Revista Ibero-Americana de Humanidades, Ciências e Educação — REASE

OPEN ACCESS

doi.org/10.51891/rease.v12i5.26309

O *CHATGPT* É CONFIÁVEL PARA PRESCRIÇÃO DIETÉTICA? AVALIAÇÃO DE LLMS VIA *CHAIN-OF-THOUGHT* E LIMITAÇÕES NUTRICIONAIS NO CONTEXTO BRASILEIRO

CAN CHATGPT PROVIDE RELIABLE DIETARY ADVICE? EVALUATION OF LLMS VIA CHAIN-OF-THOUGHT AND NUTRITIONAL LIMITATIONS IN THE BRAZILIAN CONTEXT

¿ES CHATGPT CONFIABLE PARA LA PRESCRIPCIÓN DIETÉTICA? EVALUACIÓN DE LLMS VÍA CHAIN-OF-THOUGHT Y LIMITACIONES NUTRICIONALES EN EL CONTEXTO BRASILEÑO

Thalyson Gomes Nepomuceno da Silva[1]
Gustavo Augusto Lima de Campos[2]
Bonfim Amaro Júnior[3]
Ana Luiza Bessa de Paula Barros[4]

RESUMO: O uso de Grandes Modelos de Linguagem (LLMs) para prescrição dietética levanta questões de segurança clínica. Este estudo compara a eficácia do ChatGPT, Gemini e DeepSeek na geração de planos hipocalóricos para mulheres brasileiras com sobrepeso. Utilizando prompts one-shot e chain-of-thought, 150 planos alimentares foram gerados e analisados nutricionalmente via TBCA. Os resultados mostram variabilidade energética: o ChatGPT apresentou maior precisão (desvio <0,5%), enquanto os outros modelos superestimaram as calorias. Qualitativamente, todos os modelos falharam: a gordura saturada excedeu os limites em mais de 200% (chegando a 291% no DeepSeek) e foram observadas inadequações sistemáticas de micronutrientes, com o ferro atingindo apenas 55% a 61% das recomendações, além de déficit de cálcio. Conclui-se que a imprecisão qualitativa dos LLMs impede seu uso clínico autônomo, exigindo supervisão profissional.

Palavras-chave: Grandes Modelos de Linguagem. Prescrição Dietética. Chain-of-Thought. Inteligência Artificial. Segurança Clínica. Nutrição.

ABSTRACT: The use of Large Language Models (LLMs) for dietary prescription raises clinical safety concerns. This study compares the efficacy of ChatGPT, Gemini, and DeepSeek in generating hypocaloric plans for overweight Brazilian women. Using one-shot and chain-of-thought prompts, 150 meal plans were generated and nutritionally analyzed via TBCA. Results show energy variability: ChatGPT presented superior precision (<0.5% deviation), while the other models overestimated calories. Qualitatively, all models failed: saturated fat exceeded limits by over 200% (reaching 291% in DeepSeek), and systematic micronutrient inadequacies were observed, with iron reaching only 55% to 61% of the recommendations, alongside a calcium deficit. It is concluded that the qualitative inaccuracy of LLMs prevents their autonomous clinical use, requiring professional supervision.

Keywords: Large Language Models. Dietary Prescription. Chain-of-Thought. Artificial Intelligence. Clinical Safety. Nutrition.

[1] Instituto Federal de Educação Ciência e Tecnologia do Ceará (IFCE) – Fortaleza –CE – Brasil.
[2] Universidade Estadual do Ceará (UECE) – Fortaleza–CE – Brasil.
[3] Universidade Estadual do Ceará (UECE) – Fortaleza–CE – Brasil.
[4] Universidade Estadual do Ceará (UECE) – Fortaleza–CE – Brasil.

RESUMEN: El uso de Grandes Modelos de Lenguaje (LLMs) para la prescripción dietética plantea problemas de seguridad clínica. Este estudio compara la eficacia de ChatGPT, Gemini y DeepSeek en la generación de planes hipocalóricos para mujeres brasileñas con sobrepeso. Utilizando prompts one-shot y chain-of-thought, se generaron 150 planes de alimentación y se analizaron nutricionalmente a través de la TBCA. Los resultados muestran variabilidad energética: ChatGPT presentó mayor precisión (desviación <0,5%), mientras que los otros modelos sobreestimaron las calorías. Cualitativamente, todos los modelos fallaron: la grasa saturada excedió los límites en más de un 200% (llegando al 291% en DeepSeek), y se observaron inadecuaciones sistemáticas de micronutrientes, con el hierro alcanzando solo entre el 55% y el 61% de las recomendaciones, junto con un déficit de calcio. Se concluye que la imprecisión cualitativa de los LLMs impide su uso clínico autónomo, requiriendo supervisión profesional.

Palabras clave: Grandes Modelos de Lenguaje. Prescripción Dietética. Chain-of-Thought. Inteligencia Artificial. Seguridad Clínica. Nutrición.

1. INTRODUÇÃO

As doenças crônicas não transmissíveis (DCNTs) e a crescente busca online por informações de emagrecimento popularizaram o uso de Inteligências Artificiais generativas, como ChatGPT e Gemini, para a obtenção de planos alimentares personalizados de forma rápida e direta (PONZO et al., 2024). Contudo, o uso não supervisionado dessas ferramentas na nutrição levanta sérias preocupações entre profissionais de saúde quanto à qualidade e segurança clínica dessas recomendações (COELHO et al., 2024).

A literatura aponta que, embora os Modelos de Linguagem de Grande Escala (LLMs) consigam estruturar cardápios variados baseados em princípios nutricionais básicos, eles falham frequentemente no cumprimento de metas calóricas e no balanceamento de macro e micronutrientes. Além disso, a negligência em relação a fatores individuais, como restrições clínicas e contextos socioeconômicos e culturais, representa uma limitação crítica (BELKHOURIBCHIA; PEN, 2025; KOPITAR et al., 2025; GUO et al., 2025).

