International Journal of Food Science and Technology, 2026, 61(2), vvag142
https://doi.org/10.1093/ijfood/vvag142
Published: 15 July 2026
Original Article

Can machines detect ultra-processed foods? A head-to-head evaluation of large language models using NOVA classification

Hatice Merve Bayram [1,*], Sedat Arslan [2], and Arda Ozturkcan [1]

[1]Department of Nutrition and Dietetics, Faculty of Health Sciences, Istanbul Gelisim University, Istanbul, Turkiye
[2]Department of Nutrition and Dietetics, Faculty of Health Sciences, Bursa Uludag University, Bursa, Turkiye

*Corresponding author: Cihangir Mah., Şehit Komando Er Hakan Öner Street, No: 1, Avcilar, 34310, Turkiye. Email: merve.bayrm@gmail.com

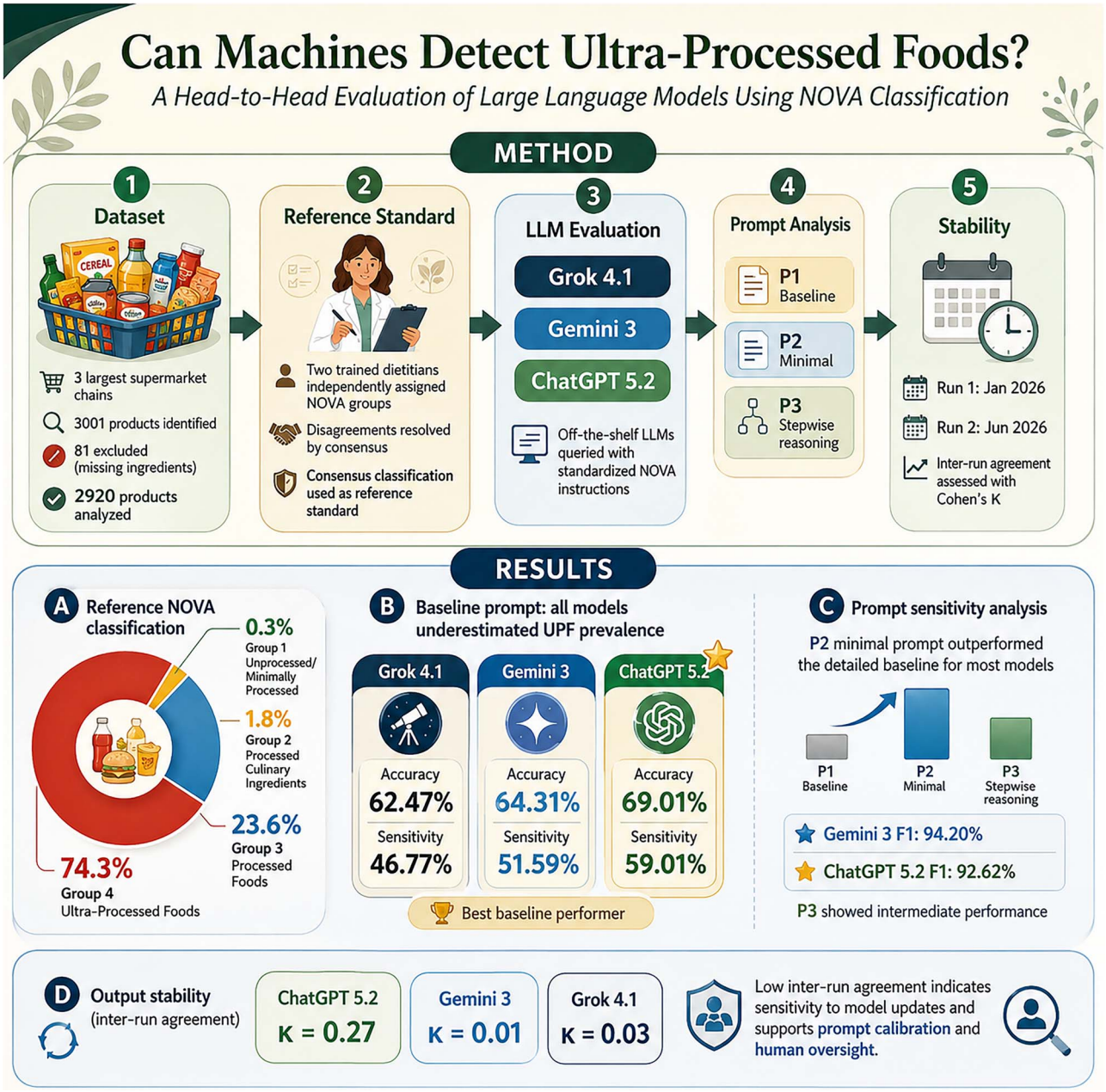
Abstract

This cross-sectional study compared three large language models (LLMs) (Grok 4.1, Gemini 3, and ChatGPT 5.2) in classifying ultra-processed foods (UPF) using best-selling products from leading supermarket chains covering 53.2% of the national market. Of 3,001 products, 2,920 with complete ingredient information were included; two trained dietitians assigned NOVA groups as the reference standard. In the reference classification, 74.3% of products were UPF. Under the baseline prompt, all models underestimated UPF prevalence compared with the reference standard ($p < .001$). ChatGPT 5.2 yielded the highest binary UPF detection performance (accuracy: 69.01%; sensitivity: 59.01%; specificity: 98.00%; F1: 73.88%). Prompt sensitivity analyses revealed that a minimal prompt substantially outperformed the detailed baseline for most models (Gemini 3 F1: 94.20%; ChatGPT 5.2 F1: 92.62%), though Grok 4.1's gain reflected a specificity trade-off (sensitivity: 97.79%; specificity: 21.09%). Low inter-run agreement (κ: 0.01–0.27) indicated sensitivity to model updates, supporting prompt calibration and human oversight. These findings suggest that off-the-shelf LLMs require prompt calibration and human oversight before UPF surveillance workflows.

Keywords NOVA classification, ultra-processed foods, large language models, food labelling, supermarket foods

Received: 28 March 2026. **Accepted:** 6 July 2026

Graphical abstract



Introduction

Ultra-processed foods (UPF) have become a defining feature of contemporary food systems and are increasingly examined as a driver of diet-related non-communicable diseases and as a policy-relevant marker of food environments. A widely used framework for identifying UPF is the NOVA classification, which categorizes foods according to the nature, extent, and purpose of industrial processing rather than nutrient content alone (Monteiro et al., 2010). Within NOVA, UPF (Group 4) are typically industrial formulations made largely from substances derived from foods and additives, often designed to intensify palatability, convenience, and durability; practical identification therefore frequently relies on ingredient-list interpretation and the detection of processing markers and cosmetic additives (Monteiro et al., 2019).

The public health relevance of UPF is supported by converging evidence from experimental and epidemiological research. Consumption of a diet rich in UPF has been reported to increase *ad libitum* energy intake and lead to weight gain compared to an unprocessed diet under strict control conditions (Hall et al., 2019). At the evidence-synthesis level, an umbrella review of meta-analyses reported that greater UPF exposure is associated with higher risks across multiple adverse health outcomes—particularly cardiometabolic outcomes, common mental

