade de apenas 31,57% para identificação de alto risco.

Do ponto de vista metodológico, os dados sugerem que a simples condição de "modelo de linguagem" não foi suficiente para predizer desempenho homogêneo na tarefa avaliada. Ao contrário, houve diferenças marcantes entre modelos da mesma família e entre plataformas distintas. Do ponto de vista prático, o achado mais relevante não foi apenas o percentual global

de acerto, mas a capacidade de identificar corretamente os casos de alto risco. Nesse aspecto, o **ChatGPT 5.4 Thinking** apresentou sensibilidade substancialmente superior à observada no **Gemini 3 Raciocínio**, no **ChatGPT Instant 5.3** e no **Claude Sonnet 4.6**, reforçando a heterogeneidade de desempenho entre os modelos avaliados.

CONSIDERAÇÕES FINAIS

O presente estudo evidenciou que modelos de linguagem de uso amplo apresentam desempenho heterogêneo quando aplicados à estratificação do risco cardiovascular pelo Escore Global de Risco de Framingham segundo a tabela por pontos adotada pelo SUS. Os achados reforçam que essas ferramentas não devem ser tratadas como equivalentes entre si, uma vez que houve diferenças relevantes de concordância, estabilidade e capacidade de reconhecer corretamente categorias clinicamente mais sensíveis, especialmente o alto risco.

De forma geral, os resultados sugerem que o uso de inteligência artificial generativa em tarefas estruturadas de apoio à decisão clínica pode ser promissor, mas permanece dependente de validação específica, padronização metodológica e supervisão profissional. A boa performance observada em alguns modelos não autoriza, por si só, sua utilização indiscriminada como substitutos de instrumentos validados, enquanto o desempenho inferior de outros reforça a necessidade de cautela diante do uso não supervisionado dessas tecnologias em contextos assistenciais.

Este estudo também destaca a importância de avaliar modelos de linguagem com base em instrumentos efetivamente utilizados no contexto nacional, e não apenas em métricas genéricas de desempenho. Ao utilizar o Escore Global de Risco de Framingham conforme a linha de cuidado do SUS, a pesquisa contribui para aproximar a discussão sobre inteligência artificial da realidade da prática clínica brasileira, particularmente na atenção primária e na prevenção cardiovascular.

Como limitação, trata-se de um estudo baseado em cenários simulados, o que impede extrapolação direta para populações reais e para situações clínicas mais complexas, nas quais fatores adicionais podem influenciar a estratificação do risco. Ainda assim, essa estratégia permitiu padronização das entradas, comparação controlada entre modelos e avaliação objetiva de concordância com o método de referência.

Dessa forma, conclui-se que modelos de linguagem podem apresentar utilidade potencial como ferramentas auxiliares em tarefas de cálculo e classificação de risco cardiovascular, mas seu emprego seguro depende de desempenho consistente, validação local

e clara delimitação de seu papel como suporte, e não como substituto, do julgamento profissional e dos instrumentos clínicos consolidados. Estudos futuros com bases clínicas reais, diferentes versões de escores e avaliação prospectiva do impacto dessas ferramentas na tomada de decisão poderão ampliar a compreensão sobre sua aplicabilidade em saúde.

REFERÊNCIAS

WORLD HEALTH ORGANIZATION. **Cardiovascular diseases (CVDs)**. Geneva: WHO, 2025. Disponível em: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-%28cvds%29. Acesso em: 9 mar. 2026.

BRASIL. Ministério da Saúde. Secretaria de Atenção Primária à Saúde**. Escore de risco global (ERG) de Framingham.** Brasília, DF: Ministério da Saúde, [s.d.]. Disponível em: https://linhasdecuidado.saude.gov.br/portal/obesidade-no-adulto/unidade-de-atencao-primaria/planejamento-terapeutico/escore-risco-global-framingham/. Acesso em: 9 mar. 2026.

WORLD HEALTH ORGANIZATION. **Ethics and governance of artificial intelligence for health: guidance on large multi-modal models.** Geneva: WHO, 2025. Disponível em: https://www.who.int/publications/i/item/9789240084759. Acesso em: 9 mar. 2026.

OPENAI. ***What is ChatGPT Plus?*** OpenAI Help Center, [2026]. Disponível em: https://help.openai.com/en/articles/6950777-what-is-chatgpt-plus