No contexto nacional, propostas de sistemas inteligentes para apoio a politicas alimentares culturalmente adaptadas reforcam a necessidade de combinar modelagem computacional com dados de consumo reais da populacao brasileira (SILVA et al., 2025a, SILVA et al., 2025b).

Considerando as limitações relatadas na literatura, este artigo analisa o desempenho dos modelos ChatGPT, Gemini e DeepSeek na elaboração de prescrições dietéticas hipocalóricas. O estudo avalia a precisão no cumprimento de metas nutricionais e a adequação cultural das recomendações para a população brasileira.

2. REVISÃO DE LITERATURA

O uso de Modelos de Linguagem de Grande Escala, como ChatGPT, para a obtenção de informações de saúde tem se expandido (BELKHOURIBCHIA; PEN, 2025). No campo da nutrição, esses modelos têm sido objeto de crescente investigação científica devido ao seu potencial para gerar planos alimentares personalizados, o que poderia ampliar o acesso a recomendações dietéticas (GUO et al., 2025).

2.1 CAPACIDADES E DESEMPENHO GERAL DOS LLMS EM NUTRIÇÃO

Estudos demonstram que os LLMs são capazes de gerar planos alimentares com diversidade e complexidade, muitas vezes alinhados a princípios básicos de uma dieta saudável. Em diversas avaliações, os planos gerados por IAs, especialmente por versões mais avançadas como o ChatGPT-4, alcançaram altas pontuações em índices de qualidade da dieta, como o DQI-I (KIM et al., 2003), principalmente nos quesitos de variedade e adequação de grupos alimentares (DERGAA et al., 2024; KAYA KACAR et al., 2025).

A qualidade das respostas tem se mostrado satisfatória a ponto de, em alguns cenários, os planos gerados serem indistinguíveis daqueles elaborados por nutricionistas humanos. Estudos comparativos revelam que, para perguntas nutricionais comuns, o ChatGPT pode superar o desempenho de nutricionistas em critérios de correção científica, compreensibilidade e acionabilidade (GUO et al., 2025; PONZO et al., 2024). A performance varia entre os diferentes modelos, enquanto o ChatGPT-4 demonstra maior precisão calórica, o DeepSeek se destaca em seguir metas de proteínas mais restritivas e o ChatGPT-4o tende a gerar dietas com maior potencial anti-inflamatório (KAYA KACAR et al., 2025; YOU et al., 2025). Esses achados sugerem potencial para o uso dessas ferramentas como recurso auxiliar na prática clínica (BELKHOURIBCHIA; PEN, 2025; YOU et al., 2025).

2.2 Limitações Críticas e Riscos à Saúde

Uma limitação recorrente entre os modelos é a dificuldade em atingir com precisão as metas calóricas solicitadas (KAYA KACAR et al., 2025).

Apesar do potencial promissor, a literatura aponta de forma consistente para limitações críticas que representam riscos à saúde do usuário. Dentre elas, a imprecisão nutricional. Estudos mostraram que os planos gerados podem apresentar desvios significativos, tanto para mais quanto para menos, em relação ao valor energético alvo. O estudo (KAYA KACAR et

al., 2025) demonstrou que, enquanto o ChatGPT-4 apresentou maior precisão, 50% dos planos do Gemini desviaram em mais de 20% da meta calórica.

A distribuição de carboidratos, proteínas e gorduras é frequentemente desequilibrada, sendo este o componente com a pior avaliação nos índices de qualidade da dieta. Adicionalmente, são comuns as deficiências de micronutrientes essenciais como ferro, cálcio, folato, vitamina D e B12, o que pode levar a problemas de saúde em longo prazo (GUO et al., 2025; NDUKA et al., 2025).

Foram documentados erros graves em relação à segurança do paciente, como a inclusão de alimentos alergênicos (por exemplo, leite de amêndoas) em dietas destinadas a indivíduos com alergia a nozes. Em cenários clínicos complexos, com múltiplas comorbidades, a capacidade dos LLMs de integrar todas as variáveis e restrições diminui drasticamente, resultando em recomendações contraditórias ou inapropriadas (BELKHOURIBCHIA; PEN, 2025; GUO et al., 2025; DERGAA et al., 2024).

Outro ponto fraco identificado nos LLMs é a sua incapacidade de gerenciar interações droga-nutriente, além da falha em estabelecer metas realistas de perda de peso. A interação droga-nutriente ocorre quando um alimento, nutriente ou suplemento afeta a ação de um medicamento, ou, inversamente, quando um medicamento interfere na absorção, metabolismo ou excreção de um nutriente pelo corpo. O ChatGPT, por exemplo, falhou em alertar sobre metas de emagrecimento excessivamente rápidas e, de forma crítica, não considerou potenciais interações medicamentosas que poderiam ser perigosas para o usuário (GUO et al., 2025).

2.3 A Ausência de Cuidado Humanizado e Adequação Cultural

Outra limitação apontada dos LLMs é a dificuldade em replicar o raciocínio clínico e o cuidado humanizado (COELHO et al., 2024). Contudo, estudos recentes demonstram que a aplicação de técnicas de engenharia de prompts pode mitigar parcialmente essa lacuna (SILVA et al., 2024).

Essa lacuna se manifesta na falta de adequação cultural e socioeconômica das recomendações. Os planos gerados frequentemente prescrevem alimentos de alto custo (como salmão, quinoa e abacate) sem oferecer alternativas acessíveis, tornando-os inviáveis para a maioria da população. Além disso, a não inclusão de alimentos básicos e culturalmente relevantes, como o feijão na dieta brasileira, evidencia um viés que compromete a adesão e a efetividade das recomendações (DERGAA et al., 2024; BELKHOURIBCHIA; PEN, 2025).

Em suma, a literatura científica atual posiciona os LLMs como ferramentas de grande potencial, mas que ainda não são suficientemente seguras ou eficazes para serem utilizadas de forma autônoma para a prescrição dietética. As falhas em acurácia, segurança e personalização contextual reforçam a necessidade de supervisão por um profissional de nutrição qualificado.