disorders, and mortality endpoints—while also highlighting that evidence quality varies across outcomes (Lane et al., 2024; Monteiro et al., 2025). These findings have increased demand for scalable and reproducible methods to quantify UPF within national food supplies and retail inventories.

Despite its broad uptake, implementing NOVA at scale is operationally challenging. Product-level NOVA classification requires detailed and up-to-date ingredient information, and manual assignment can be time-consuming, particularly in dynamic retail settings where products are frequently reformulated, and labelling conventions vary (Monteiro et al., 2019). Additionally, NOVA is subject to ongoing scientific debate regarding conceptual boundaries and interpretation from a food science perspective, underscoring the importance of transparent, replicable classification procedures (Petrus et al., 2021). These realities motivate methodological innovation to reduce the burden of classification while preserving alignment with expert judgement.

Recent advances in natural language processing and transformer-based models provide a plausible route to automate processing-level classification using ingredient-list text. For example, Hu et al. demonstrated that a fine-tuned language model can accelerate NOVA classification across multiple national food databases, supporting feasibility for large-scale applications (Hu et al., 2023). In parallel, general-purpose large language models (LLMs) offer a potentially more flexible approach: without task-specific training, they can follow standardized instructions and process long ingredient lists rapidly. However, LLM outputs may be sensitive to prompt framing and may exhibit systematic misclassification; therefore, head-to-head benchmarking against a human-coded reference standard is essential before LLMs are used for surveillance or etiologic research. More broadly, recent nutrition-focused commentaries have highlighted both the promise of general-purpose LLMs for personalized nutrition support and the need for careful validation and methodological safeguards before clinical or research deployment (Arslan, 2023; Arslan, 2024; Bayram & Ozturkcan, 2025).

Country-specific validation is particularly important because labelling language, ingredient terminology, and the composition of the packaged food supply differ across contexts. In Türkiye, recent studies have examined UPF in chain-market settings and reported associations between NOVA-based processed food consumption and cardiometabolic risk among adults (Ozkan et al., 2025; Şimşek et al., 2025). Complementing this evidence, Turkish observational studies have also linked higher UPF consumption with food insecurity among adults with overweight/obesity, greater menopausal symptom burden in postmenopausal women, higher premenstrual syndrome symptom severity, and greater social media addiction—highlighting the relevance of UPF exposure across socioeconomic vulnerability, women's health, and digital-behavioural determinants (Arslan et al., 2025; Bekar et al., 2025; Durmaz et al., 2025; Karabiyik et al., 2025). Nevertheless, evidence remains limited on whether off-the-shelf LLMs can reproduce expert NOVA decisions at scale within a contemporary national supermarket inventory.

