Large Language Models in Nutritional Recognition: A Comprehensive Review of Applications

Qingfeng Tian
School of Computer Science,
Xiangtan University, 411015,
Xiangtan, China
202205566725@smail.xtu.edu.cn

Boyuan Wang*
The Chinese University of Hong Kong, Shenzhen (CUHK-Shenzhen), China
boyuan422@qq.com
**Corresponding author*

Shanquan Chen*
The Chinese University of Hong Kong, Shenzhen (CUHK-Shenzhen), China
chenshanquan@cuhk.edu.cn
**Corresponding author*

2025 10th International Conference on Computer and Communication System (ICCCS) | 979-8-3315-2314-5/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICCCS65393.2025.11069607

***Abstract*—Nutritional recognition has become a critical component in modern healthcare, which is used in personalized dietary management and chronic disease prevention. Traditional approaches, ranging from manual analysis to deep learning methods, face limitations in scalability, real-world adaptability. The advent of large language models (LLMs) shows transformative potential to address these challenges, which leverage multimodal integration and cross-modal attention mechanisms. This paper systematically reviews the applications of LLMs in nutritional recognition. Key innovations such as nutrition-specific tokenization and vision-language alignment demonstrate significant improvements in accuracy and practical utility. However, LLMs face challenges, including dataset dependency, computational inefficiency, and Hallucination. By synthesizing recent advancements, this paper also points out the future directions.**

Keywords—Large Language Models, Nutritional Recognition, Multimodal fusion, Dietary Management, Personalized Nutrition, Health

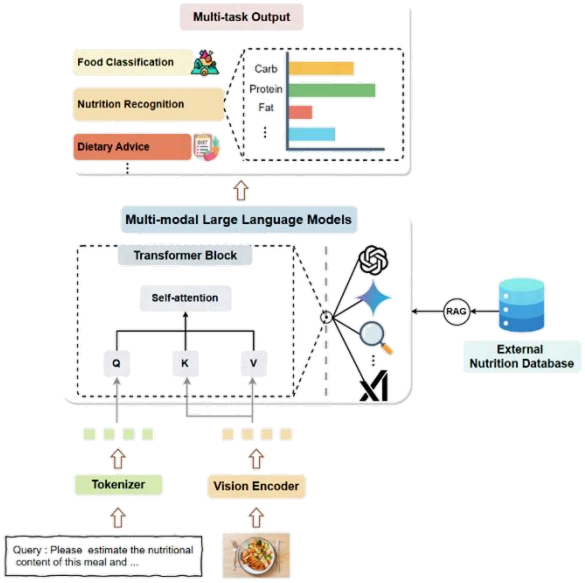

Figure 1. MLLMs paradigm in Nutrition Recognition

I. Introduction

Nutritional analysis plays a pivotal role in public health, chronic disease prevention, and personalized dietary guidance. Chronic diseases like diabetes and cancers are closely linked to modifiable dietary factors and imbalanced nutrient intake would exacerbate inflammatory and metabolic dysregulation[1]. It demonstrates the bidirectional relationship between nutrition and health. Thus, it is crucial to establish an efficient and dynamic nutritional recognition framework.

Early methods predominantly relied on manual analysis by dietitians and chemical assays, which achieved high accuracy but were impractical for large-scale applications. The advent of ML introduced data-driven solutions, where classifiers utilized handcrafted features to categorize food [2]. Although these ML-based systems improved operational efficiency, their performance hinged critically on domain expertise for feature engineering. Recent advances in deep learning, particularly CNNs, revolutionized the field by enabling end-to-end feature extraction from raw visual data[3]. These DL models demonstrated superior capabilities in handling complex food morphologies. Nevertheless, persisting limitations still remain, such as, suboptimal performance in real-world scenarios involving occluded or ambiguous food items.

The emergence of LLMs offers a transformative opportunity to address the limitations of traditional methods. With the advanced NLP capabilities and vast knowledge bases, LLMs can potentially revolutionize dietary analysis through multimodal integration, contextual understanding, and adaptive

learning[4]. In addition, these models also excel in interpreting complex nutritional queries and processing diverse data formats.

This survey aims to provide a comprehensive review of the emerging applications of LLMs in nutritional recognition, presenting their transformative potential in dietary analysis and personalized health management. And the paper is structured as follows: Section II introduces the fundamental principles of LLMs; Section III delves into the specific applications of LLMs in nutritional recognition; Section IV conclusion.