3. METODOLOGIA

Este estudo foi definido como uma pesquisa quantitativa, descritiva e comparativa, com o objetivo de avaliar e comparar a qualidade de planos alimentares para o contexto brasileiro gerados por três Modelos de Linguagem de Grande Escala: ChatGPT (versão GPT-4), Gemini (versão Gemini-3) e DeepSeek (versão DeepSeek-V3.2).

O modelo ChatGPT foi incluído por ser extensivamente avaliado na literatura científica para geração e análise de recomendações dietéticas, servindo como linha de base para a comparação (GUO et al., 2025; COELHO et al., 2024). O modelo Gemini representa uma arquitetura concorrente que tem sido objeto de estudos recentes (KAYA KACAR et al., 2025).

A inclusão do DeepSeek justifica-se por seu desempenho de destaque em trabalhos recentes, onde evidenciam bom desempenho em estruturar decisões em saúde na atenção primária e apontam superioridade no cumprimento de metas nutricionais restritivas (YOU et al., 2025; FRANCA et al., 2025). A avaliação conjunta destes três modelos permite mapear, com alta fidelidade, as atuais capacidades e limitações tecnológicas na automação do planejamento dietético.

A escolha desses três modelos permite uma análise das capacidades e limitações das diferentes tecnologias de IA disponíveis ao público.

3.1 Desenvolvimento e Administração do Prompt

Para realização dos experimentos, optou-se por utilizar um único prompt para geração de cada plano alimentar, aplicando as técnicas textbfone-shot e textbfchain-of-thought (CoT) estruturado.

Essas técnicas foram selecionadas por permitirem uma avaliação padronizada do desempenho de cada LLM, simulando um usuário especialista que guia o modelo por meio de um raciocínio lógico. A abordagem é classificada como one-shot por fornecer ao modelo um exemplo de dieta pronta e utiliza chain-of-thought, que consiste em instruir explicitamente o

modelo a seguir uma cadeia de passos predefinida pelo usuário antes de gerar o plano alimentar final.

O prompt utilizou as seguintes etapas de raciocínio:

> **1.** Cálculo das Necessidades Energéticas: calcular o Gasto Energético Basal (GEB) e o Gasto Energético Total (GET) para o indivíduo de referência, utilizando a equação de predição de Mifflin-St Jeor.
> **2.** Definição da Meta Calórica: estabelecer a meta calórica da dieta aplicando déficit energético de 500 kcal para o objetivo de emagrecimento.
> **3.** Distribuição de Macronutrientes: distribuir o valor calórico total em percentuais de carboidratos, proteínas e lipídios, conforme as recomendações para a população brasileira.
> **4.** Tradução para Alimentos: converter as metas nutricionais em combinações de alimentos com massas em gramas, priorizando itens comuns na cultura alimentar brasileira, com diversidade entre grupos alimentares e com custo acessível.
> **5.** Montagem do Plano Alimentar: organizar os alimentos selecionados em um plano de 5 dias, com 6 refeições diárias (café da manhã, lanche da manhã, almoço, lanche da tarde, jantar e ceia).
> **6.** Checagem do Plano Alimentar: recalcular os nutrientes diários obtidos em cada dia e, caso a distribuição não seja adequada, retornar ao passo 3.
> **7.** Apresentação do Plano Alimentar: apresentar o plano alimentar de 5 dias, com 6 refeições diárias, informando as quantidades dos alimentos em gramas para viabilizar o cálculo nutricional de cada refeição.

3.2 Escolha do Perfil

Para garantir reprodutibilidade e relevância epidemiológica, estabeleceu-se um perfil representativo de usuários que buscam dietas online: mulher, 30 anos, 71 kg, 1,62 m, com sobrepeso (IMC 27 kg/m²), atividade física leve e objetivo de emagrecimento saudável. A escolha do sexo feminino justifica-se pela maior prevalência de excesso de peso neste grupo (STOPA et al., 2020), que também lidera as buscas online por dietas (DOS SANTOS PORTO et al., 2019). A idade de 30 anos representa adultos jovens digitalmente engajados, principal faixa etária na procura por soluções nutricionais na internet (MOURA et al., 2022). Por fim, a condição de sobrepeso e o foco em emagrecimento refletem a realidade de mais de 61% dos adultos no Brasil (SAÚDE, 2024), que constitui um desafio de saúde pública nacional e a principal demanda por orientação dietética (DIAS et al., 2025).

3.3 Análise da Composição Nutricional

Após a geração dos planos alimentares, a composição nutricional de cada um foi sistematicamente analisada. A composição de energia (kcal), macronutrientes e micronutrientes selecionados foi determinada utilizando como referência a Tabela Brasileira de Composição de Alimentos (TBCA) (Universidade de São Paulo, 2025). Os valores obtidos foram comparados com as metas de referência.

4. Experimentos e Análise de Resultados

Esta seção apresenta os resultados da avaliação nutricional dos planos gerados pelos três modelos e organiza a discussão de forma comparativa, com foco em precisão energética, adequação de macro e micronutrientes, originalidade e aderência ao contexto alimentar brasileiro.

4.1 Geração dos Planos Alimentares e Processamento de Dados

O experimento consistiu em 10 requisições de planos alimentares de cinco dias para cada um dos três modelos, resultando em 50 dias de dietas por modelo e um total de 150 dias analisados. A *Tabela 1* apresenta uma amostra de um dia de plano alimentar gerado por cada ferramenta.