Moreover, the Lancet's current series of reports identifies UPF consumption as a global public health issue. The globalized food industry is normalizing UPF consumption through aggressive marketing and policy strategies, particularly in low- and middle-income countries, while weakening regulatory interventions. Therefore, the transition to low-UPF diets must be addressed through comprehensive policies that support fair, inclusive, and sustainable food systems, while prioritizing food security and gender equality (Baker et al., 2025; Monteiro et al., 2025; Scrinis et al., 2025). This study aimed to evaluate three LLMs against a two-dietitian NOVA reference standard using a large supermarket product dataset. By quantifying differences in NOVA group distributions and evaluating binary UPF detection performance across multiple prompt formulations and repeated query conditions, this work aims to clarify the feasibility and limitations of LLM-assisted NOVA classification for scalable UPF monitoring, assess output stability across runs, and delineate where calibration and human oversight remain necessary.

Materials and methods

Study design and data collection

This study was a cross-sectional comparative study assessing three artificial intelligence (AI) models (Grok 4.1, Gemini 3, ChatGPT 5.2) in identifying UPF from a food list containing the best-selling products in supermarkets. For this study, information on the contents of packaged food and beverage products sold in the three largest supermarkets, representing 53.2% of the Turkish market share, was collected. The food products sold in these stores are similar to those sold in other supermarket chains across the country.

Data were gathered between November 2025 and January 2026. A total of 3,001 products were identified. 81 products were excluded due to missing ingredient information. A total of 2,920 products were included in the analysis. Information on identification (product name and type) and ingredients was obtained in the stores from labels of all of the food and beverage products. Then, their information (nutritional label and ingredients) was manually entered into Microsoft Office Excel 2016.

Food categorization according to the NOVA classification

The NOVA classification system is a framework used to categorize foods based on the nature, extent, and purpose of industrial processing they undergo before consumption (Monteiro, 2009). The NOVA system classifies foods into four categories: (i) unprocessed or minimally processed foods, (ii) processed culinary ingredients, (iii) processed foods, and (iv) UPF (Monteiro, 2009; Monteiro et al., 2018; Monteiro et al., 2019).

Two trained dietitians independently classified each food item according to standard NOVA criteria and prespecified decision rules. Disagreements between the two dietitians were discussed and resolved by consensus, and the final consensus classification was used as the reference standard for model evaluation. This manual classification was used as the reference (or "gold") standard to assess the reliability of the classifications in this study.

Assessment of AI model performance

All AI models were queried using the same standardized prompt, which was entered on 25 January 2026, to ensure temporal consistency.

Prior to the prompt, information about the NOVA classification was provided to each AI model as below:

Group 1. Unprocessed and minimally processed foods: Unprocessed or Natural foods are obtained directly from plants or animals and do not undergo any alteration following their removal from nature. Minimally processed foods are natural foods that have been submitted

to cleaning, removal of inedible or unwanted parts, fractioning, grinding, drying, fermentation, pasteurization, cooling, freezing, or other processes that may subtract part of the food, but which do not add oils, fats, sugar, salt, or other substances to the original food. Examples include fresh, frozen, or chilled vegetables, fresh or dried legumes (lentils, chickpeas, dry beans, and peas), whole grains and cereals (rice, corn, bulgur, wheat, etc.), fresh, chilled, or frozen meat, poultry, and fish (raw chicken breast, fresh fish fillets, lean beef cuts, etc.), nuts and seeds without added salt, sugar, or oils (almonds, walnuts, hazelnuts, peanuts, sunflower seeds, etc.), fresh or pasteurized milk and plain yogurt, fresh or dried herbs and spices, tea, coffee, and water.

Group 2. Processed culinary ingredients: These foods are products obtained from Group 1 foods or nature through processes such as pressing, grinding, crushing, pulverizing, and refining. They are used in homes and restaurants to season and cook food, creating all kinds of delicious meals and dishes, including soups, salads, pies, breads, cakes, desserts, and preserves. Examples include vegetable oils extracted from seeds or fruits (olive oil, sunflower oil, canola oil, soybean oil, etc.), butter and other animal fats, sugar, honey, maple syrup, and similar natural sweeteners, salt, starch extracted.

Group 3. Processed foods: These foods are Group 1 foods combined with salt, sugar, fat, or other Group 2 components using preservation methods such as canning, bottling, or non-alcoholic fermentation. These processes aim to extend shelf life and improve sensory characteristics. These products typically contain a limited number of ingredients and may be subject to methods such as canning, bottling, or fermentation. Examples include canned or bottled legumes, vegetables, and fish, salted, dried, smoked, or cured meat or fish, fruits in syrup, fresh breads, and cheeses, salted or sugared nuts.

