Towards Automated Recipe Reconstruction: Optimization of Dietary Data Collection using Information Retrieval, Large Language Models and Mathematical Optimization

Svetlana Schmidt*[2], Linda Klasen[2], Ute Nöthlings[2], and Rafet Sifa[1,2,3]

[1]Fraunhofer IAIS, Sankt Augustin, Germany
[2]University of Bonn, Bonn, Germany
[3]Lamarr Institute for Machine Learning and Artificial Intelligence, Germany

Abstract

Accurate and scalable collection of dietary data is vital for advancing nutritional epidemiology and understanding links between diet, public health, and environmental sustainability. A key challenge is the collection of the detailed nutrition data on the product level which currently largely relies on manual recipe reconstruction.

We propose computational approaches to optimize this workflow. First, an information retrieval (IR)–based recommender system integrates food-category prediction with retrieval over product text, ingredients, and nutrient profiles to streamline food item matching and reduce redundancy across the database. Second, we outline a roadmap for automated recipe reconstruction that combines large language models (LLMs) for ingredient parsing with nutrient-constrained mathematical optimization for recipes reconstruction.

By integrating machine learning, generative modeling, and optimization, our work enhances the efficiency, transparency, and scalability of nutrition data collection, laying a foundation for sustainable practices in nutritional epidemiology and research on interactions of the diet, health and environment.

Keywords: nutrition data collection, nutritional epidemiology, recipe reconstruction, information retrieval, food categorization, machine learning, large language models, mathematical optimization.

*First and corresponding author: sschmid4@uni-bonn.de

1 Introduction

Dietary intake plays a crucial role towards a sustainable future. Thus, examining dietary behaviors has gained increasing importance in recent years. Notable trends include reducing the consumption of animal-based foods, transitioning toward more sustainable dietary patterns, and placing greater emphasis on proper nutrition. These objectives are the critical components in the achievement of the Sustainable Development Goals (SDGs) from the United Nations 2030 Agenda for Sustainable Development[1]. The influence of dietary habits on the public health and the environment constitutes a central focus within the field of nutritional epidemiology.

Research on planetary health and environmental impacts demands highly accurate nutritional data assessment. Given the accelerating pace of climate change, efficient collection and evaluation of such data are essential for investigating the interconnections between dietary habits, human health, and the environment.

Nutritional data are typically collected through established dietary data collection methodologies, such as weighed dietary records (WDR) [12]. In this approach, participants are instructed to weigh and document all food and drink intake over a specific period, often spanning several days. The NutriDiary app [12] developed at the University of Bonn facilitates the data collection by enabling participants to report the weighted quantities of consumed food and beverages using their smartphones. Users can select foods and beverages from a drop-down menu within the app, scan the barcode from the packaging, or upload photographs of the packaging when specific foods are not available in the database.

Accurate representation of both macro- and micronutrient content is essential to preserve the validity and internal consistency of nutritional analyses, which in turn makes rigorous procedures for dataset expansion indispensable. The NutriDiary database constitutes a unique and comprehensive resource, encompassing food items and their associated nutritional profiles at both macro- and micronutrient levels. For processed foods, entries are further linked to reconstructed recipes, detailing ingredient lists and corresponding quantities ([12], [22]). As emphasized by [1], recipe reconstruction enables a more accurate and holistic estimation of nutrient intake, thereby supporting the reliability of nutritional epidemiological analyses.

Although recipe reconstruction is critical for precise and consistent documentation of food intake, it represents a considerable challenge. The integration of new food items into the database is

[1]2030 Agenda for Sustainable Development `https://sdgs.un.org/2030agenda`

particularly labor-intensive, as packaging typically does not disclose ingredient proportions. Consequently, dietitians must manually estimate ingredient compositions based on the declared nutrient content, relying on their expertise. To meet the increasing demands of nutritional research and maintain scalability, this process requires methodological optimization.

A large number of recent nutritional data collection applications focus on obtaining nutritional information through food image recognition, barcode scanning of packaged products, or manual selection from predefined food item lists, for subsequent food type classification and nutrient intake analysis ([29], [15], [6]). These tools frequently integrate with proprietary databases containing nutritional profiles and enable direct computation of nutrient and food intake. However, a major drawback of such systems is the lack of transparency, which makes it unclear whether the nutrient information of packaged foods is reconstructed via ingredient proportion estimation or entered directly from the packaging information.

Complementary approaches utilize the machine learning techniques to estimate selected nutrients or ingredients ([4], [16]). Nonetheless, these studies typically limit their scope to the information available from food packaging. This makes it unclear whether they account for the full spectrum of nutrients—up to 82 distinct macro- and micronutrients—that are relevant for comprehensive dietary data assessment [12].

The recipe reconstruction for the NutriDiary database can be divided into two main components: *recipe matching* and *recipe simulation*. A substantial part of new food products closely resemble existing entries in NutriDiary [12]. Those products are integrated by aligning their nutritional profiles. To optimize this process, we propose an information retrieval (IR) system that combines food category prediction with scoped retrieval based on product text, ingredients, and nutrient profiles. Unlike traditional term frequency–inverse document frequency (TF–IDF) or Best Matching 25 (BM25) methods ([19], [13]), our approach follows a two-stage IR design [2], reducing the search space before applying fine-grained similarity measures.

When new food items cannot be incorporated into the database by direct matching, their nutrient values and ingredient proportions must be reconstructed through recipe simulation, which is currently performed manually [12]. Mathematical optimization has long been recognized as an effective method for estimating ingredient proportions. Linear and quadratic programming outperform manual trial-and-error approaches by achieving comparable accuracy and efficiency as it was demonstrated in [28]. More recently, [1] applied optimization to over 1,800 German food products, using automated mapping from packaging ingredient lists to nutritional tables as a preprocessing step. Building on these foundations, recent advances in large language models (LLMs) highlight the potential of Chain-of-Thought (CoT) prompting to support ingredient parsing and alignment, as demonstrated on the NutriBench benchmark [9]. We therefore propose a hybrid pipeline that leverages CoT-guided models for ingredient normalization and mapping, followed by nutrient-constrained mathematical optimization for proportion estimation.