Refeição	Gemini	ChatGPT	DeepSeek
Café Manhã	Leite Integral (100g), Aveia (30g), Maçã (130g)	Leite Desnatado (200g), Cuscuz (80g), Ovos (100g)	Cuscuz Milho (120g), Ovo Cozido (50g), Melancia (150g)
Lanche Manhã	Banana (100g), Amendoim s/ sal (10g)	Mac¸ã (130g), Castanha-do-Pará (10g)	Castanha-do-Pará (20g), Pera (110g)
Almoço	Arroz (100g), Feijão (80g), Frango Grelhado (110g), Salada (150g), Azeite (5g)	Arroz Integral (100g), Feijão (80g), Frango Desfiado (120g), Abo´bora (100g), Azeite (5g)	Arroz Integral (95g), Lentilha (75g), Frango Ensopado (115g), Berinjela Assada (140g), Azeite (5g)
Lanche Tarde	Pão Francês (50g), Queijo Minas (40g)	Iogurte Desnatado (170g), Aveia (15g)	Coalhada Seca (130g), Ameixa Seca (30g)
Jantar	Batata Doce (150g), Ovos (80g), Salada Bro´colis/Cenoura (150g)	Macarrão Integral (120g), Carne Moída (100g), Molho Tomate (50g)	Omelete Espinafre (120g), Batata Doce (140g), Salada Rúcula (90g)
Ceia	Chá s/ açucar (200g)	Leite Desnatado (150g)	Leite Desnatado (180g)

Tabela 1: *Exemplos de planos alimentares diários recomendados pelos modelos Gemini, ChatGPT e DeepSeek.*

Para a análise da composição nutricional, os alimentos sugeridos pelos modelos foram mapeados para os seus códigos correspondentes na TBCA.

Para garantir a padronização dos dados no mapeamento, adotaram-se duas regras para resolver ambiguidades:

Quando o método de preparo não foi especificado pela ferramenta, selecionou-se na base de dados a opção com o menor valor de *Gordura em Adição*;

Em saladas compostas por múltiplos vegetais sem a especificação da massa individual, o peso total da porção foi dividido igualmente entre os ingredientes citados (por exemplo, para 100g de salada de tomate com cenoura, considerou-se 50g de cada).

4.2 Parâmetros de Referência

A avaliação da adequação nutricional foi baseada no perfil estabelecido na *Subseção 3.2*. A Taxa Metabólica Basal (TMB) calculada pela equação de Mifflin-St Jeor resultou em 1411,5 kcal. Com a aplicação do Fator de Atividade (FA) de 1,375, o Gasto Energético Total (GET) atingiu 1940,8 kcal. O Valor Energético Total (VET) utilizado como meta calórica diária, após a aplicação do déficit de 500 kcal para emagrecimento, foi de 1440,8 kcal. A *Tabela 2* resume as metas de ingestão utilizadas para a comparação com os valores obtidos nas dietas sugeridas pelos modelos.

Nutriente	Alvo	Gemini	ChatGPT	DeepSeek
Energia (kcal)	1440.8	1523.8±129.3	1446.0±147.0	1487.9±165.8
Proteínas (g)	106.5	98.2±11.5	100.2±13.6	99.1±11.3
Carboidratos (g)	163.7	184.7±21.3	181.9±24.3	165.1±20.6
Lipídios (g)	40.0	49.0±7.7	41.3±8.4	53.1±15.8
Gordura Sat. (g)	<16.0	16.5±2.7	12.7±3.5	18.5±6.7
Fibras (g)	>25.0	23.6±4.0	27.2±3.0	24.8±4.2
Cálcio (mg)	1000	858.2±186.1	966.4±205.4	972.2±310.2
Zinco (mg)	8.0	9.9±2.6	11.2±3.1	11.3±2.7
Selênio (µg)	55.0	55.3±27.5	55.9±21.1	62.9±32.9
Ferro (mg)	18.0	11.0±1.9	10.0±2.0	10.0±2.8
Magnésio (mg)	310	266.4±35.5	335.1±37.7	297.9±56.6
Sódio (mg)	<2000	2271.3±392.0	2018.8±1323.5	1869.2±735.9
Colesterol (mg)	<300	470.6±179.3	395.8±183.7	397.0±157.4
Vitamina C (mg)	75.0	91.7±47.4	139.8±64.0	144.0±62.4
Vit. B12 (µg)	2.4	5.0±2.4	4.5±1.6	8.0±7.0
Folato (µg)	400	355.1±71.7	377.1±78.1	340.9±80.5

Tabela 2: *Comparativo entre a ingestão nutricional recomendada e o perfil médio simulado pelos modelos.*

4.3 Consumo Energético e Macronutrientes

Observa-se, na *Figura 1*, que o modelo ChatGPT apresentou a maior precisão em relação à meta energética, com média de 1446,04 kcal (± 146,96), o que representa variação inferior a 0,5% do alvo estabelecido. Os modelos Gemini e DeepSeek apresentaram médias superiores, com 1523,80 kcal e 1487,93 kcal, respectivamente. Na figura, a sigla GBT corresponde ao ChatGPT.