Group 4. Ultra-processed foods: UPF are industrial formulations of substances derived from foods. These foods typically contain additives for flavour, colour, texture, and shelf life. Production involves numerous industrial processes such as fractionation, chemical modification, extrusion, moulding, frying, and sophisticated packaging. Ingredients commonly include sugars, refined oils, salt combinations, high-fructose corn syrup, hydrogenated or esterified oils, protein isolates, and cosmetic additives such as flavourings, colourings, emulsifiers, sweeteners, thickeners, anti-foaming agents, and gelling agents. Examples include fatty, sweet, savoury or salty packaged snacks, biscuits (cookies), ice creams and frozen desserts, chocolates, candies, cola, soda and other carbonated soft drinks, sports drinks, instant soups, noodles, sauces, desserts, drink mixes and seasonings, sweetened and flavoured yogurts, dairy drinks, sweetened juices, pre-prepared (packaged) meat, fish and vegetables, pre-prepared pizza, pasta dishes, burgers, hot dogs, sausages, poultry and fish, other animal products made from remnants, packaged breads, breakfast cereals and bars.

Then the prompt was asked, and it was in English and it was as follows:

"Now, classify the foods in the file I am providing you and present the classification every 100 items." To assess output stability, each model was queried independently on two separate occasions using the original prompt (P1): the first query was conducted in January 2026 (Run 1) and the second query was conducted in June 2026 (Run 2). As model versions may have been updated between query dates, version information was documented for both runs. For each product, identical NOVA group assignments across the two runs were considered stable outputs. In cases of disagreement between runs, the Run 1 classification was retained for the primary performance analysis, consistent with the original study design. Inter-run agreement was quantified using Cohen's kappa.

No model training, fine-tuning, or parameter optimization was performed in this study. The evaluated systems were used as off-the-shelf LLMs, and their outputs were compared with the dietitian-coded NOVA reference classification. The methodological focus was to assess how currently available general-purpose LLMs perform when given identical product information and standardized NOVA instructions. This approach allowed the models to be evaluated under conditions that approximate real-world use by researchers or practitioners who may rely on LLMs without task-specific retraining.

Prompt sensitivity analysis

To evaluate whether classification performance varied with prompt design, each model was additionally queried using two alternative prompt formulations alongside P1. The minimal prompt (P2) provided only NOVA group names and a brief definition without examples: "Classify each food into one of four NOVA groups: (1) Unprocessed/minimally processed, (2) Processed culinary ingredients, (3) Processed foods, (4) Ultra-processed foods. Apply standard NOVA criteria".

The stepwise reasoning prompt (P3) instructed the model to reason explicitly before assigning a category: "For each food item, follow these steps: (1) List the ingredients that indicate industrial processing. (2) Identify any cosmetic additives, flavourings, emulsifiers, or substances of rare culinary use. (3) Based on these, assign the most appropriate NOVA group (1–4)".

All three prompts were applied to the same dataset. Performance metrics (accuracy, sensitivity, specificity, precision, and F1-score) were calculated for each model–prompt combination against the dietitian-coded reference standard.

Statistical analysis

Data was analysed using SPSS 24.0 and Microsoft Excel. Differences in UPF content across AI models and reference were examined using Chi-square test. Binary UPF/non-UPF agreement between the reference and each model was assessed using McNemar's test with continuity correction, as this accounts for the paired structure of the data. Statistical significance was set at a p-value $\leq .05$.

Results

Table 1 presents the distribution of 2,920 food items across NOVA classification groups according to the reference classification and each AI model. Based on the reference, the majority of food items were classified as UPF (Group 4, 74.3%), followed by processed foods (Group 3, 23.6%), while unprocessed or minimally processed foods (Group 1) and processed culinary ingredients (Group 2) accounted for a very small proportion of the dataset (0.3% and 1.8%, respectively). These values were 44.4%, 21.7%, 32.2, and 1.7% for ChatGPT 5.2. However, Grok 4.1 and Gemini 3 markedly overclassified foods as Group 1. The differences between each AI model and the reference classification were statistically significant ($p < .001$ for all).

Table 2 presents the distribution of food items across NOVA classification groups according to the reference classification and each AI model for Run 2. Compared with Run 1, ChatGPT 5.2 classified a higher proportion of items as UPF (70.0%), while Grok 4.1 and Gemini 3 also showed altered classification distributions. The differences between each AI model and the reference classification remained statistically significant ($p < .001$ for Grok 4.1, ChatGPT 5.2; and Gemini 3).

Table 1 Distribution of food items across NOVA classification according to each AI model and reference for Run 1.

NOVA classification	Reference (*n*, %)	AI models		
		Grok 4.1 (*n*, %)	Gemini 3 (*n*, %)	ChatGPT 5.2 (*n*, %)
Group 1	10 (0.3)	1659 (56.8)	822 (28.2)	941 (32.2)
Group 2	51 (1.8)	2 (0.1)	114 (3.8)	50 (1.7)
Group 3	689 (23.6)	32 (1.1)	621 (21.3)	633 (21.7)
Group 4	2170 (74.3)	1227 (42.0)	1363 (46.7)	1296 (44.4)
McNemar χ^2		*637.91*	*434.33*	*842.06*
p-value		*<.001**	*<.001**	*<.001**

*$p < .001$. *Note.* For statistical comparison, NOVA groups were dichotomized as UPF *vs.* non-UPF, and McNemar's test with continuity correction was applied.