II. Fundamental Principles of LLMs

As a versatile framework for processing complex multimodal data, LLMs leverage core components such as self-attention mechanism, positional encoding, cross-modal attention mechanisms, multi-task learning paradigm and Generative modeling. This section focuses on the foundational principles of LLMs in nutritional recognition, elucidating their theoretical mechanisms and technical pathways in addressing domain-specific challenges.

1. Self-Attention Mechanism for Global Context Modeling

The self-attention mechanism, a core component of the Transformer architecture, enables models to capture global contextual dependencies by dynamically computing relationships between different positions in input sequences. Its fundamental principle involves mapping input sequences into Query (Q), Key (K), and Value (V) matrices, followed by calculating the similarity between queries and keys, normalizing the scores, and aggregating the value matrix with weighted sums. The formula is defined as:

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

where Q, K, and V are linear transformations of the input, and $\sqrt{d_k}$ scales the dot product to mitigate gradient vanishing. In nutritional recognition tasks, self-attention effectively models contextual associations between keywords in food descriptions, such as linking "low-sugar" to "recommended for diabetic patients." Additionally, its global modeling capability supports efficient processing of long-text inputs, ensuring semantic coherence when analyzing multi-paragraph dietary information.

2. Positional Encoding for Temporal Modeling

Positional Encoding improves the lack of sequential information by assigning unique positional embeddings to each position. And this mechanism employs a combination of sine and cosine functions to generate position-dependent embeddings. The formula is as follows:

$$PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{2i/d}}\right), PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{2i/d}}\right)$$

Here, *pos* is the position index, *d* is the embedding dimension, and *i* is the dimension index. By using exponentially decaying frequencies, this design allows the model to capture both absolute positions and relative distances.

3. Cross-Modal Attention for Alignment of Multimodal Information

As a pivotal mechanism in MLLMs, Cross-modal attention, bridges feature spaces across modalities to enable cross-modal interaction. It dynamically aligns queries (Q) from one modality with keys (K) and values (V) from another, formulated as:

$$\text{CrossAttention}(Q_{\text{text}}, K_{\text{image}}, V_{\text{image}}) = \text{softmax}\left(\frac{Q_{\text{text}}K_{\text{image}}^T}{\sqrt{d_k}}\right)V_{\text{image}}$$

Here, Q_{text} represents the text-based query matrix, while K_{image} and V_{image} denote the image-based key-value matrices. It enables the model to align visual features with textual nutrient labels and support joint reasoning for generating dietary advice and health risk assessments.

4. Synergistic Optimization in Multi-Task Learning

Multi-task learning enhances model generalization in complex nutritional recognition scenarios by jointly optimizing loss functions of multiple related tasks. Its core principle dynamically combines classification, regression, and other tasks losses into a unified objective through shared representations and task-specific coefficients (λ), with the following formula:

$$\mathcal{L}_{\text{total}} = \lambda_1\mathcal{L}_{\text{classify}} + \lambda_2\mathcal{L}_{\text{regress}} + \cdots$$

where λ are tunable weights. In nutritional recognition, this framework enables simultaneous predictions of food categories, nutrient values, and other dietary tasks to achieve multi-objective synergy.

5. Generative Modeling for Nutritional Reports

Generative nutritional report modeling leverages conditional generation techniques to dynamically produce nutrient analyses or personalized dietary recommendations based on input data. The objective function employs maximum likelihood estimation (MLE) to optimize model parameters by maximizing the conditional probability of the generated sequence, with the following formula:

$$\mathcal{L}_{\text{gen}} = -\sum_{t=1}^{T}\log P(y_t|y_{<t}, x)$$

Here, x denotes the input data, y_t represents the t-th token in the generated sequence, and $y_{<t}$ refers to previously generated tokens. This framework generates nutritional reports autoregressively, ensuring contextual coherence. In nutritional recognition, it supports personalized dietary advice or inferring nutrient tables from recipe descriptions, providing interpretable decision support for health management.

III. Applications of LLMs in Nutrition Recognition

The application of LLMs in nutritional recognition can be categorized into unimodal and multimodal paradigms based on input modalities. In unimodal processing pipelines, LLMs typically receive textual inputs which undergo semantic parsing and structured information extraction to identify nutritional components, portion sizes, and other dietary parameters through the model's inherent linguistic understanding capabilities.