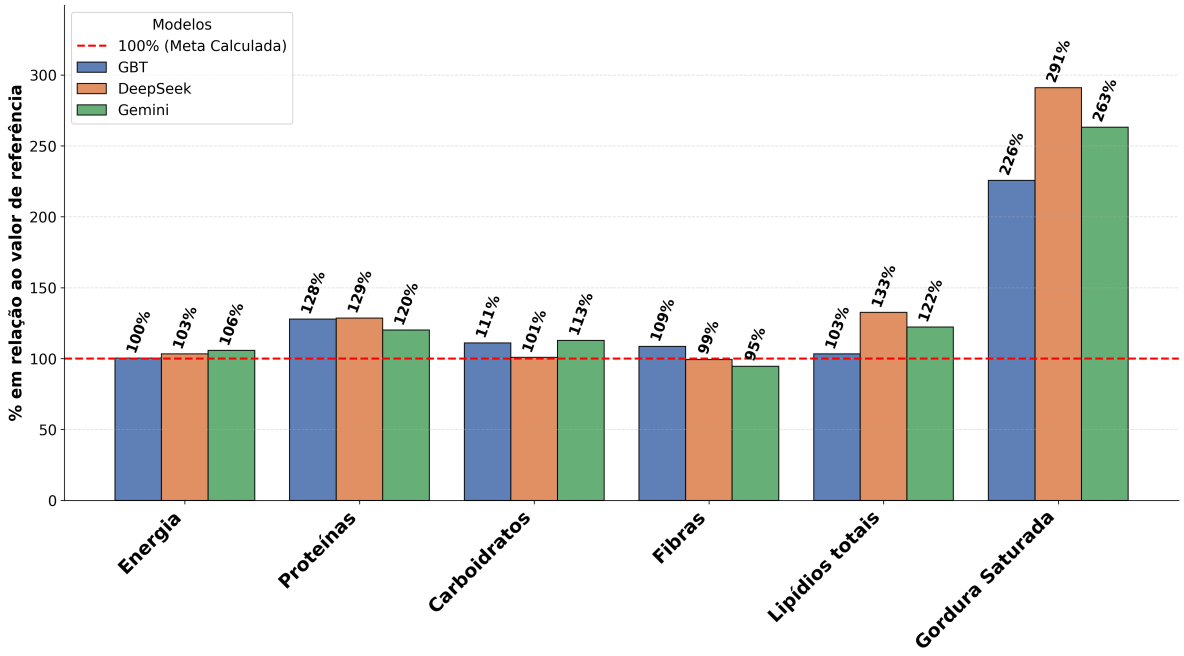

Figura 1: Adequação do consumo de macronutrientes e fibras.

Quanto à distribuição de macronutrientes, todos os modelos apresentaram um consumo proteico próximo, porém ligeiramente inferior à meta de 106,5g, calculada para preservação de massa magra, variando entre 98,23g (Gemini) e 100,19g (ChatGPT). O modelo DeepSeek destacou-se pelo maior teor de lipídios totais (53,06g ± 15,77), ultrapassando a referência de 40g em 33%, conforme ilustrado na *Figura 1*.

Um dado relevante refere-se à quantidade de gordura saturada nas dietas recomendadas. Conforme demonstrado na *Figura 1*, todos os modelos ficaram acima da referência relativa adotada na análise. O modelo DeepSeek apresentou a maior discrepância, atingindo 291% do valor de referência, seguido por Gemini (263%) e ChatGPT (226%).

4.4 Análise de Micronutrientes

A análise dos micronutrientes, detalhada na *Tabela 2* e visualizada na *Figura 2*, revela padrões de inadequação consistentes entre os três modelos. No que diz respeito ao ferro, nenhum modelo atingiu a recomendação dietética de 18mg para mulheres em idade fértil. As médias variaram entre 9,97mg e 11,02mg, o que representa uma adequação de apenas 55% a 61% da necessidade diária. Este achado reafirma limitações identificadas em literatura prévia sobre a dificuldade de LLMs em garantir a densidade nutricional de ferro (COELHO et al., 2024). De maneira semelhante, a ingestão de cálcio permaneceu abaixo da recomendação de 1000mg

em todos os cenários, com o modelo Gemini apresentando a menor média (858,21mg), correspondendo a 86% da adequação.

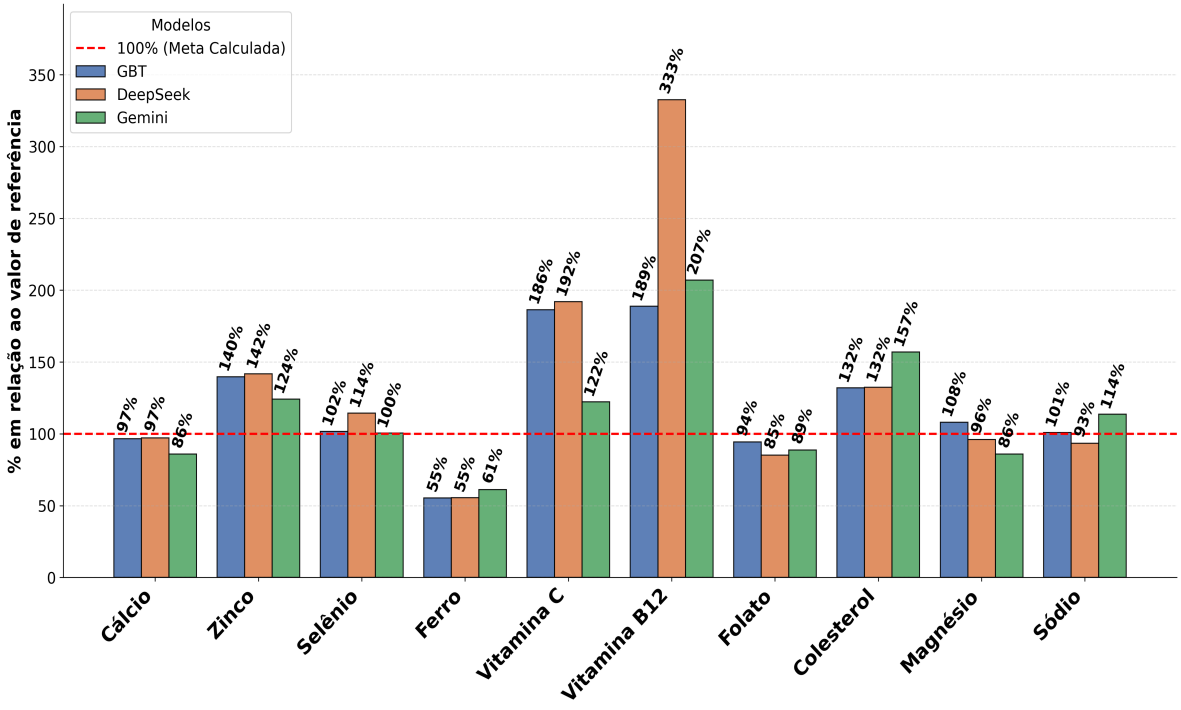

Figura 2: Adequação do consumo de micronutrientes e colesterol.

Em contraste com as deficiências observadas, a Vitamina B12 obteve níveis altos de consumo. O modelo DeepSeek, particularmente, atingiu 333% do valor de referência (7,98 mcg), o que sugere uma priorização acentuada de fontes de proteína animal na composição dos cardápios. Em relação aos demais parâmetros avaliados, os níveis de sódio mantiveram-se próximos aos limites de referência em todos os modelos. Por fim, a quantidade de fibras oscilou em torno da meta de 25g, com o modelo ChatGPT apresentando o melhor desempenho neste quesito, alcançando uma média de 27,15g.