Table 2 Distribution of food items across NOVA classification according to each AI model and reference for Run 2.

NOVA classification	Reference (*n*, %)	AI models		
		Grok 4.1 (*n*, %)	Gemini 3 (*n*, %)	ChatGPT 5.2 (*n*, %)
Group 1	10 (0.3)	148 (5.1)	336 (11.5)	330 (11.3)
Group 2	51 (1.8)	759 (26.0)	0 (0.0)	0 (0.0)
Group 3	689 (23.6)	2 (0.1)	1073 (36.7)	545 (18.7)
Group 4	2170 (74.3)	2011 (68.9)	1511 (51.7)	2045 (70.0)
McNemar χ^2		*29.26*	*640.53*	*53.95*
p-value		*<.001**	*<.001**	*<.001**

*$p < .001$. *Note.* For statistical comparison, NOVA groups were dichotomized as UPF *vs.* non-UPF, and McNemar's test with continuity correction was applied.

Table 3 Output stability metrics across two independent query runs.

AI model	Agreement rate (%)	Cohen's κ	*p*-value
Grok 4.1	33.8	0.03	<.001**
Gemini 3	36.1	0.01	.002*
ChatGPT 5.2	55.2	0.27	<.001**

*$p < .05$. **$p < .001$. *Note.* Run 1: January 2026; Run 2: June 2026.

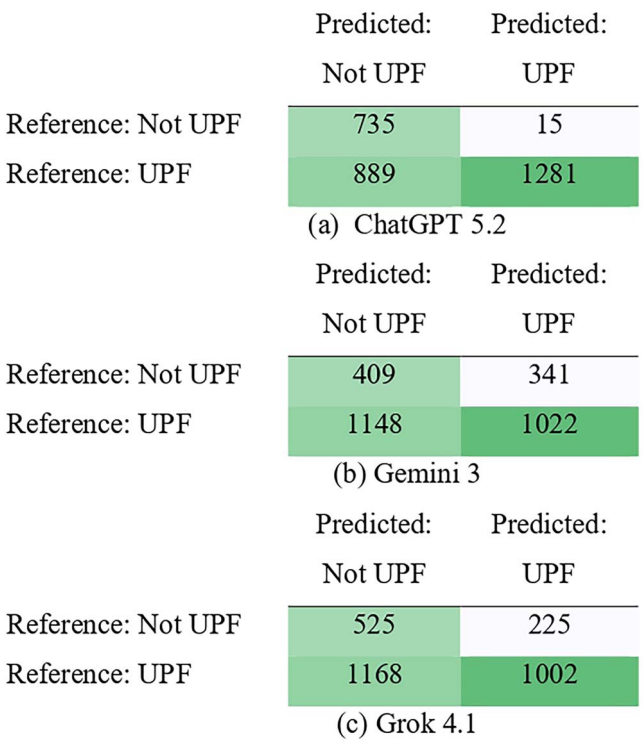


Figure 1 Comparison of AI model predictions with reference UPF classification for run 1 (UPF vs. non-UPF). UPF = ultra-processed foods.

Figure 1 shows the comparison of AI model predictions with reference UPF classification for Run 1. The observed classification pattern indicated that ChatGPT 5.2 had the closest agreement with the reference classification among the evaluated models, correctly identifying a larger proportion of UPF items and showing fewer false-positive classifications than Gemini 3 and Grok 4.1. Gemini 3 and Grok 4.1 showed more frequent misclassification, particularly by labelling UPF items as non-UPF.

Figure 2 presents the comparison of AI model predictions with reference UPF classification for Run 2. Compared to Run 1, ChatGPT 5.2 and Grok 4.1 showed increased true positives alongside higher false positives. However, Gemini 3 demonstrated markedly improved performance in Run 2, with both higher specificity and higher sensitivity, indicating a general improvement rather than a trade-off.

Table 3 shows the output stability metrics across two independent query runs. The agreement rates between Run 1 and Run 2 were 55.2% for ChatGPT 5.2, 36.1% for Gemini 3, and 33.8% for Grok 4.1. Cohen's kappa indicated weak agreement for ChatGPT 5.2 ($\kappa = 0.27$) and near-chance agreement for Gemini 3 ($\kappa = 0.01$) and Grok 4.1 ($\kappa = 0.03$).

Table 4 presents the performance metrics by prompt formulation across AI models. Under P1, ChatGPT 5.2 yielded the highest observed accuracy (69.01%), sensitivity (59.01%), specificity (98.00%), and precision (98.84%). In comparison, Grok 4.1 and Gemini 3 demonstrated

Table 4 Performance metrics by prompt formulation across AI models.