The overarching aim of our work is to optimize the methods of expanding the dietary dataset and data analysis. Integrating machine learning methods into data collection will open new avenues for future research in nutritional science. A central long-term research question of our work concerns how sustainable approaches to data collection and assessment can enhance our understanding of environmental impacts and public health. We are not aiming to supplant established data assessment practices in nutritional epidemiology, but rather to optimize the recipe reconstruction process by integrating advanced computational approaches, including machine learning and generative modeling into the practical toolset of experts.

In this work, we propose a set of solutions tailored to the specific needs and challenges connected to the institute of Nutrition and Food Sciences at the University of Bonn. The contributions of our work are twofold:

- We introduce an information retrieval–based recommender system, augmented with food-category filtering, to enhance the relevance of search results and thereby support and streamline the recipe matching task. Initial experiments and results highlight the potential of this approach.
- We outline a roadmap for automated recipe reconstruction that integrates language models with mathematical optimization techniques.

2 Related Work

The dietary data assessment is the key part of the nutritional epidemiological studies conducted in University of Bonn. The NutriDiary app [12], developed at the Institute of Nutritional and Food Sciences [2] at University of Bonn, supports the collection of WDR. Users can select food and beverage items from a drop-down menu or upload photographs of packaging when specific items are unavailable in the database. The app relies on the proprietary LEBTAB database ([22], [17]), a comprehensive resource containing detailed macro- and micronutrient profiles for a wide range of foods. LEBTAB was compiled within the Dortmund Nutritional and Anthropometric Longitudinally Designed (DONALD) study ([12, 17, 22]). Ongoing projects, such as the Cohort on Plant-based Nutrition (COPLANT) study[3], further extend this work by collecting dietary data from approximately 6.400 participants. The basic food entries mainly correspond to German standard food table "Bundeslebensmittelschlüssel" (BLS)[4]. Some food items missing in BLS were added from the U.S. Department of Agriculture (USDA)[5] database.

A substantial proportion of new products with nearly identical compositions are incorporated into the NutriDiary database by matching their ingredient proportions and nutritional profiles with existing entries. Implementing the information retrieval system would improve the process of adding new food items into the dataset. Traditional approaches such as TF-IDF and BM25 have long served as strong baselines for lexical similarity matching [19], from the information retrieval perspective.

[2] Institut für Ernährungs- und Lebensmittelwissenschaften (IEL): https://www.iel.uni-bonn.de/de/startseite-iel?set_language=de

[3] COPLANT studyhttps://www.epi.uni-bonn.de/forschung/coplant-studie

[4] https://blsdb.de/

[5] https://fdc.nal.usda.gov/food-search?type=Foundation

Current IR systems often adopt a two-stage architecture: a retriever first narrows the search space, followed by a more precise re-ranking stage using a larger set of textual features [2]. We propose a system that mirrors this architecture by first predicting the food category and then conducting scoped retrieval using both textual product and ingredient descriptions and numeric nutrient profile similarity.

We aim to enhance the IR system by incorporating searches based on intrinsic features of the food item, such as its food category. A similar food categorization approach was proposed by [14], who applied both machine learning and deep learning techniques for food category and nutrient prediction, and demonstrated that deep learning models achieved strong performance in predicting the food category of packaged food products from the United States. A comprehensive survey of food and nutrition recommender systems was provided by [24], showing that the field has largely concentrated on consumer-oriented applications such as recipe or meal recommendation. The absence of the expert-facing systems suggests that there is little research on information retrieval tools designed for use by nutritional epidemiology experts domain. Our work addresses this gap by designing an IR system intended for expert use.

Mathematical optimization techniques offer a solid solution to ingredient proportion estimation. This methodology has been established in the nutritional epidemiology field for years. Reference [28] proposed using linear and quadratic programming for this purpose, noting that iterative trial-and-error methods are time-consuming, error-prone, and may lead dietitians to overlook discrepancies in nutrient calculations to expedite results. Their study demonstrates that mathematical optimization can achieve comparable or greater accuracy more quickly than trial-and-error estimation. Building on the concept of mathematical optimization [1] utilize this method for estimating the ingredient proportions for the dataset of the 1804 food products present on German market. The proposed approach utilizes the automated mapping of the ingredients list provided on the packaging to nutritional tables, thus pre-processing data for the mathematical optimization.

Recent research demonstrates that Chain-of-Thought (CoT) prompting enhances LLM performance by enabling the decomposition of complex problems into smaller tasks and improving explainability. The CoT approach closely mirrors the human-like logical reasoning processes [27]. The capacity of LLMs to solve mathematical and decision-making problems through the CoT approach was investigated in [7], demonstrating that autoregressive Transformers can reliably produce correct solutions to complex tasks when provided with sufficient CoT demonstrations.

Finally, [9] proposed CoT-prompting approach for solving the nutrition estimation problem. They proposed the NutriBench dataset, which was developed for LLM evaluation, and was used to train models to estimate daily caloric intake. However, the task of ingredients estimation is more complex than computing one of the nutritional values. To address this, we propose to utilize the generative LLMs with CoT prompting for aligning the ingredients from the packaging with those from the proprietary database and their nutritional profiles for further mathematical optimization.

3 Data in nutritional epidemiology

3.1 Data

NutriDiary database is the core dietary data collection of epidemiological nutritional studies at the Bonn University. Each food or beverage contains the food item code, the manufacturer name from the packaging and the unique food item identifier. Using the food item code, each entry is linked to the LEBTAB database, which provides the corresponding nutritional profile and ingredient information. The LEBTAB database contains around 19,000 generic and branded food items [12]. The nutritional data comprises energy content and values for 82 nutrients, including the Big7, - nutrients typically found on the food packaging. Table 1 illustrates the Big7 values for quark dessert for children. Multiple food items may share the same food item code, indicating that they are nearly identical in composition. The ingredient lists and proportions of such items are highly similar, resulting in identical nutritional profiles. Such equivalent items are added to the dataset by trained dietitians through a recipe matching procedure, which involves manually comparing the ingredient lists and nutritional values from packaging with those already present in the database. The entries from NutriDiary database are linked to their corresponding entries from the LEBTAB dataset based on the food item code. As a result we obtain the dataset with 30,336 items in total.

Table 1: Example of nutrient profile (Big 7 + key micronutrients) for "Quark dessert Vanilla with Calcium and Vitamin D".