4.5 Originalidade e Adequação Cultural

A geração de texto por modelos de linguagem tem natureza estocástica, o que significa que o mesmo comando pode gerar respostas diferentes a cada uso (GUO et al., 2025; DERGAA et al., 2024). No estudo, o comando instruiu que o modelo não deveria repetir cardápios fornecidos anteriormente. Ao avaliar a originalidade dos cardápios gerados, o ChatGPT apresentou o melhor desempenho, entregando 100% de dias diferentes. O Gemini repetiu refeições em 7 dias (86% de originalidade) e o DeepSeek repetiu em 14 dias (72% de

originalidade). Esse resultado mostra que alguns modelos têm maior dificuldade em manter variedade de refeições quando precisam seguir restrições nutricionais.

Além de variar os pratos, uma dieta precisa respeitar a cultura de quem vai consumi-la (BELKHOURIBCHIA; PEN, 2025). Estudos reforçam a importância do feijão na alimentação diária do brasileiro como fator de proteção à saúde (SAÚDE, 2024). No estudo, o comando instruiu que a dieta era direcionada para brasileiros, mas sem exigir explicitamente a inclusão de arroz e feijão. Como resultado, o ChatGPT incluiu essa combinação no almoço em 100% dos dias. O Gemini fez isso em 72% das vezes e o DeepSeek em 58%.

Esses achados indicam um avanço de alguns modelos em relação a estudos anteriores, que criticavam as LLMs por esquecerem as leguminosas e sugerirem dietas fora da realidade local (COELHO et al., 2024).

4.6 Distribuição Energética das Refeições

A *Figura 3* ilustra a distribuição percentual média do Valor Energético Total (VET) entre as seis refeições diárias. Todos os modelos adotaram padrão em que a maior parte da energia se concentra nas refeições principais. O almoço recebeu o maior aporte calórico em todos os cenários, variando de 28,4% (Gemini) a 29,7% (GBT, correspondente ao ChatGPT), seguido por jantar (21,0% a 25,7%) e café da manhã (18,1% a 21,9%). Lanches intermediários e ceia receberam proporções menores.

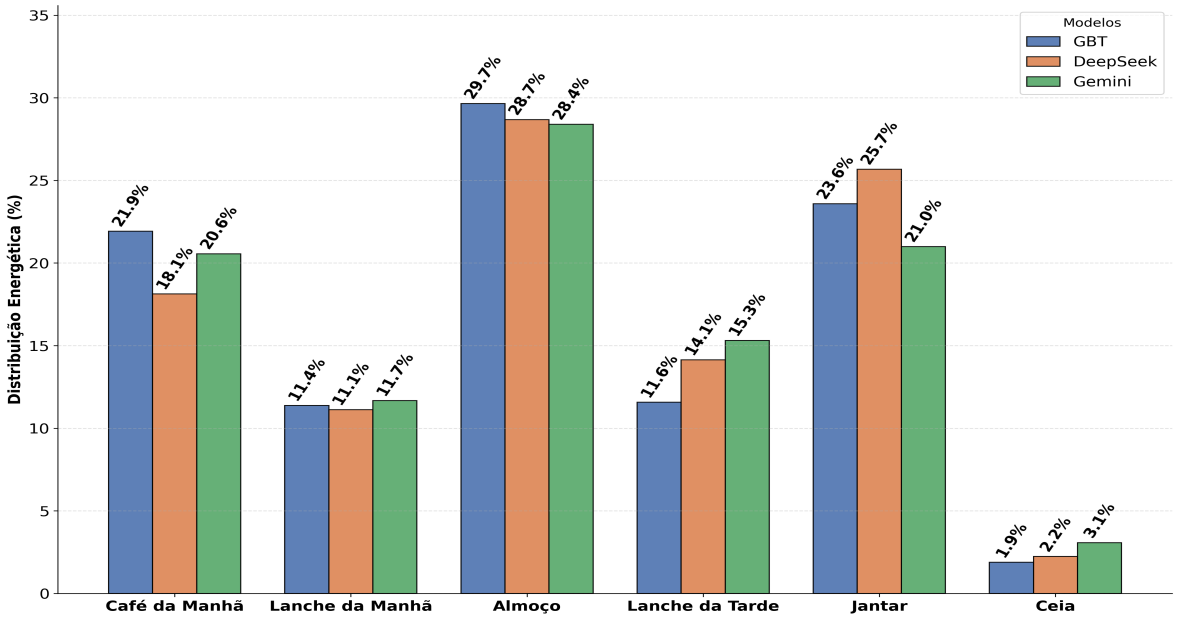

Figura 3: Distribuição energética das refeições

Um ponto positivo nas dietas geradas foi a baixa alocação de calorias para a ceia, que variou de 1,9% (GBT, correspondente ao ChatGPT) a 3,1% (Gemini) da energia diária, respeitando o princípio de evitar refeições pesadas no período noturno tardio.

Esses resultados mostram que, embora os modelos apresentem falhas qualitativas na escolha de algumas fontes de alimentos, eles conseguem replicar as regras matemáticas e estruturais de uma prescrição (GUO et al., 2025; PONZO et al., 2024). A capacidade de distribuir a energia de forma lógica ao longo do dia confirma que os modelos compreendem o formato geral esperado para um plano alimentar.

4.7 Síntese Comparativa entre os Modelos

Em termos de precisão energética, o ChatGPT apresentou o melhor desempenho médio, seguido de DeepSeek e Gemini. Na adequação de macronutrientes, todos os modelos mostraram desvios relevantes: DeepSeek apresentou os maiores valores relativos para lipídios totais e gordura saturada, enquanto o ChatGPT manteve melhor proximidade da meta energética.

Na análise de micronutrientes, os três modelos repetiram inadequações para ferro e cálcio, com diferenças de magnitude entre eles. DeepSeek apresentou maior excesso relativo de vitamina B12 e colesterol. Gemini apresentou os menores valores relativos de cálcio e magnésio.