By contrast, multimodal architectures (**Fig.1**) employ heterogeneous data fusion through coordinated visual-linguistic

TABLE I.
COMPARATIVE ANALYSIS OF ADVANCED MODELS FOR NUTRITION AND FOOD INFORMATION PROCESSING

Method Name	Dataset	LLMs Model	Experimental Result
FoodLMM [5]	Nutrition5k, FoodDialogues, FoodReasonSeg	LLaVA-7B	Total Caloric MAE: 67.2 kcal
Knowledge-Infused LLM-Powered Conversational Health Agent (CHA) [6]	Nutritionix API, 100 diabetes-related meal questions	ChatGPT (3.5 Turbo)	Accuracy: CHA outperformed GPT4
AI Dietician System [7]	USDA FoodData Central	ChatGPT (3.5 Turbo)	Energy: 96.29% within 40% bounds, CV=3.3%
NutriBench [8]	NutriBench, WWEIA and FAO/WHO GIFT	GPT-4o, GPT-4o-mini, Llama3.1 (8B-405B), Gemma2 (9B-27B), Qwen2 (7B-72B), OpenBioLLM-70B	Best Accuracy: GPT-4o (CoT) achieved 66.82% Acc@7.5 (MAE: 8.61g).
The Food Product Nutrition Assistant [9]	Manufacturer-provided Nutrition Facts labels and ingredient lists	GPT-4	Likert scores improved from Level 1 (accuracy: 3.77) to Level 3 (accuracy: 4.40).
CaLoRAify [10]	CalData	MiniGPT-v2	ROUGE-2: 0.0947; BLEU: 0.0218; Aggregate Metrics: 0.4662.
NGQA (Nutritional Graph Question Answering Benchmark) [11]	NGQA benchmark, NHANES, FNDDS	GPT-4o-mini, Llama-3.1-70b-instruct, GPT-3.5-turbo	Text Generation: ToG scored ROUGE-L=0.82, BERT=0.97.
CalorieVoL [12]	Nutrition5k	GPT-4V, LLaVA-v1.5	GPT-4V + CalorieVoL: MAE = 101.7 kcal, correlation coefficient = 0.708.

processing. The framework involves 2 parallel feature extraction streams: visual inputs are encoded through Vision Encoder(e.g., ViT) , while textual descriptions undergo tokenization and embedding. These modality-specific representations are subsequently fused via cross-modal attention mechanisms. To enhance precision, the system also

integrates Retrieval-Augmented Generation (RAG) that leverages fused multimodal representations to query external nutritional databases and calibrates model outputs with evidence-based dietary references.

A. Unimodal LLMs

1. Knowledge-Enhanced Dietary Risk Assessment

Abbasian et al. [6]proposed a knowledge-infused LLM-powered conversational health agent (CHA) for precise nutritional risk assessment in diabetic patients. By integrating the American Diabetes Association dietary guidelines and the Nutritionix database, along with analytical tools for nutrient intake calculation, the system leverages the openCHA framework and GPT-3.5-turbo with Tree of Thought prompting as the core LLM. And this method retrieves external knowledge

and performs threshold-based comparisons to identify dietary risks. Evaluated on 100 diabetes-related dietary questions, the agent outperformed GPT4 across all nutrients (e.g., 15% higher accuracy in carbohydrate risk identification).

2. API-Based Nutrition Value Estimation

Haman et al. [7]proposed an AI Dietician System leveraging ChatGPT to estimate nutritional values by querying its API for 236 food items, each repeated five times. The model demonstrated high accuracy for energy values (97% within a 40% deviation from USDA data), moderate consistency (average coefficient of variation: 3.3% for energy), and efficiency in generating meal plans (all 15 meals within 30% caloric bounds). However, accuracy varied across nutrients, with lipids (69.69%) and carbohydrates (80.96%) underperforming compared to energy and protein.

3. Benchmark-Driven Nutrition Estimation from Meal Descriptions

Hua et al. [8]introduced NutriBench, the first publicly available benchmark for evaluating LLMs on nutrition estimation from natural language meal descriptions. They evaluated 12 LLMs (e.g., GPT-4o, Llama3.1) using standard, Chain-of-Thought (CoT), and RAG prompting strategies. Results demonstrated that GPT-4o with CoT achieved the highest accuracy (66.82% within ±7.5g error) for carbohydrate estimation, outperforming professional nutritionists in both speed and precision.