Nutrient	**Per 100 g**	**Unit**
Energy (kcal)	86.86	kcal
Fat	2.79	g
Saturated Fat	1.68	g
Carbohydrates	12.09	g
Sugars	10.81	g
Protein	3.34	g
Salt	0.10	g
Calcium	180.00	mg
Vitamin D	1.25	μg

Processed foods are also linked to the ingredients database via the special food code, with the corresponding ingredient weights specified. These ingredients, likewise identified by a food code, can be matched to their own nutritional profiles, enabling comprehensive compositional analysis. This ingredient database is a result of the recipe simulations conducted by dietitians.

Moreover each item is linked with the corresponding food category. The LEBTAB taxonomy comprises 122 food categories in total ranging from fresh dairy products to specialized infant formulas and supplements, covering the full spectrum of packaged and processed foods in the German market. Figure 1 demonstrates top seven most populated food categories in the database which we utilize for IR system.

The distribution of products across food categories in our dataset is highly imbalanced. Several categories contain a large number of items (e.g. bread or soft drinks), whereas many categories are sparsely populated. Specifically, there are 18 categories with fewer than 10 products, indicating a long-tail distribution that may affect classification performance. Cooked legumes, for example canned bean mix, is an underrepresented category with 11 items, but important for recipe matching, so we include it for training to enable the recommender to recognize legume products.

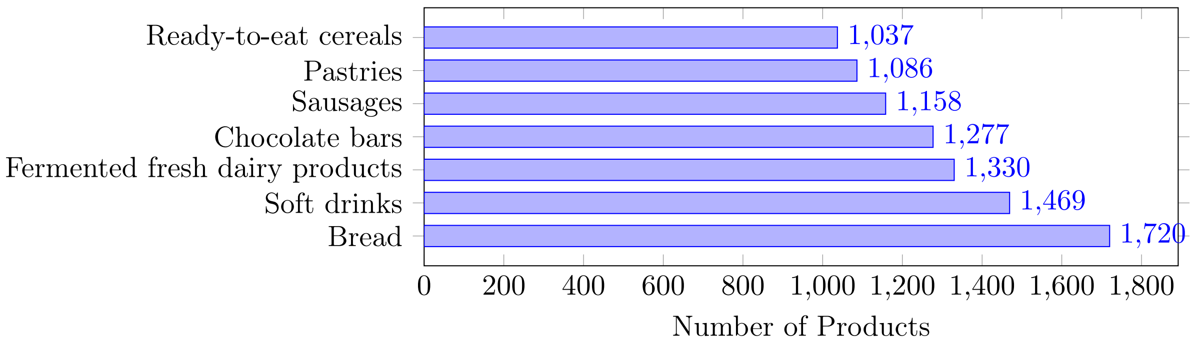

Figure 1: Counts of selected food categories in in the LEBTAB dataset.

3.2 Recipe matching and recipe simulation

The expansion of nutritional database for the NutriDiary application is managed by dietitians and nutritional scientists. A substantial proportion of new products with nearly identical compositions are incorporated into the database by matching their ingredients and nutritional profiles with existing entries [12]. To be considered a valid match, food products must align with an existing entry in terms of food category, listed ingredients, and overall nutritional composition.

In cases where no suitable equivalent is identified within the database, the new product is added via recipe simulation, - ingredient weights estimation based on the product information provided on the packaging. The manufacturer is obliged to provide the nutrient profile and ingredients' list, but not the ingredient proportions. The recipes are simulated utilizing the data analytics software SAS[6].

Dietitians input the approximate initial ingredient portions, based on the order of ingredients listed on the packaging, into the SAS software. The software has an access to the LEBTAB ingredient database, thus the recipe is simulated based on the food entries from this proprietary dataset. The calculated nutrient values indicate whether the initial estimated ingredient portions are accurate. The ingredient weights are then adjusted iteratively until the calculated nutrient values closely match those provided on the packaging.

Both recipe matching and recipe simulation are currently performed manually, which makes these processes labor-intensive and error-prone.

Nutritional constraints

Dietitians apply strict constraints when adding new food products to the database via recipe matching and simulation:

[6] https://www.sas.com/en_us/home.html

- The nutrient profile of a candidate match must not deviate by more than 20% for each nutrient relative to the query product.
- The main ingredients of the query and the candidate match, in particular, the top-listed ingredients must be consistent. Products with differing key ingredients are not considered equivalent while recipe matching.

Applying these constraints while IR-based recipe matching ensures that retrieved matches closely align with expert judgment, thereby maintaining the quality and reliability of the database.

4 Recipe matching with information retrieval system

4.1 Methodology

4.1.1 Information retrieval

The first component of our content-based information retrieval system is the text pre-processing module. This involves normalization, such as lowercasing and tokenization of the text. Nutritional and food product information representation includes:

- TF-IDF representations of the food product name and ingredients for each entry in the dataset.
- Each food item is aligned with the Big7 nutrients and ingredients proportions.
- Measurement units are normalized, with all values represented in grams.
- The nutrient profile of each entry is represented as a normalized vector.

Table 2: Performance comparison of models on food category classification. Reported values are macro-averaged metrics; values in parentheses correspond to weighted-averaged metrics. Logistic Regression, Linear SVM, and the neural network were trained with class balancing from the outset, while BERT and DistilBERT were additionally evaluated with and without class-weighted loss (CWL) to assess its impact.

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	84.6%	84.1% (88.6%)	85.7% (84.6)	82.9% (85.4%)
LinearSVC	**96.9%**	**95.4% (97.8%)**	**94.7% (96.9%)**	**94.5% (97.1%)**
SVC (RBF kernel)	98.6%	98.5% (99%)	97.4% (98.6%)	97.3% (98.5%)
SVC (Polynomial kernel)	**99%**	**98.7% (99.3%)**	**97.8% (99%)**	**97.7% (98.9%)**
Neural Network	92.6%	87.1% (92.7%)	85.2% (92.6%)	84.7% (91.9%)
DistilBERT	92.3%	81.3% (92.1%)	81.1% (92.3%)	80.3% (91.6%)
DistilBERT CWL	93%	88.1% (93.1%)	89.1% (93%)	87.5% (92.4%)
BERT	94.3%	86.7% (94.7%)	85.9% (94.3%)	85.4% (93.9%)
BERT CWL	94%	90.3% (95.2%)	90.6% (94%)	89.8% (94.1%)

Nutritional constraints are combined with TF-IDF textual representations and nutritional vectors to compute deviations and rank the best matches using cosine similarity. The constraints allow up to a 20% deviation between the nutritional profiles of matching food products, restricting consideration to items within this threshold. Cosine similarity is used to rank the best matches based on food product and ingredient representations. The system outputs the top k matches. If the food item already exists in the database, it appears among the recommended results.