Na dimensão de aderência cultural e originalidade, o ChatGPT obteve melhor resultado, seguido por Gemini e DeepSeek. Em conjunto, os resultados indicam que nenhum modelo é superior em todos os critérios: ChatGPT se destaca em precisão energética e consistência de estrutura, DeepSeek apresenta maior variabilidade e maiores excessos em nutrientes críticos, e Gemini fica em posição intermediária com melhor desempenho em alguns indicadores de micronutrientes.

5. Ameaça a validade

Os resultados deste estudo devem ser interpretados considerando limitações do escopo e à metodologia. A restrição a um único perfil demográfico (mulher, 30 anos, sobrepeso) limita a generalização dos achados para populações com necessidades metabólicas distintas ou comorbidades complexas.

Além disso, a análise é estritamente teórica, não avaliando fatores como palatabilidade, custo real da cesta de compras ou a eficácia clínica em longo prazo.

Do ponto de vista analítico, o estudo focou em métricas quantitativas isoladas, sem a aplicação de índices compostos de qualidade, como o Diet Quality Index-International (DQI-I) (KIM et al., 2003). A literatura recente aponta que, mesmo quando LLMs atingem metas energéticas, eles frequentemente falham nos critérios de equilíbrio e moderação avaliados por esses índices (KAYA KACAR et al., 2025).

A natureza estocástica dos modelos impede a reprodutibilidade exata dos cardápios, porém disponibilizamos o prompt one-shot utilizado, garantindo resposta com formatação uniforme para comparações futuras.

Por fim, a necessidade de decomposição manual de pratos introduz um grau de subjetividade para a classificação dos alimentos, porém utilizamos critérios descritos na *Subseção 4.1* para reduzir a subjetividade da classificação, além da disponibilização da classificação de todos os alimentos avaliados.

6. Disponibilização dos dados

Para fins de reprodutibilidade, o conjunto de dados completo da pesquisa (o prompt utilizado, os planos gerados, a classificação dos alimentos e relatórios estatísticos) está acessível em: https://github.com/thalyson004/llm-evaluation-diet

7. CONCLUSÕES

Este estudo avaliou a eficácia de três Modelos de Linguagem de Grande Escala (LLMs), ChatGPT, Gemini e DeepSeek, na elaboração de planos alimentares personalizados para o contexto brasileiro, utilizando a técnica de chain-of-thought estruturada. A análise dos dados revela que, embora os modelos demonstrem competência na estruturação lógica das refeições e na aproximação das metas energéticas globais, a qualidade nutricional detalhada dos planos gerados apresenta deficiências críticas que comprometem sua utilização clínica autônoma.

Quanto à precisão calórica, o modelo ChatGPT destacou-se com a menor variação em relação à meta estabelecida (<0,5%), corroborando achados recentes de (KAYA KACAR et al., 2025), que identificaram a superioridade deste modelo na adesão a alvos energéticos em comparação a concorrentes. Em contrapartida, os modelos Gemini e DeepSeek tenderam à superestimação calórica, o que, no contexto do tratamento da obesidade poderia prejudicar a eficácia de intervenções de emagrecimento.

Entretanto, o cumprimento da meta calórica não se traduziu em equilíbrio qualitativo. Uma falha sistemática observada em todos os modelos foi o excesso relativo de gordura saturada, com o modelo DeepSeek atingindo 291% da referência no comparativo. Esse padrão sugere que os algoritmos, mesmo quando guiados por prompts complexos, priorizam a contabilidade de macronutrientes totais em detrimento da seleção qualitativa das fontes lipídicas, elevando potencialmente o risco cardiovascular dos usuários.

A análise de micronutrientes reforça limitações documentadas na literatura. A inadequação consistente de ferro (55% a 61% da recomendação) e cálcio nos cardápios gerados alinha-se aos resultados de (COELHO et al., 2024), que alertaram para a incapacidade do ChatGPT em garantir a densidade nutricional necessária para alguns grupos, como mulheres em idade fértil. Divergindo parcialmente da literatura que aponta deficiências generalizadas, observou-se no modelo DeepSeek um aporte excessivo de Vitamina B12 (333% da referência), indicando um possível viés do modelo para a inclusão desproporcional de proteínas de origem animal, o que pode impactar a viabilidade econômica da dieta para a população brasileira média.

Conclui-se que, no estágio tecnológico atual, os LLMs avaliados operam como ferramentas de suporte para ideação e estruturação de cardápios, mas carecem da precisão clínica e do discernimento contextual necessários para a prescrição dietética segura. A supervisão humana por nutricionistas permanece indispensável para corrigir desequilíbrios de micronutrientes e moderar o perfil lipídico das sugestões.

Como trabalhos futuros, propõe-se o desenvolvimento de arquitetura híbrida que integre a capacidade de processamento de linguagem natural dos LLMs com algoritmos de otimização matemática, mecanismos de recuperação de conhecimento por retrieval-augmented generation (RAG) e coordenação por agentes especializados para ampliar segurança, padronização e rastreabilidade dos resultados. Recomenda-se também estender o estudo para outros grupos de pacientes, além de mulheres adultas com sobrepeso, incluindo diferentes faixas etárias, sexo biológico, perfis clínicos e comorbidades. Essa agenda pode consolidar diretrizes mais abrangentes para aplicação prática em nutrição clínica e saúde pública.

REFERÊNCIAS

BELKHOURIBCHIA, J. and Pen, J. J. (2025). Large language models in clinical nutrition: an overview of its applications, capabilities, limitations, and potential future prospects. Frontiers in Nutrition, 12:1635682.

COELHO, P. K., Santos, A. M. S., das Neves Santos, A. C., and Machado, V. C. (2024). Inteligência artificial e nutrição: Um estudo da composição nutricional de cardápio de emagrecimento gerado por chatgpt. Revista PET Brasil, 3(01):50–62.