4. Expert-Validated Personalized Nutrition Assistant

Szymanski et al. [9]proposed a GPT-4-based customized nutrition assistant ("The Food Product Nutrition Assistant") by collaborating with registered dietitians (RDs) to validate model outputs through mixed-methods (quantitative Likert scoring and qualitative feedback) and refine responses via structured template instructions. Results showed improved accuracy (mean score 4.40 at Level 3 specificity) and relevance, though challenges remain in generalizing to diverse dietary contexts.

5. Graph-RAG for Complex Nutritional Reasoning

Zhang et al. [11]proposed the Nutritional Graph Question Answering (NGQA) benchmark, which integrates the NHANES and FNDDS datasets to construct a knowledge graph. Leveraging LLMs with graph retrieval-augmented generation (Graph-RAG), the method evaluates the health impact of foods on specific users by linking medical profiles to nutritional tags. The benchmark includes sparse, standard, and complex question settings, achieving F1=0.87 in multi-label classification and ROUGE-L=0.82 in text generation tasks. Experiments demonstrated that LLM-based models excel in complex reasoning but suffer from low efficiency.

B. Multimodal LLMs

1. Nutrition-Specific Tokens for Enhanced Dialogue and Reasoning

Yin et al. [5]proposed FoodLMM, a food assistant system based on a Large Multimodal Model (LMM). This method introduces nutrition-specific tokens (e.g., <total_cal>, <mass>) and regression heads, combined with a two-stage training strategy Experimental results demonstrate that FoodLMM accurately predicts nutrient values (e.g., fat, protein) for specific ingredients and achieves a Mean Absolute Error (MAE) of 67.2 kcal for total calorie estimation in nutrition tasks, outperforming existing methods by 4.5%.

2. Automated Nutritional Profiling via Gemini-Integrated Framework

P. Ushashree et al.[13] proposed a multimodal LLM-based nutrition recognition method that integrates Google Gemini's image recognition capabilities with a structured knowledge base to automate nutritional analysis of food images. The approach employs image segmentation and semantic embedding to extract food features, combined with LLMs for generating nutritional reports. Experiments demonstrated accurate identification of common food components and calorie estimation (average response time <3 seconds), though limitations persist in complex dish recognition and sensitivity to lighting conditions.

3. Vision-Language Alignment for Ingredient and Calorie Estimation

Dongyu Yao et al. [10]proposed CaLoRAify, a framework integrating vision-language models (MiniGPT-v2) with Retrieval-Augmented Generation (RAG) for ingredient recognition and calorie estimation from monocular food images. By aligning visual-text features via Low-rank Adaptation (LoRA) fine-tuning and retrieving external nutritional databases through RAG, the method achieved significant improvements of 55.01% in ROUGE-2 and 61.48% in BLEU metrics compared to baseline models.

4. Nutrient Ranking and Portion Estimation from Meal Images

O'Hara et al. [14]proposed a multimodal LLM-based approach using ChatGPT-4 to estimate nutrient content from meal photographs. The method involved uploading standardized meal images (n=114) to ChatGPT-4 and prompting it to identify foods, estimate portion weights, and calculate nutrient values for 16 nutrients. ChatGPT-4 demonstrated high precision (93.0%) in food identification and adequate ranking of meals by nutrient content (Spearman correlations: 0.29–0.83). However, it systematically underestimated medium/large portion weights ($p<0.001$) and 11/16 nutrients (mean error: 26.9%).

5. Volume Information for Enhanced Food Calorie Estimation

Tanabe and Yanai [12] developed the CalorieVoL framework to enhance calorie estimation. This method combines monocular depth estimation with segmentation models to extract volumetric context, which enables reasoning-based calorie calculation. In zero-shot evaluations on the Nutrition5k dataset, CalorieVoL reduced the MAE from 106.6 kcal to 101.7 kcal and improved the correlation coefficient from 0.688 to 0.708.

Collectively, the exploration of unimodal and multimodal LLMs distinguished by their respective data input reveals remarkable strengths in the application of nutritional recognition. Unimodal approaches, primarily anchored in textual inputs, excel in structured knowledge extraction and semantic parsing due to their linguistic proficiency. These unimodal approaches illustrate a trajectory of external knowledge integration and sophisticated graph augmentation to enhance accuracy in nutritional recognition. Conversely, multimodal approaches integrate heterogeneous visual-textual data through cross-modal attention mechanisms and excel in addressing real-world complexities by aligning visual features with nutritional semantics. Key innovations include nutrition-specific tokenization, real-time image analysis, vision-language alignment, generative reasoning and volume reasoning. While unimodal systems demonstrate efficiency in rapid, text-centric scenarios (e.g., API-based nutrient queries), multimodal frameworks, by contrast, offer superior adaptability to unstructured, real-world inputs, enabling holistic tasks such as nutrient estimation, ingredient analysis and dietary suggestion.