For each product, we compute similarity to the query by combining nutrient-based and text-based measures, as it is shown in (1). Nutrient similarity is computed using the Euclidean distance.

$$S_{\text{nutrients}}(q, x) = 1 - \frac{\|\mathbf{q} - \mathbf{x}\|_2}{\sqrt{d}} \tag{1}$$

The $\mathbf{q} \in \mathbb{R}^d$ denote the query nutrient profile (vector of length d), where each value is scaled to [0,1] and $\mathbf{x} \in \mathbb{R}^d$ the product nutritional information. Textual similarity in (2) of two terms is measured via cosine similarity between the vectorized with TF–IDF representations of product name and ingredients.

$$S_{\text{name}}(q, x) = \cos\big(\mathbf{t}_q^{\text{name}}, \mathbf{t}_x^{\text{name}}\big) \tag{2}$$

$$S_{\text{ingredients}}(q, x) = \cos\big(\mathbf{t}_q^{\text{ing}}, \mathbf{t}_x^{\text{ing}}\big). \tag{3}$$

The combined text similarity interpolates both sources of food item similarity:

$$S_{\text{text}}(q, x) = \beta\, S_{\text{name}}(q, x) + (1 - \beta)\, S_{\text{ingredients}}(q, x) \tag{4}$$

Finally, the overall similarity of the food items is a weighted combination of nutrients and textual similarity:

$$S_{\text{final}}(q, x) = \alpha\, S_{\text{nutrients}}(q, x) + (1 - \alpha)\, S_{\text{text}}(q, x) \tag{5}$$

We empirically set $\alpha = 0.7$ and $\beta = 0.5$ to prioritize nutrient similarity while giving equal weight to product name and ingredient text within the textual similarity component.

4.1.2 Food category based information retrieval

Food category classification models can facilitate for the retrieval of food items that belong to the same food group as a given query. We assume that items within the same category tend to have similar nutritional profiles, ingredient composition, and product names. This categorization approach could not only improve the relevance of retrieved matches but also accelerate the search process by narrowing the search space. Given the large number of food categories, with each product assigned to a single food group, we trained a multinomial logistic regression[7] (LR) classifier. Additionally, we trained the linear Support Vector Classification[8] (LinearSVC) model for food categorization task. The input to those models was generated using TF-IDF representations derived from the combined textual content of product names and ingredient lists. To assess the benefits of non-linear decision boundaries, we additionally trained SVC[9] models with polynomial and radial

[7] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html
[8] https://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html
[9] https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html

basis function (RBF) kernels under the same pre-processing pipeline.

As a potential enhancement to the classification framework, we implemented simple feedforward neural network to assign food items to their corresponding food groups. The model was trained on input features generated using TF-IDF representations derived from the concatenated product names and ingredient lists. Food categories were encoded as numerical labels, enabling the use of sparse categorical cross-entropy loss. To mitigate class imbalance, class weights inversely proportional to category frequencies were applied during training. Hyperparameters, such as layer sizes and dropout were optimized with Keras Tuner[10]. The network architecture employed the rectified linear unit (ReLU)[11] activation function in the hidden layers, while the softmax function was applied in the output layer to produce normalized probability distributions over the set of possible classes.

Further we finetuned a pre-trained for german language DistilBERT[12] [20] and BERT[13] [5] models for food group prediction as we assumed the improvement in classification performance due to leveraging the contextual language understanding. DistilBERT[14] and BERT[15] tokenizers are used for tokenization of the input text. To prevent the overfitting, the model was trained for three epochs using the Hugging Face Trainer[16] API with cross-entropy loss and evaluated on a held-out set of manually annotated queries.

To ensure robustness under class imbalance, all classifiers, except transformer-based models, were trained with class-balancing techniques, implemented through class-weight adjustments in their respective loss functions or regularization terms. The impact of these balancing strategies is compared against transformer models trained without balancing in a later section.

4.2 Experiments and Results

4.2.1 Data annotation

To evaluate the performance of the IR approaches, we initially retrieved random samples using our baseline algorithm. These samples were then assessed by dietitians for the relevance of the retrieved matches, adhering to the established constraints for recipe matching. Each sample was annotated by a single dietitian authorized for database management and expansion.

To verify inter-rater consistency, we conducted a control cross-validation task on a small subset (5%) of the annotated data. The resulting average Jaccard similarity measure indicated that in 97% of cases, the dietitians reached the same decisions, confirming a high level of agreement among experts. The dietitians were instructed to annotate all relevant matches among the retrieved and

[10] https://www.tensorflow.org/tutorials/keras/keras_tuner
[11] https://www.tensorflow.org/api_docs/python/tf/keras/layers/ReLU
[12] distilbert-base-german-cased https://huggingface.co/distilbert/distilbert-base-german-cased
[13] bert-base-german-cased https://huggingface.co/google-bert/bert-base-german-cased
[14] https://huggingface.co/docs/transformers/en/model_doc/distilbert#transformers.DistilBertTokenizer
[15] https://huggingface.co/docs/transformers/en/model_doc/bert#transformers.BertTokenizer
[16] https://huggingface.co/docs/transformers/en/main_classes/trainer#api-reference%20][%20transformers.Trainer

choose the option "no matches" then no retrieved variants were relevant. The annotated data was subsequently used to evaluate and compare the performance of the information retrieval models presented in this work.

4.2.2 Food categorization

Food categorization models were trained on a dataset of 30.036 food items with 300 items which were annotated by the dietitians were preserved for the test set.

The multinomial Logistic Regression classifier was trained using scikit-learn with balanced class weights and a maximum of 5000 iterations. The Linear SVC was also trained with balanced class weights, using up to 5000 iterations. Additionaly, we trained and evaluated the SVC models with polynomial and RBF kernels. To reduce computational cost and avoid overfitting, we fixed the SVC penalty to C=2.0, a value that provided a stable trade-off between margin size and training error in preliminary sweeps.