DERGAA, I., Saad, H. B., Ghouili, H., Glenn, J. M., El Omri, A., Slim, I., Hasni, Y., Taheri, M., Aissa, M. B., Guelmami, N., et al. (2024). Evaluating the applicability and appropriateness of chatgpt as a source for tailored nutrition advice: A multi-scenario study. New Asian Journal of Medicine, 2(1):1–16.

DIAS, A. C. S., Santana, J. D., Colla, M. A., Borges, M. E. R., Tissiani, G. F., Meneses, I. S. B., dos Santos Sousa, A. C., de Morais Neto, J. M., Rodrigues, K. M., de Lima Filho, A. V., et al. (2025). O cenário passado e atual da obesidade e do sobrepeso no brasil. Studies in Health Sciences, 6(1):e13936–e13936.

FRANÇA, P. A., Sá, R. V., Alves-Costa, S., de AF Viola, P. C., Costa, S. S., de Souza, B. F., de Almeida, J. D., Diniz, J. O., and Ribeiro, C. C. (2025). Maria-deepseek: Uma proposta de assistente por modelo amplo de linguagem para agentes comunitários de saúde. Simpósio Brasileiro de Computação Aplicada a Saúde (SBCAS), pages 305–316. SBC.

GUO, P., Liu, G., Xiang, X., and An, R. (2025). From ai to the table: A systematic review of chatgpt's potential and performance in meal planning and dietary recommendations. Dietetics, 4(1):7.

KAYA Kac ̧ar, H., Kaçar, O . F., and Avery, A. (2025). Diet quality and caloric accuracy in ai-generated diet plans: A comparative study across chatbots. Nutrients, 17(2):206.

KIM, S., Haines, P. S., Siega-Riz, A. M., and Popkin, B. M. (2003). The diet quality index-international (dqi-i) provides an effective tool for cross-national comparison of diet quality as illustrated by china and the united states. The Journal of nutrition, 133(11):3476–3484.

KOPITAR, L., Bedrac, L., Strath, L. J., Bian, J., and Stiglic, G. (2025). Improving personalized meal planning with large language models: Identifying and decomposing compound ingredients. Nutrients, 17(9):1492.

MINISTÉRIO da Saúde (2024). Vigitel brasil 2006-2023: vigilância de fatores de risco e proteção para doenças crônicas por inquérito telefônico. Technical report, Secretaria de Vigilância em Saúde e Ambiente, Brasília, DF. Disponível em: https://www.gov.br/saude/pt-br/centrais-de-conteudo/publicacoes/svsa/vigitel/vigitel-brasil-2023-vigilancia-de-fatores-de-risco-e-protecao-para-doencas-cronicas-por-inquerito-telefonico/view. Acesso em: 20 abril. 2026.

MOURA, C. A. S., dos Santos, Y. M., da Silva Neto, J. G., Cavalcante, S. K. C. C., de Sousa, E. F. G., Honório Filho, S. M., Alves, M. E. P., de Freitas Pereira, T. E., de Freitas Pereira, T. C., and de Brito, A. N. M. (2022). Os perigos das dietas milagrosas sem acompanhamento do profissional nutricionista. RECIMA21-Revista Científica Multidisciplinar-ISSN 2675-6218, 3(2):e321106–e321106.

NDUKA, T. C., Ndakotsu, A., Nriagu, V. C., Karikalan, S., Abdulkareem, L., Omede, F. O., and Bob-Manuel, T. (2025). Ai-generated diet and exercise recommendations for cardiovascular health compared to established cardiology society guidelines. Cureus, 17(8):e90968.

PONZO, V., Goitre, I., Favaro, E., Merlo, F. D., Mancino, M. V., Riso, S., and Bo, S. (2024). Is chatgpt an effective tool for providing dietary advice?. Nutrients, 16(4):469.

SANTOS Porto, T. N. R., da Rocha Cardoso, C. L., Baldoino, L. S., de Sousa Martins, V., Alcântara, S. M. L., and Carvalho, D. P. (2019). Prevalência do excesso de peso e fatores de risco para obesidade em adultos. Revista Eletrônica Acervo Saúde, 22(22):e308–e308.

SILVA, V., Furtado, E. S., Oliveira, J., and Furtado, V. (2024). Engenharia de prompts em assistentes conversacionais para promoção de autocuidado baseados em modelos amplos de linguagem. Simpósio Brasileiro de Computação Aplicada a Saúde (SBCAS), pages 377–388. SBC.

SILVA, T. G., de Campos, G. A., Júnior, B. A., and de Paula Barros, A. L. B. (2025a). A bio-inspired ai approach to personalized dietary planning for chronic disease prevention. Congresso Latino-Americano de Software Livre e Tecnologias Abertas (Latinoware), páginas 194–201. SBC.

SILVA, T. G., de Campos, G. A., Júnior, B. A., and de Paula Barros, A. L. B. (2025b). Intelligent systems for public health: A multi-agent system for culturally tailored dietary policy. Congresso Latino-Americano de Software Livre e Tecnologias Abertas (Latinoware), páginas 33–41. SBC.

STOPA, S. R., Szwarcwald, C. L., Oliveira, M. M. d., Gouvea, E. d. C. D. P., Vieira, M. L. F. P., Freitas, M. P. S. d., Sardinha, L. M. V., and Macário, E. M. (2020). Pesquisa nacional de saúde 2019: histórico, métodos e perspectivas. Epidemiologia e Serviços de Saúde, 29:e2020315.

UNIVERSIDADE DE SÃO PAULO (USP); CENTRO DE PESQUISA EM ALIMENTOS (FoRC). Tabela Brasileira de Composição de Alimentos (TBCA). Versão 7.3. São Paulo, 2025. Disponível em: http://www.fcf.usp.br/tbca. Acesso em: 20 abr. 2026.

YOU, Q., Zhou, L., Ma, Y., Guo, J., Wang, Y., Shi, L., Deng, Y., Rao, Z., and Li, X. (2025). Comparison of chatgpt-3.5, chatgpt-4o and deepseek in generating dietary plans for patients with chronic kidney disease: a focus on nutritional accuracy and dietary inflammation. Nutrition, page 112957.