Methods and Applications

Large Language Model-Based Text Recognition and Structured Data Extraction for Dietary Surveys

Fangxu Guan[1,&]; Ruixue Niu[2,&]; Feifei Huang[1]; Xiaofan Zhang[1]; Yanli Wei[1]; Jiguo Zhang[1]; Xiaofang Jia[1]; Yifei Ouyang[1]; Jing Bai[1]; Chang Su[1]; Li Li[1]; Wenwen Du[1,#]; Honglei Liu[3,#]; Huijun Wang[1]

ABSTRACT

Introduction: Traditional dietary surveys are time-consuming, and manual recording may lead to omissions. Improvement during data collection is essential to enhance accuracy of nutritional surveys. In recent years, large language models (LLMs) have been rapidly developed, which can provide text-processing functions and assist investigators in conducting dietary surveys.

Methods: Thirty-eight participants from 15 families in the Huangpu and Jiading districts of Shanghai were selected. A standardized 24-hour dietary recall protocol was conducted using an intelligent recording pen that simultaneously captured audio data. These recordings were then transcribed into text. After preprocessing, we used GLM-4 for prompt engineering and chain-of-thought for collaborative reasoning, output structured data, and analyzed its integrity and consistency. Model performance was evaluated using precision and F1 scores.

Results: The overall integrity rate of the LLM-based structured data reached 92.5%, and the overall consistency rate compared with manual recording was 86%. The LLM can accurately and completely recognize the names of ingredients and dining and production locations during the transcription. The LLM achieved 94% precision and an F1 score of 89.7% for the full dataset.

Conclusion: LLM-based text recognition and structured data extraction can serve as effective auxiliary tools to improve efficiency and accuracy in traditional dietary surveys. With the rapid advancement of artificial intelligence, more accurate and efficient auxiliary tools can be developed for more precise and efficient data collection in nutrition research.

The dietary survey evaluated energy and nutrient intake of individuals and groups by collecting food intake data of respondents within a certain period. Traditional dietary survey methods commonly used in large-scale nutrition surveys and monitoring domestically and internationally include food frequency questionnaires (*1*–*3*) and food-weighing accounting methods combined with 24-hour dietary recall (*4*–*5*). At the current stage, traditional dietary surveys face issues such as large recall bias, complex data processing procedures, and a heavy burden on surveyors.

Large language models (LLMs) are artificial intelligence (AI) technologies based on deep learning. In recent years, LLMs have experienced rapid development from theoretical breakthroughs to applications (*6*–*7*). These models exhibit strong capabilities in contextual understanding, knowledge, reasoning, and text generation. In particular, large multimodal models that emerged after 2020 have realized the joint modeling of multidimensional data, such as text, images, and speech, providing technical support for complex scene applications (*8*). These technological breakthroughs have provided new solutions for complex text processing in the medical and healthcare fields. The innovative development of LLMs has provided a breakthrough approach for dietary surveys.

This study intends to conduct semantic recognition and data information structure research on 24-hour dietary survey text information based on LLM and explore the feasibility of AI in improving efficiency and assisting the investigators in completing the dietary surveys.

METHODS

Study Population

The study subjects were recruited from 15 families in the Huangpu and Jiading districts of Shanghai. In total, 38 individuals (86 person-times) completed the 24-hour dietary recall survey between October and November 2024.

Dietary Data Collection Methods and Processing

Referring to standardized logic and asking for a 24-hour dietary recall method of a 24-hour dietary survey, the investigators used a computer-assisted dietary survey interviewing system to conduct a face-to-face dietary survey at home. The survey included questions regarding food and beverage intake by the participants in the past 24 h, including meal time, food/ingredient name, consumption, edible part, dining, and production location. Simultaneously, an intelligent recording pen (SR101T, iFLYTEK Co., Ltd.) recorded the interview process and transcribed it into text, which was then used for text recognition and data structuring with the LLM.