The feedforward neural network architecture was optimized using Keras Tuner, exploring different hidden layer widths and dropout rates. The final network comprised two fully connected layers with ReLU activations, interleaved with dropout, followed by a softmax output layer. Training was performed with the Adam optimizer, categorical cross-entropy loss, using a batch size of 64 for up to 15 epochs. Early stopping with patience = 3 was applied to prevent overfitting.

Finally, transformer-based models, BERT and DistilBERT, were fine-tuned for the task. BERT was trained for five epochs with a learning rate of 3×10^{-5} and batch size of 32, while DistilBERT was fine-tuned for three epochs with a learning rate of 5×10^{-5} and the same batch size. Both were evaluated with and without class-weighted loss functions to assess the effect of explicit loss re-balancing.

Table 2 demonstrates that polynomial SVC model achieves accuracy of 99.0% and macro-F1 score of 97.7%, closely followed by the RBF SVC with accuracy 98.6%. The kernel SVC models perform the best on the food classification task, despite our assumption regarding the linear separability of data. Among linear models, LinearSVC is the strongest (acc. 96.9%, macro-F1 94.5%), clearly outperforming logistic regression. The small gaps between macro- and weighted-averaged scores for the kernel SVC models indicate better minority-class recall under the class-balanced training strategy, despite class imbalance among the dataset. Transformer-based models underperform compared to SVC models on this dataset. As shown in Table 2, the introduction of class-weighted loss functions for BERT and DistilBERT substantially improved both macro- and weighted-averaged metrics, raising precision, recall, and F1-scores while maintaining or slightly improving accuracy.

Although the taxonomy suggests the categories should be linearly separable, the dataset exhibits label noise. Kernel SVC models outperform the linear baseline on our data. This pattern is consistent with the presence of annotation noise and class imbalance. We assume that on our data, the polynomial kernel can capture pairwise interactions between food item descriptions, and the RBF kernel can form localized decision regions that recover minority classes more effectively than a single linear boundary.

To identify which classes exhibited label inconsistencies, we audited misclassifications and found annotation noise. Fig. 2 shows near-identical *Quetschbeutel* (fruit puree) from the same manufacturer labeled inconsistently as *Smoothies* versus *Getreide-Obst-Breie (cereal–fruit puree)*. Although the taxonomy requires a single category per item, entries were created independently by multiple nutrition experts without secondary validation, leading to label drift. Such inconsistencies can inflate the apparent benefit of models that exploit feature interactions (polynomial SVC), and underscore the need for label harmonization to obtain cleaner estimates of model performance. On the training set, 270 items were misclassified, indicating label noise in the annotations.

	Product 1	**Product 2**
Product	*Quetschbeutel Erdbeere–Apfel*	*Quetschbeutel Apfel–Waldfrucht*
Dataset label	Getreide-Obst-Breie	Smoothies
Model prediction	Smoothies	Smoothies

Figure 2: Labeling inconsistency for the *Quetschbeutel* (fruit puree): *Erdbeere–Apfel* (strawberry–apple) and *Apfel–Waldfrucht* (apple–wild berries).

Table 3: IR performance with different food category predictors. Precision@Retrieved is the average precision among retrieved matches; Success@K is the fraction of queries with at least one correct match among the top-K retrieved.

Categorizer	**Precision@Retrieved**	**Hit Rate (Success@K)**
Baseline A: Text Preprocessing	73.74%	92%
Baseline B: Logistic Regression	78.86%	92%
Linear SVC	80.14%	92%
Polynomial kernel SVC	**80.48%**	92%
RBF kernel SVC	**80.48%**	92%
Neural Network Classifier	77.08%	92%
DistilBERT	80.10%	92%
BERT	80.15%	92%

4.2.3 Information Retrieval

As a first baseline, we implemented an IR system without category filtering, ranking candidate products by cosine similarity of their TF–IDF representations, and their nutritional profiles with at most 20% deviation per nutrient. We then compared this to IR methods augmented with food category filtering, beginning with LR classifier and extending to more advanced predictors. In these settings, retrieval was restricted to candidates within the predicted category. If no matches were found within the predicted category the search was extended to the complete dataset.

System effectiveness was first assessed in terms of *Success@k*, which measures the proportion of queries for which at least one relevant item is retrieved among the top-k results. For computing *Success@1* as shown in (6), the IR system retrieved at least one correct match for 92% of queries:

$$\text{Success@1} = \frac{M}{N} \tag{6}$$

where M is the number of queries with a correct top-ranked match and N the total number of queries.

To compare how our IR approaches retrieve relevant matches, we additionally report *Precision@Retrieved*, defined for a query q as in (7), with $R(q)$ the retrieved set and $Rel(q)$ the set of relevant items.

$$\text{Precision}(q) = \frac{|R(q) \cap Rel(q)|}{|R(q)|} \tag{7}$$

Table 3 shows that augmenting the IR system with a category predictor consistently improves retrieval. The best Precision@Retrieved is achieved by the two kernel SVC models (Polynomial and RBF), tied at **80.48%**, followed by BERT (80.15%), LinearSVC (80.14%), DistilBERT (80.10%) and Logistic Regression (78.86%). The IR combined with neural network achieves only 77.08%. All models yield the same Success@K (92%), indicating that food category primarily reduce the searching space and refine ranking rather than increase coverage of matches. The identical scores for polynomial and RBF kernel-based SVC models suggest their category cues are effectively interchangeable for this benchmark. Overall, incorporating food-category filtering substantially improves retrieval quality over the Baseline A (73.74%).

5 Recipe simulation optimization

The long-term objective of our work is to develop the efficient and automated system for recipe reconstruction based on the information provided on the food packaging. We are aiming to combine the machine learning, mathematical optimization, large and small language models, and expert knowledge for supporting the transparent and effective dietary data assessment.

The first component of the future recipe reconstruction pipeline is the IR-based recipe matching system proposed in our work. This module allows experts to verify whether a newly encountered food product is already represented in the dataset. This prevents the database from adding duplicates. When a product with an identical nutritional profile, food category and ingredient composition is identified, the new item can be directly linked to the existing entry, thereby reusing its nutritional information. Expert knowledge remains central to this process, as dietitians ultimately validate the equivalence of nutritional and ingredient information. This procedure reduces redundancy, streamlines data management, and enhances the overall consistency of the database.