AI model	Prompt	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	F1 score (%)
Grok 4.1	P1 (baseline)	52.26	46.15	69.96	81.66	58.95
	P2 (Minimal)	78.11	97.79	21.09	78.22	86.91
	P3 (chain-of-thought)	77.60	76.08	81.98	92.44	83.47
Gemini 3	P1 (baseline)	48.90	47.08	54.21	74.87	57.76
	P2 (Minimal)	91.64	91.34	92.52	97.25	94.20
	P3 (chain-of-thought)	88.90	87.74	92.26	97.04	92.16
ChatGPT 5.2	P1 (baseline)	69.01	59.01	98.00	98.84	73.88
	P2 (Minimal)	89.44	89.16	90.25	96.36	92.62
	P3 (chain-of-thought)	84.17	81.47	91.99	96.72	88.44

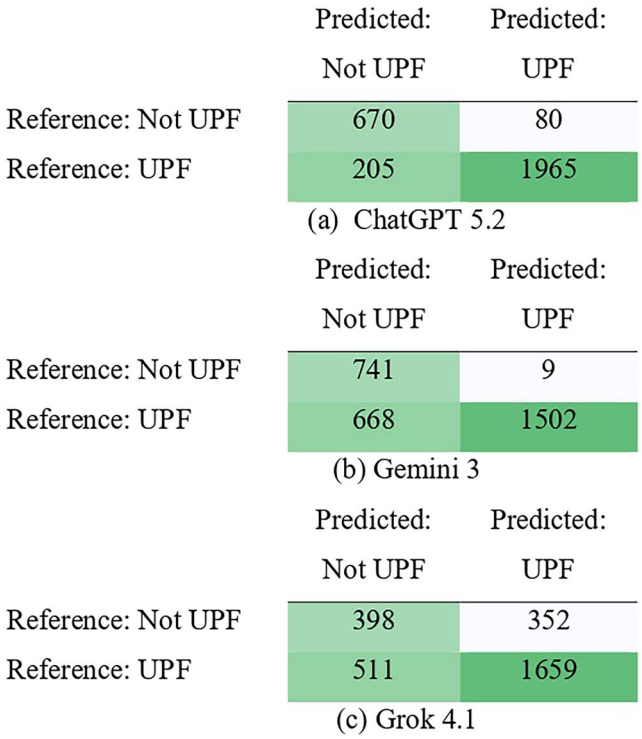


Figure 2 Comparison of AI model predictions with reference UPF classification for run 2 (UPF vs. non-UPF). UPF = ultra-processed foods.

lower accuracy and sensitivity, with moderate specificity and precision. Under P2, Gemini 3 yielded the highest overall performance (accuracy: 91.64%, sensitivity: 91.34%, F1: 94.20%), followed by ChatGPT 5.2 (accuracy: 89.44%, sensitivity: 89.16%, F1: 92.62%). Under P3, performance improved substantially relative to P1 for all models, though remained slightly below P2 for ChatGPT 5.2 (accuracy: 84.17%, sensitivity: 81.47%, F1: 88.44%) and Gemini 3 (accuracy: 88.90%, sensitivity: 87.74%, F1: 92.16%).

Discussion

In this study, we evaluated the ability of three LLMs to classify packaged foods into NOVA groups using ingredient lists, benchmarking their outputs against expert consensus classifications. The findings should be interpreted as indicative patterns observed within the present supermarket product dataset rather than as definitive evidence of generalizable model performance. Across a dataset of 2,920 products, UPF (NOVA 4) constituted the majority of items, while unprocessed/minimally processed foods (NOVA 1) and processed foods (NOVA 3) represented smaller shares. Under the baseline prompt (P1), model performance was generally characterized by higher precision and specificity than sensitivity, with ChatGPT 5.2 showing particularly high precision and specificity. This pattern indicates that UPF predictions were often correct when made, especially by ChatGPT 5.2, although a substantial proportion of true UPF items remained undetected. However, this performance profile did not persist across all alternative prompt formulations. Notably, prompt formulation substantially modulated this performance: simpler prompt designs yielded markedly higher sensitivity and overall accuracy than the detailed baseline prompt for most models, while the chain-of-thought approach produced intermediate results, underscoring that LLM classification behaviour is sensitive not only to model architecture but also to how the task is framed. Misclassification patterns were not uniform across models: the distribution of predicted NOVA categories differed significantly between the three systems, suggesting model-specific decision heuristics rather than a stable "processing" concept consistently applied across LLMs. This interpretation is further supported by the output stability analysis: when each model was queried a second time 5 months later using the identical prompt, agreement rates were low across all three systems, and NOVA group distributions differed significantly between runs for all models, suggesting that intervening model version updates substantially altered classification tendencies within the same prompt framework.