IV. Conclusion

This paper comprehensively reviews the recent advancements and application potential of LLMs in nutrition recognition. The findings demonstrate that LLMs, through multimodal data integration, cross-modal attention mechanisms, and RAG, significantly enhance the accuracy, scalability, and personalization of nutritional analysis. However, critical challenges remain: (1) models heavily rely on meticulously labeled datasets; (2) the trade-off between computational intensity and real-time efficiency remains unresolved; (3) Hallucination in LLMs still exists. Future research should prioritize: (1) developing lightweight architectures and incremental learning strategies to reduce deployment costs; (2) enhancing cross-modal alignment to improve parsing of unstructured inputs; (3) constructing dynamic knowledge graphs that fuse real-time nutritional databases with individual health metrics for precise closed-loop interventions. In conclusion, LLMs present a paradigm-shifting opportunity for nutritional science, yet their applications demand interdisciplinary collaboration and rigorous validation.

Acknowledgment

This research was funded by the Shenzhen Fundamental Research Program (Natural Science Foundation) Stability Support Program for Higher Education Institutions: Intelligent Food Nutrition Analysis Model Based on Large Models.

References

[1] Martinon, P., et al., *Nutrition as a Key Modifiable Factor for Periodontitis and Main Chronic Diseases.* Journal of Clinical Medicine, 2021. **10**(2): p. 197.

[2] Pouladzadeh, P., S. Shirmohammadi, and R. Al-Maghrabi, *Measuring Calorie and Nutrition From Food Image.* IEEE Transactions on Instrumentation and Measurement, 2014. **63**(8): p. 1947-1956.

[3] Shao, W., et al., *Vision-based food* nutrition *estimation via RGB-D fusion network.* Food Chemistry, 2023. **424**: p. 136309.

[4] Zhao, W.X., et al., *A survey of large language models.* arXiv preprint arXiv:2303.18223, 2023. **1**(2).

[5] Yin, Y., et al., *Foodlmm: A versatile food assistant using large multi-modal model.* arXiv preprint arXiv:2312.14991, 2023.

[6] Abbasian, M., et al., *Knowledge-Infused LLM-Powered Conversational Health Agent: A Case Study for Diabetes Patients.* arXiv preprint arXiv:2402.10153, 2024.

[7] Haman, M., M. Školník, and M. Lošťák, *AI dietician: Unveiling the accuracy of ChatGPT's nutritional estimations.* Nutrition, 2024. **119**: p. 112325.

[8] Hua, A., et al., *NutriBench: A Dataset for Evaluating Large Language Models on Nutrition Estimation from Meal Descriptions.* arXiv preprint arXiv:2407.12843, 2024.

[9] Szymanski, A., et al. *Integrating expertise in llms:* crafting *a customized nutrition assistant with refined template instructions*. in *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 2024.

[10] Yao, D., et al., *CaLoRAify: Calorie Estimation with Visual-Text Pairing and LoRA-Driven Visual Language Models.* arXiv preprint arXiv:2412.09936, 2024.

[11] Zhang, Z., et al., *NGQA: A Nutritional Graph Question Answering Benchmark for Personalized Health-aware Nutritional Reasoning.* arXiv preprint arXiv:2412.15547, 2024.

[12] Tanabe, H. and K. Yanai. *CalorieVoL: Integrating Volumetric Context Into Multimodal Large Language Models for Image-Based Calorie Estimation*. in *International Conference on Multimedia Modeling*. 2025. Springer.

[13] Ushashree, P., A. Naik, and P.A.S. Sri. *AI-Driven Health: A Web App for Enhanced Healthcare Queries and Nutrition Analysis*. in *2024 5th International Conference on Smart Electronics and Communication (ICOSEC)*. 2024.

[14] O'Hara, C., et al., *An Evaluation of ChatGPT for Nutrient* Content *Estimation from Meal Photographs.* Nutrients, 2025. **17**(4): p. 607.