As discussed in Section 2, mathematical optimization is an established method for estimating ingredient proportions. A critical prerequisite for optimization is the alignment of ingredients listed on product packaging with their counterparts in the proprietary NutriDiary database. This alignment step ensures that optimization operates on standardized ingredient representations and reliable nutritional profiles. We propose a hybrid approach in which mathematical optimization remains the central component of recipe simulation, while generative models complement the process.

1. Ingredient parsing using generative modeling

[1] proposed fuzzy matching as a method for mapping ingredients listed on product packaging to entries in food composition databases. However, they emphasized that this approach often fails to identify the correct database entry, as food manufacturers are not obliged to employ standardized

terminology for ingredients.

We propose the use of generative LMs for parsing and normalization of ingredient text, including synonym resolution and lexical standardization. In the work of [21] it is demonstrated that the LLMs can be used for normalization. They proposed a self-improving model which learned to match schema elements via multi-step reasoning, including candidate generation, refinement, and confidence scoring. Recent work such as in [9] has demonstrated the potential of LLMs for nutrition-related reasoning tasks, highlighting the usefulness of CoT-Prompting to improve reliability and interpretability in nutrient estimation. Inspired by these findings, we aim to employ CoT-based strategies to guide ingredient parsing, enabling the model to explicitly reason through synonym resolution and hierarchical ingredient structures.

After normalization step, we aim to retrieve the candidate matches from the food composition database. Since the nutritional data is not highly structured, the best strategy is not to rely completely on the LLM performance. We propose to incorporate other similarity-based methods, for example the k-nearest neighbors algorithm. Similar approach was proposed by [3]. They used kNN to estimate added sugar content based on nutritional similarity from the US Label Insight dataset (70,500 packaged foods). We adopt the idea of identifying nutritionally similar ingredients for retrieving the ingredients from the same food category for further ingredient parsing.

The *1 Ingredient Parsing* module as shown in Fig. 3 is the workflow for aligning product packaging information with standardized entries in the NutriDiary database. The outcome of this step is a normalized product representation that integrates the declared product name and ingredient list with corresponding nutritional profiles from the database. Accounting for the cooking and processing methods is the necessary step for the accurate ingredient estimation. Such optimization step can be integrated into the ingredient–database mapping. As it is demonstrated in Fig. 3, mapping "oats" in oat milk directly to a dedicated "oats-for-oat-milk" entry, which already reflects the increased sugar content due to the fermentation during processing, ensures that the nutrient profile is more accurate from the outset. This approach embeds processing knowledge at the mapping stage and reduces the reliance on post-hoc nutrient-constrained optimization.

2. Nutrient-constrained optimization

Once the ingredients are mapped, their proportions are determined through mathematical optimization aligning the weighted sum of ingredient nutrient values with the declared product profile. As illustrated in module *2 Nutrient-constraint optimization* (Fig. 3) the nutritional information such as nutrient table of products serves as constraints during optimization. By leveraging both the nutrient composition of each ingredient and the overall product profile, it becomes possible to estimate ingredient proportions using linear or quadratic programming approaches [28].

The central objective of this step is to minimize the deviation between the nutrient profile reconstructed from the estimated ingredient proportions and the declared nutrient values on the packaging (commonly referred to as the Big7). Conceptually, each nutrient value of the final product can be expressed as a weighted sum of the nutrient values of its constituent ingredients, where the weights correspond to the unknown ingredient proportions. For example, the fat content of a product

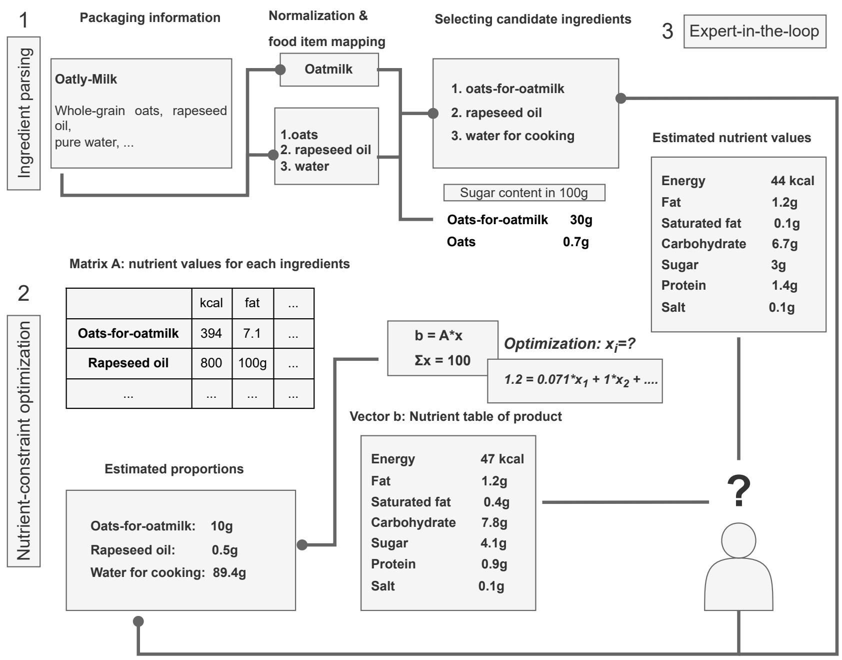

Figure 3: Recipe simulation optimization using LM-based ingredient parsing, mathematical optimization, and expertise of dietitians.

is represented as the sum of fat contributions from each ingredient (per 1 gram), each multiplied by its respective proportion within the product.

This established method, detailed in the works of [28] and [1], enables reliable estimation of ingredient proportions that are not disclosed on packaging and constitutes a key step in bridging ingredient mapping with downstream analyses.

3. Expert-in-the-loop validation

The reconstructed recipes will undergo systematic review by trained dietitians. Expert involvement is essential for refining model outputs and ensuring the long-term improvement of the system. As shown in Fig. 3, the experts evaluate (i) whether the system parses and maps ingredients accurately, (ii) whether food categories and processing methods are correctly assigned, and (iii) whether the nutritional values derived from the reconstructed recipes are consistent with those declared on the packaging. For example, Fig. 3 illustrates a case in which the reconstructed recipe (*Estimated nutrient values*) underestimates the sugar content compared to the declared value (*Nutrient table of product*), even after ingredient adjustments. In such cases, expert judgment is required to determine whether the deviation falls within acceptable limits or whether further refinement of the reconstruction is necessary.