The output variability observed across query occasions in the present study aligns with a broader concern in the clinical AI literature. Since LLMs generate responses by probabilistically selecting tokens, the same prompt can produce different outputs when submitted at different times (Bayram & Ozturkcan, 2025; Shyr et al., 2025). It was reported that accuracy and consistency are not necessarily aligned; a model may generate a correct classification on one occasion while failing to reproduce the same result in subsequent runs (Shyr et al., 2025). For NOVA-based UPF surveillance, this issue is particularly relevant because reproducibility is essential for accurate exposure assessment and meaningful comparisons across studies. Variability in model outputs, together with the possibility of unannounced updates to underlying model versions, may limit the reliability of LLMs for routine monitoring purposes. To support their use in research and surveillance settings, measures such as version-controlled model access, systematic output archiving, and regular validation against expert-coded reference datasets will likely be necessary.

These findings matter because NOVA-based UPF exposure is increasingly used in nutritional epidemiology and policy discussions,

and misclassification can bias both surveillance and etiologic inference. NOVA was originally proposed to capture the extent and purpose of processing rather than nutrient composition alone, emphasizing industrial formulations, cosmetic additives, and substances of rare culinary use as distinguishing features of UPF (Monteiro et al., 2010; Monteiro et al., 2019). Moreover, the expanding evidence suggests a link between high UPF consumption in observational syntheses and adverse cardiometabolic and mental health outcomes (Lane et al., 2024; Monteiro et al., 2025), and experimental data demonstrate that *ad libitum* diets that are rich in UPF can increase energy intake and promote weight gain even when matched for presented macronutrients and calories (Hall et al., 2019). In this context, tools that systematically underestimate UPF prevalence, particularly by missing NOVA 4 items, could attenuate associations (bias towards the null) in exposure–outcome models and could understate the population-level UPF burden used to motivate interventions.

The key performance signature we observed under the baseline prompt, particularly for ChatGPT 5.2, suggests that general-purpose LLMs may behave more like conservative "UPF flaggers" than comprehensive UPF detectors under certain prompting conditions. One plausible explanation is that ingredient-list cues that are salient and unambiguous to an LLM (e.g., explicit emulsifiers, flavour enhancers, intense sweeteners, modified starches, protein isolates, or long multi-component additive strings) trigger NOVA 4 labelling, whereas UPF with shorter or more "culinary-sounding" lists (yet still industrial formulations) may be interpreted as less processed. This aligns with methodological proposals emphasizing discriminative ingredient markers, especially cosmetic additives and substances of rare culinary use, as reproducible identifiers of UPF when ingredient data are available (Grilo et al., 2025). It also resonates with the broader methodological push to increase transparency and standardization when applying NOVA, including explicit decision rules for ambiguous mixed items and careful documentation of classification procedures (Martinez-Steele et al., 2023).

Our results are consistent with the well-documented challenge that NOVA assignment is not always straightforward and can show variable agreement across coders and datasets. Studies examining functionality and robustness of NOVA have reported that consistency can be moderate in some settings and improves with structured training and prespecified rules, particularly for mixed dishes and borderline cases (Astrup & Monteiro, 2022). Therefore, part of the sensitivity gap we observed may reflect genuine ambiguity in some products, but the magnitude and systematic direction of errors suggest an additional model-related limitation: LLMs infer "processing" through linguistic and semantic proxies rather than through an explicit, validated processing ontology. This limitation becomes more salient when the task requires distinguishing NOVA 3 vs. NOVA 4, often the most policy-relevant boundary, where industrial formulation cues can be subtle.

Importantly, our findings should not be interpreted as evidence that LLMs cannot contribute to NOVA classification workflows. Instead, they indicate that out-of-the-box LLMs may be best positioned as components of hybrid pipelines: e.g., (i) high-precision triage to prioritize manual review of likely UPF, combined with (ii) rule-based screens for additive and "rare culinary use" markers and (iii) human adjudication for ambiguous products. Similarly, domain-specific machine learning approaches trained explicitly on labelled food databases have shown promise for quantifying processing in more continuous and reproducible ways. For instance, machine learning models have been proposed to predict degree of processing at scale and link processing metrics to health risk indicators (Menichetti et al., 2023). Additionally, transformer-based natural language processing methods have been explored to accelerate NOVA classification in national food databases, reflecting a broader trend towards task-specific modelling rather than relying on general-purpose generation models.

The prompt sensitivity analysis showed that prompt wording had a clear influence on classification performance. Simple prompts performed better than the more detailed baseline prompt in most models. Contrary to our initial expectation, the minimal prompt (P2), which included only the NOVA group names and a brief definition without examples, produced the highest accuracy and F1-scores for ChatGPT 5.2 and Gemini 3. The stepwise reasoning prompt (P3) resulted in intermediate performance. Our findings suggest that providing extensive instructions may not always improve LLM-based classification. Instead, highly detailed prompts may introduce unnecessary complexity or competing cues, which can reduce consistency in decision-making. Grok 4.1 showed a different pattern under P2. Although it achieved very high sensitivity, its specificity was markedly low, indicating a tendency to classify a large proportion of products as UPF rather than applying more discriminative criteria.