Text Information Recognition and Data Structuring of LLM

We used GLM-4 to generate the structured data. GLM-4 is a fourth-generation base large model released by Zhipu AI (Beijing Zhipu Huazhang Technology Co., Ltd.) on January 16, 2024. The dietary survey text data processing constructed in this study covers the following core steps: First, data cleaning, preprocessing, and de-identification were conducted on the recorded text of the on-site dietary survey to ensure data quality and compliance. Second, we combined two key technologies: prompt engineering and chain-of-thought prompts. Finally, the structured output was generated. We designed a complete quality control process, including a logic consistency check, statistical alignment analysis of results and manual annotations, and error analysis (Figure 1).

Statistical Analysis

The dietary survey data collated and uploaded to the server by the investigator were downloaded and analyzed to describe the basic situation of the research object. The structured dietary data output by the LLM was compared with the dietary survey recording text, and the integrity of the LLM in the identification of key information was analyzed, such as whether the text

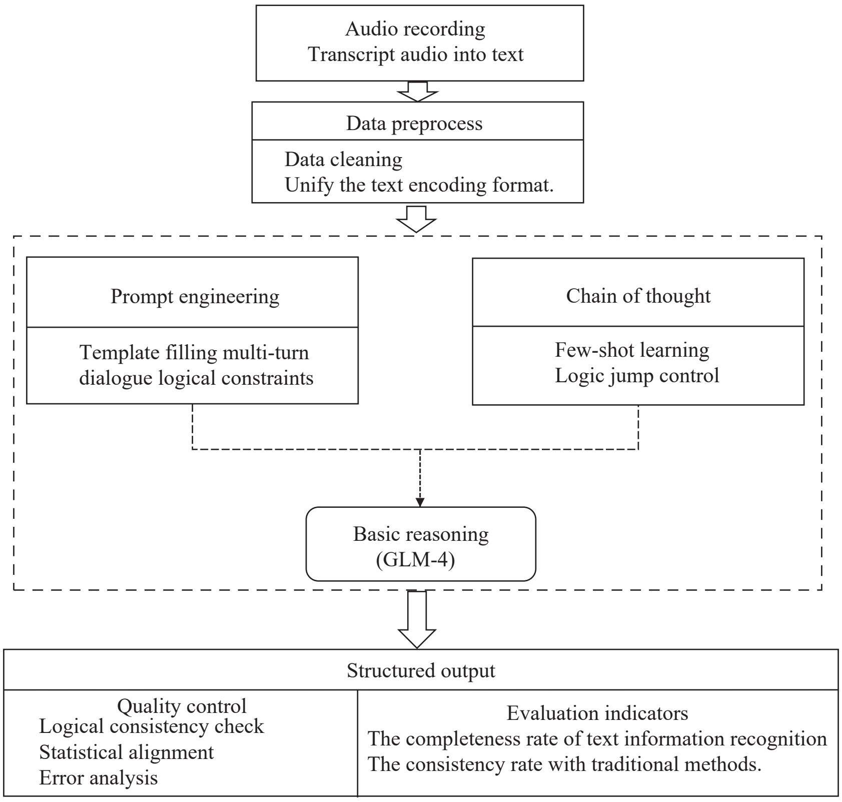
Audio recording
Transcript audio into text
Data preprocess
Data cleaning
Unify the text encoding format.
Prompt engineering
Template filling multi-turn dialogue logical constraints
Chain of thought
Few-shot learning
Logic jump control
Basic reasoning (GLM-4)
Structured output
Quality control
Logical consistency check
Statistical alignment
Error analysis
Evaluation indicators
The completeness rate of text information recognition
The consistency rate with traditional methods.

FIGURE 1. Flowchart of text information recognition and data structuring based on LLM.

mentioned "Mealtime" and the LLM recognized and produced effective structured data of "Mealtime." Using the transcripts with complete identification information, the dietary data structured by the LLM were compared with those collated and uploaded by the investigator, and the consistency between the two methods in assigning key variables was analyzed.