4. Limitations

Large-scale pretraining necessitates access to extensive amounts of high-quality data. In particular, training models to parse food packaging information into standardized ingredient representations requires datasets that ensure reliable alignment between packaging data and proprietary resources.

However, the subset of NutriDiary entries that can be directly matched with packaging information remains limited. To address this constraint, our goal is to expand the nutritional database with systematically collected food packaging data and subsequently align these entries with NutriDiary items, for example through barcodes or product name matching. To mitigate missing or incomplete information, we further aim to incorporate advanced data collection and enrichment techniques, such as imputation methods proposed in the work of [10] and [11].

We argue that advanced fine-tuning strategies are essential for adapting language models to the nutrition domain. Beyond CoT-prompting, we plan to employ efficient fine-tuning techniques such as Low-Rank Adaptation (LoRA) [8], which significantly reduces the number of trainable parameters while maintaining model performance. Another promising approach is Data Whisperer [26], an attention-based method that identifies optimal subsets of training data. Using this approach LLaMA-3-8B has been shown to achieve high performance using only 10% of the available training data.

In addition, knowledge distillation provides a pathway to train compact student models on limited datasets with performance comparable to their larger teachers [23]. This is particularly relevant for our pipeline, where lightweight models are required for efficient inference during recipe reconstruction. As ingredient parsing constitutes a key module of our system, deploying distilled models offers a practical balance between computational efficiency and predictive accuracy.

6 Discussion

Table 2 shows that, among all food categorization models, the SVC models with polynomial and RBF kernels clearly outperform other approaches, including transformer-based models. Among linear models the LinearSVC performed best, outperforming LR classification model. Applying different cost functions to minority and majority classes can improve SVC model performance on unbalanced datasets, which may explain why the LinearSVC outperformed LR model [18]. Introducing the polynomial and RBF kernels further enhanced SVC performance for food categorization. Although each item in the dataset should belong to a single food category, we observe the inconsistency in assigned categories. Example in Fig. 2 illustrates near identical fruit purée that was labeled both as *fruit-puree* and *smoothie*. This annotation noise in the NutriDiary dataset likely stems from manual labeling and dataset expansion performed by the annotators without systematic validation, underscoring the need for label harmonization.

Moreover, the TF–IDF representation, which explicitly encodes correlations between ingredients and food categories and aligns well with the strength of SVC models in handling high-dimensional, sparse features [25]. In contrast, transformer models rely on contextual embeddings from general-domain corpora, which are less effective for datasets dominated by domain-specific lexical terms with limited contextual variation.

Introducing class-weighted loss for neural and transformer models mitigated the bias toward frequent classes, improving both macro-averaged classification metrics and retrieval precision as shown in Table 3. This indicates that reweighting enhances fairness across categories while also benefiting overall performance.

Finally, integrating food category filtering into the IR system further improved retrieval, with *Precision@Retrieved* increasing from 73.74% to over 80%. Restricting the candidate pool to predicted categories reduces noise and ensures greater consistency with domain-specific classification constraints.

Beyond information retrieval, our pipeline lays the groundwork for automated recipe simulation by integrating IR-based recipe matching, LM-driven ingredient parsing, and nutrient-constrained optimization. CoT-Prompting, as demonstrated in NutriBench [9], enables more reliable synonym resolution, hierarchical ingredient handling, and incorporation of processing knowledge directly at the mapping stage.

Full automation, however, remains challenging since LLM-based parsing risks hallucination. To mitigate this issue we will restrict mappings to database-derived candidate sets and untilize human-in-the-loop validation which ensures transparency, reproducibility, and continuous refinement. Overall, a hybrid approach of combining LLMs for structured parsing, mathematical optimization, and expert oversight appears as a scalable and trustworthy solution for dietary data assessment.

The observed inconsistencies in the dataset indicate a need for label harmonization. Discrepant labels should be validated by a second reviewer using written guidelines, and inter-annotator agreement reported. Such harmonization is expected to reduce annotation noise and yield more reliable model performance.

7 Conclusion and Outlook

In this work, we introduced the concept of the data collection optimization project at the Institute of Nutrition and Food Sciences, University of Bonn. First component of this effort is the implementation of the IR-based recommender system designed to enable efficient and accurate searches across the food item database. The system supports nutritional experts by providing relevant candidate matches, while the final decisions regarding data integration remain under expert supervision.

Our experiments demonstrate that enriching the information retrieval process with nutrition-based features improves performance by narrowing the search space and thereby increasing the accuracy of retrieved matches. Among the linear classification models, the SVC with polynomial and RBF kernels achieved the highest performance on the food categorization task. Furthermore, the kernel SVC–based retrieval pipeline yielded the highest overall Precision@Retrieved (80.48%), confirming its effectiveness for food categorization-based recipe matching.

We proposed a hybrid framework for recipe reconstruction that integrates information retrieval, generative language models, mathematical optimization, and expert validation. The pipeline begins with IR-based recipe matching to prevent redundancy in the NutriDiary database and proceeds to recipe simulation when new items cannot be matched directly. Core components include LM-based ingredient parsing and normalization, mathematical optimization for ingredient proportion utilizing nutritional constraints, and systematic expert-in-the-loop validation.

To address current limitations, we outlined strategies for expanding aligned food packaging data, adopting efficient fine-tuning methods such as LoRA and Data Whisperer, and applying knowledge distillation to develop lightweight models suitable for inference. These steps provide a roadmap toward scalable, transparent, and accurate dietary data collection, supporting both nutritional epidemiology and research on sustainable nutritional data collection methods.

In future work, we plan to extend our research toward the practical implementation of recipe reconstruction pipeline by developing an intelligent tool that combines mathematical optimization with LM-based ingredient parsing for ingredient weight estimation. Beyond nutritional accuracy, we envision broadening the scope of dietary data assessment to include sustainability aspects, thereby aligning nutritional data collection with pressing environmental and public health challenges. Moreover, additional data validation is necessary to prevent the discrepancies in category labels.

Acknowledgment

We would like to thank the Institute of Nutritional Epidemiology at the University of Bonn for their support in data validation and for providing expert consultations in the field of nutritional epidemiology. This research has been partially funded by the state of North-Rhine Westphalia as part of the Lamarr Institute for Machine Learning and Artificial Intelligence.

References

[1] Kristin Bohn, Michael Amberg, Toni Meier, Frank Forner, Gabriele I Stangl, and Patrick Mäder. Estimating food ingredient compositions based on mandatory product labeling. *Journal of Food Composition and Analysis*, 110:104508, 2022.

[2] Van Dang, Michael Bendersky, and W Bruce Croft. Two-stage learning to rank for information retrieval. In *European Conference on Information Retrieval*, pages 423–434. Springer, 2013.

[3] Tazman Davies, Jimmy Chun Yu Louie, Rhoda Ndanuko, Sebastiano Barbieri, Oscar Perez-Concha, and Jason HY Wu. A machine learning approach to predict the added-sugar content of packaged foods. *The Journal of Nutrition*, 152(1):343–349, 2022.

[4] Tazman Davies, Jimmy Chun Yu Louie, Tailane Scapin, Simone Pettigrew, Jason HY Wu, Matti Marklund, and Daisy H Coyle. An innovative machine learning approach to predict the dietary fiber content of packaged foods. *Nutrients*, 13(9):3195, 2021.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[6] Charlotte Evenepoel, Egbert Clevers, Lise Deroover, Wendy Van Loo, Christophe Matthys, and Kristin Verbeke. Accuracy of nutrient calculations using the consumer-focused online app myfitnesspal: validation study. *Journal of medical Internet research*, 22(10):e18237, 2020.

[7] Guhao Feng, Bohang Zhang, Yuntian Gu, Haotian Ye, Di He, and Liwei Wang. Towards revealing the mystery behind chain of thought: A theoretical perspective. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 70757–70798. Curran Associates, Inc., 2023.

[8] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021.

[9] Andong Hua, Mehak Preet Dhaliwal, Ryan Burke, Laya Pullela, and Yao Qin. Nutribench: A dataset for evaluating large language models on nutrition estimation from meal descriptions, 2025.

[10] Yuru Huang, Thomas Burgoine, Michael Essman, Dolly RZ Theis, Tom RP Bishop, and Jean Adams. Monitoring the nutrient composition of food prepared out-of-home in the united kingdom: database development and case study. *JMIR Public Health and Surveillance*, 8(9):e39033, 2022.

[11] Gordana Ispirova, Tome Eftimov, and Barbara Koroušić Seljak. Evaluating missing value imputation methods for food composition databases. *Food and Chemical Toxicology*, 141:111368, 2020.

[12] Linda Klasen, Stefanie Anna Julia Koch, Maike Elena Benz, Johanna Conrad, Ute Alexy, Konrad Blaszkiewicz, Ionut Andone, and Ute Nöthlings. Nutridiary, a smartphone-based dietary record app: Description and usability evaluation. *JMIR Hum Factors*, 12:e62776, Feb 2025.

[13] Jimmy Lin. A proposed conceptual framework for a representational approach to information retrieval. In *ACM Sigir Forum*, volume 55, pages 1–29. ACM New York, NY, USA, 2022.

[14] Peihua Ma, Zhikun Zhang, Ying Li, Ning Yu, Jiping Sheng, Hande Küçük McGinty, Qin Wang, and Jaspreet KC Ahuja. Deep learning accurately predicts food categories and nutrients based on ingredient statements. *Food Chemistry*, 391:133243, 2022.

[15] Anand Mariappan, Marc Bosch, Fengqing Zhu, Carol J Boushey, Deborah A Kerr, David S Ebert, and Edward J Delp. Personal dietary assessment using mobile devices. In *Computational imaging VII*, volume 7246, pages 294–305. SPIE, 2009.

[16] Shu Wen Ng, Gregory Bricker, Kuo-ping Li, Emily Ford Yoon, Jiyoung Kang, and Brian Westrich. Estimating added sugars in us consumer packaged goods: an application to beverages in 2007–08. *Journal of Food Composition and Analysis*, 43:7–17, 2015.

[17] Ines Perrar, Ute Alexy, and Ute Nöthlings. Cohort profile update–overview of over 35 years of research in the dortmund nutritional and anthropometric longitudinally designed (donald) study. *European Journal of Nutrition*, 63(3):727–740, 2024.

[18] Salim Rezvani, Farhad Pourpanah, Chee Peng Lim, and QM Wu. Methods for class-imbalanced learning with support vector machines: a review and an empirical evaluation. *arXiv preprint arXiv:2406.03398*, 2024.

[19] Stephen Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. Okapi at trec-3. pages 0–, 01 1994.

[20] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, 2020.

[21] Nabeel Seedat and Mihaela van der Schaar. Matchmaker: Schema matching with self-improving compositional llm programs. 2025.

[22] Wolfgang Sichert-Hellert, Mathilde Kersting, Christa Chahda, Ruth Schäfer, and Anja Kroke. German food composition database for dietary evaluations in children and adolescents. *Journal of Food Composition and Analysis*, 20(1):63–70, 2007.

[23] Inar Timiryasov and Jean-Loup Tastet. Baby llama: knowledge distillation from an ensemble of teachers trained on a small dataset with no performance penalty, 2023.

[24] Christoph Trattner and David Elsweiler. Health recommender systems: Harnessing user and contextual data for health promotion. *User Modeling and User-Adapted Interaction*, 27(3–5):1–66, 2017.

[25] Yasmen Wahba, Nazim Madhavji, and John Steinbacher. A comparison of svm against pre-trained language models (plms) for text classification tasks. In *International Conference on Machine Learning, Optimization, and Data Science*, pages 304–313. Springer, 2022.

[26] Shaobo Wang, Xiangqi Jin, Ziming Wang, Jize Wang, Jiajun Zhang, Kaixin Li, Zichen Wen, Zhong Li, Conghui He, Xuming Hu, and Linfeng Zhang. Data whisperer: Efficient data selection for task-specific llm fine-tuning via few-shot in-context learning, 2025.

[27] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.

[28] Brian J Westrich, I Marilyn Buzzard, Laël C Gatewood, and Paul G McGovern. Accuracy and efficiency of estimating nutrient values in commercial food products using mathematical optimization. *Journal of Food Composition and Analysis*, 7(4):223–239, 1994.

[29] Lei Zhou, Chu Zhang, Fei Liu, Zhengjun Qiu, and Yong He. Application of deep learning in food: a review. *Comprehensive reviews in food science and food safety*, 18(6):1793–1811, 2019.