RESULTS

Basic Information

Thirty-eight family members from 15 households participated in the dietary survey and provided complete dietary survey data for 1–3 days, with a total of 86 person-day recalls, as shown in Table 1. Among them, children and adolescents under 18 years accounted for 7.9%, adults aged 18–59 years for 57.9%, and older adults aged 60 years and above for 34.2%. The respondents were primarily female, accounting for 63.2% of the sample.

Integrity of Text Recognition in LLM

Table 2 shows that the integrity of key information recognition using the LLM for dietary survey recording text was generally high, with an overall integrity rate of 92.5%. Among them, the recognition integrity rate for "Mealtime" and "Dining location" were the highest reaching 100%. The second is the recognition integrity rate for "Food name," "Ingredient name," and "Production location", with recognition integrity rate above 90%. In contrast, the text recognition of "Consumption weight" and "edible or includes nonedible" was slightly lower, with the integrity rate of information extraction is 80%–86%.

Consistency Analysis of Structured and Traditional Dietary Survey Data Assisted by LLM

Table 3 further analyzes the consistency between the results of structuring the dietary survey data assisted by LLM and the data uploaded by traditional investigators. The results show that the overall

TABLE 1. Basic characteristics of the survey participants involved in the dietary survey.

Characteristics	Jiading district	Huangpu district	Total
Household survey, *N* (%)	6 (40.0)	9 (60.0)	15
Individuals, *N* (%)	15 (39.5)	23 (40.5)	38
person-times, *N* (%)	35 (40.7)	51 (59.3)	86
Age group, %			
<18 years	6.7	8.7	7.9
18–59 years	73.3	47.8	57.9
≥60 years	20.0	43.5	34.2
Gender, %			
Male	46.7	30.4	36.8
Female	53.3	69.6	63.2

TABLE 2. Recognition integrity rate of dietary survey text information using LLM.

Survey indicators	Jiading district, %	Huangpu district, %	Total, %
Mealtime	100.0	100.0	100.0
Food name	95.0	96.0	95.5
Ingredient name	93.0	92.0	92.5
Consumption weight	85.0	84.0	84.5
Edible or includes nonedible	86.0	80.0	83.0
Dining location	100.0	100.0	100.0
Production location	92.0	89.0	90.5
Overall integrity rate	93.0	92.0	92.5

Abbreviation: LLM=large language model.

consistency rate between the automatically structured data of the dietary survey recording text using the LLM and the data recorded and decomposed by traditional investigators is 86%. From the perspective of different survey indicators, the higher consistency rates are "Mealtime", "Dining location," and "Production location," and the agreement rate can reach more than 90%. The second is the agreement rate of "Food name" and "Ingredient name," which can reach more than 80%. The indicators with a relatively lower consistency rate are "consumption weight" and "edible or includes nonedible," ranging from 70% to 80%.

Advantages and Challenges of Text Recognition and Structuring of Dietary Surveys Assisted by LLM

We further evaluated the model using Pression and F1 scores, with the LLM achieving 94% precision and an F1 score of 89.7% on the full dataset. Table 4 presents the advantages and challenges of applying the LLM to nutritional surveys and our proposals to improve this method.

DISCUSSION

Our study used a dialogue survey, recorded and transcribed the dialogue into text, and used LLM-based text recognition and structured data extraction. The structured data were then compared with manual recordings. The overall recognition integrity of the survey was approximately 92.5%. The recognition completeness for food types and ingredients was 95.5% and 92.5%, respectively. The consistency rate between LLM-assisted structured dietary survey data and data from the traditional survey method was 86%. The major discrepancies occurred in the identification of the consumption weight and the ability to distinguish between edible and nonedible parts. Responses to these

TABLE 3. Consistency rate between the structured data of the LLM-assisted dietary survey and the data of the traditional survey method.

Survey indicators	Jiading district, %	Huangpu district, %	Total
Mealtime	99.0	99.0	99.0
Food name	83.0	94.0	88.5
Ingredient name	81.0	85.0	83.0
Consumption weight	78.0	77.0	77.5
edible or includes nonedible	70.0	73.0	71.5
Dining location	87.0	99.0	93.0
Production location	87.0	95.0	91.0
Overall consistency rate	83.0	89.0	86.0

Abbreviation: LLM=large language model.

TABLE 4. Advantages and challenges of text recognition and structuring of dietary surveys using LLM.

Survey phases	Advantages	Challenges	Improvement direction
Recording collection	Nondisruptive collection, reducing the workload of investigators.	There are dialect barriers in the transcription of recordings, and the scope of application needs to be improved.	The investigator should double check the key information.
Text recognition	Automatically remove invalid conversations and improve the efficiency of identifying key information.	There are certain limitations in the complex semantic recognition of quantitative or professional indicators, such as consumption amount, edible or includes nonedible food.	Continuously train the text understanding ability, logical computing ability, and professional judgment ability of LLM, and improve their adaptability.
Data structuring	Automatically extract and assign key indicators, forming a structured database in real time, reducing the need for investigators to reorganize and input data manually.	There is a lack of information about the exact amount of food or the breakdown of the amount of food consumed by multiple people in a household. There also some omissions or misjudgments in the determination of commercially available products.	Optimize the underlying architecture to provide the average weight or range of different portion sizes of common foods. Combine semantic understanding and logical computing to achieve quantitative functionality in different question and answer scenarios.

Abbreviation: LLM=large language model.

two questions were vague. The intake amount is generally described using quantifiers or gestures, and further analysis is required to convert it into grams or milliliters. Distinction between edible parts and commercially available products is currently made via manual identification. Additional data training is required to establish a database for large models, and multidimensional data (such as images and videos) support should be combined to increase the accuracy of these two problems.

This method was implemented as an auxiliary tool for dietary surveys. With the advent of the AI, a revolution has occurred in the field of medical research. In large-scale population studies, there is a need to improve survey methods without increasing their complexity and cost, considering the convenience and universality of the methods. In the future, we will attempt to combine voice transcription and image recognition based on deep learning. Sun et al. developed a deep-learning-based image recognition model for food identification at the ingredient level to conduct health management for patients with diabetes (*9*). This approach can be further advanced by testing AI architectures on a limited number of large-scale food image and nutrition databases (*10*).

However, AI has not been widely applied to large-scale population surveys. In medical ethics, surveys using AI face many challenges, owing to aspects such as user privacy, data security, and explainability. China is also accelerating the introduction of a series of ethical review norms and measures, in the hope that AI can be fully applied under ethical conditions. We will continue to monitor developments in AI (artificial intelligence) technology ethics management and research the application of large AI models to a larger population.

In the long run, large models will drive paradigm changes in nutritional research. At the public health level, the LLM can be applied to nutrition, spanning smart and personalized nutrition, dietary assessment, food recognition and tracking, predictive modeling for disease prevention, and disease diagnosis and monitoring (*11*). With the development of deep-learning technology, a distributed learning framework can solve the data island problem and accelerate cross-institutional and cross-regional nutrition and health research collaborations.

Acknowledgements: The efforts of the Shanghai Center for Disease Control and Prevention and all participants and interviewers of this study.

Conflicts of interest: No conflicts of interest.

Ethical statment: Approved by the institutional review board of the National Institute for Nutrition and Health, China CDC (No: 2022-024), and all participants provided written informed consent before participation.

Funding: Supported by the Ministry of Finance of the People's Republic of China from 2022 to 2024 (grant number 102393220020070000016).

doi: 10.46234/ccdcw2026.008

[#] Corresponding authors: Wenwen Du, duww@ninh.chinacdc.cn; Honglei Liu, liuhonglei@ccmu.edu.cn.

[1] Key Laboratory of Public Nutrition and Health, National Health Commission of the People's Republic of China; National Institute for Nutrition and Health, Chinese Center for Disease Control and Prevention & Chinese Academy of Preventive Medicine, Beijing, China; [2] School of Software Engineering, Beijing Jiaotong University, Beijing, China; [3] School of Biomedical Engineering, Capital Medical University, Beijing, China.
[&] Joint first authors.

Submitted: September 28, 2025
Accepted: January 02, 2026
Issued: January 09, 2026

REFERENCES

1. Sierra-Ruelas É, Bernal-Orozco MF, Macedo-Ojeda G, Márquez-Sandoval YF, Altamirano-Martínez MB, Vizmanos B. Validation of semiquantitative FFQ administered to adults: a systematic review. Public Health Nutr 2021;24(11):3399 – 418. https://doi.org/10.1017/S1368980020001834.
2. Cui Q, Xia Y, Liu YS, Sun YF, Ye K, Li WJ, et al. Validity and reproducibility of a FFQ for assessing dietary intake among residents of northeast China: northeast cohort study of China. Br J Nutr 2023;129(7):1252 – 65. https://doi.org/10.1017/S0007114522002318.
3. Zhao D, Gong YY, Huang LY, Lv RX, Gu YX, Ni CX, et al. Validity of food and nutrient intakes assessed by a food frequency questionnaire among Chinese adults. Nutr J 2024;23(1):23. https://doi.org/10.1186/s12937-024-00921-9.
4. Mao YK, Weng JY, Xie QY, Wu LD, Xuan YL, Zhang J, et al. Association between dietary inflammatory index and Stroke in the US population: evidence from NHANES 1999-2018. BMC Public Health 2024;24(1):50. https://doi.org/10.1186/s12889-023-17556-w.
5. Yu DM, Zhao LY, Zhao WH. Status and trends in consumption of grains and dietary fiber among Chinese adults (1982-2015). Nutr Rev 2020;78(Suppl 1):43 – 53. https://doi.org/10.1093/nutrit/nuz075.
6. Bedi S, Liu YT, Orr-Ewing L, Dash D, Koyejo S, Callahan A, et al. Testing and evaluation of health care applications of large language models: a systematic review. JAMA 2025;333(4):319 – 28. https://doi.org/10.1001/jama.2024.21700.
7. Shool S, Adimi S, Saboori Amleshi R, Bitaraf E, Golpira R, Tara M. A systematic review of large language model (LLM) evaluations in clinical medicine. BMC Med Inform Decis Mak 2025;25(1):117. https://doi.org/10.1186/s12911-025-02954-4.

8. Shen YQ, Xu YQ, Ma JJ, Rui WS, Zhao C, Heacock L, et al. Multimodal large language models in radiology: principles, applications, and potential. Abdom Radiol (NY) 2025;50(6):2745 – 57. https://doi.org/10.1007/s00261-024-04708-8.
9. Sun HN, Zhang K, Lan W, Gu QF, Jiang GX, Yang X, et al. An AI dietitian for type 2 diabetes mellitus management based on large language and image recognition models: preclinical concept validation study. J Med Internet Res 2023;25:e51300. https://doi.org/10.2196/51300.
10. Shonkoff E, Cara KC, Pei XC, Chung M, Kamath S, Panetta K, et al. AI-based digital image dietary assessment methods compared to humans and ground truth: a systematic review. Ann Med 2023;55(2):2273497. https://doi.org/10.1080/07853890.2023.2273497.
11. Theodore Armand TP, Nfor KA, Kim JI, Kim HC. Applications of artificial intelligence, machine learning, and deep learning in nutrition: a systematic review. Nutrients 2024;16(7):1073. https://doi.org/10.3390/nu16071073.