From a practical perspective, the model differences we observed have direct implications for researchers and policymakers considering automated UPF classification. When the goal is to avoid false positives (e.g., enforcement contexts where mislabelling a non-UPF as UPF has reputational or regulatory consequences), high precision/specificity is advantageous. However, when the goal is population surveillance or exposure quantification for epidemiologic associations, low sensitivity may be unacceptable because it systematically undercounts UPF exposure. In these use cases, conservative classification could lead to misestimation of UPF prevalence and potentially diluted exposure–outcome relationships. Therefore, automated systems should be evaluated not only with overall accuracy but also with sensitivity, calibration across NOVA groups, and error decomposition by product type (e.g., breads, dairy analogues, ready meals, confectionery), especially around the NOVA 3–4 boundary.

This study has several limitations. First, classification relied on ingredient lists and product metadata available in the dataset; any omissions, non-standard naming, or language-specific ingredient conventions could have affected both expert coding and model predictions. Second, while three prompt formulations were evaluated in this study, the optimal prompting strategy may vary across product categories, languages, and model versions; few-shot examples and retrieval-augmented approaches were not tested and may further improve performance. Third, while we benchmarked against expert consensus, we did not independently quantify uncertainty for borderline items with a formal adjudication protocol beyond consensus resolution; future work could incorporate structured arbitration and report interrater reliability alongside model metrics. Finally, our dataset reflects a specific product ecosystem; transferability to other markets and labelling regimes should be tested explicitly. These findings should therefore be interpreted as indicative rather than definitive evidence of LLM performance in NOVA classification, particularly because the models were not fine-tuned, calibrated, or externally validated for this task.

Despite these limitations, the study provides a clear, actionable message: under the baseline prompt, particularly for ChatGPT 5.2, UPF classifications showed high precision, although many true UPF items remained undetected. This pattern varied substantially across models and prompt formulations, and they differ meaningfully from

one another in category tendencies. Future research should test hybrid ingredient-marker + ML pipelines, evaluate cross-country generalizability, and assess whether fine-tuning on high-quality labelled datasets can improve sensitivity without sacrificing the strong specificity observed here. Given the accelerating policy relevance of UPF classification, establishing transparent, reproducible, and scalable methods, while acknowledging ongoing debate around NOVA's conceptual boundaries, remains a priority (Martinez-Steele et al., 2023; Petrus et al., 2021).

Conclusion

In a large, real-world supermarket product inventory, dietitian-led NOVA coding indicated a high prevalence of UPF. When benchmarked against this reference standard, all evaluated LLMs showed systematic deviations in NOVA group assignment and tended to underestimate UPF prevalence, with ChatGPT 5.2 demonstrating the strongest binary UPF detection performance. Prompt sensitivity analyses showed that simpler formulations substantially outperformed the detailed baseline prompt for most models, and output stability analyses revealed low inter-run agreement across all three systems, collectively indicating that LLM classification behaviour is highly sensitive to both prompt design and model version.

These findings indicate that off-the-shelf LLMs are not yet suitable as stand-alone tools for estimating UPF prevalence or replacing expert NOVA classification in research and surveillance. However, they may be valuable as components of a human-in-the-loop workflow (e.g., high-confidence UPF flagging and triage) alongside standardized decision rules and periodic calibration against expert-labelled samples. Concretely, a lightweight post-hoc classifier trained on dietitian-coded products could correct systematic underclassification at the NOVA 3–4 boundary, while a tiered human review protocol, routing discordant cases to expert adjudication and auditing a random sample periodically, would address output instability. Future work should evaluate prompt strategies, category-specific errors, and locally adapted rule/lexicon support to improve sensitivity while maintaining the strong specificity required for reliable large-scale UPF monitoring.

Supplementary material

Supplementary material is available at *International Journal of Food Science and Technology* online.

Data availability

The data underlying this article will be shared on reasonable request to the corresponding author. Formun ÜstüFormun Altı.

Author contributions

Hatice Merve Bayram (Conceptualization [equal], Formal analysis [lead], Investigation [lead], Methodology [equal], Resources [equal], Software [equal], Writing—original draft [equal], Writing—review & editing [equal], Data curation, Investigation), Sedat Arslan (Conceptualization [equal], Methodology [equal], Writing—original draft [equal], Writing—review & editing [equal], Formal analysis), and Arda Ozturkcan (Conceptualization [equal], Methodology [equal], Writing—original draft [equal], Writing—review & editing [equal])

Funding

None declared.

Conflicts of interest

None declared.

Acknowledgements

None declared.

Ethics

No ethical approval was required for this study as it involved no human participants. All data were derived from publicly available food product labels and the outputs of publicly accessible large language model interfaces.

References

Arslan, S. (2023). Exploring the potential of chat GPT in personalized obesity treatment. *Annals of Biomedical Engineering*, *51*, 1887–1888. https://doi.org/10.1007/s10439-023-03227-9

Arslan, S. (2024). Decoding dietary myths: The role of ChatGPT in modern nutrition. *Clinical Nutrition ESPEN*, *60*, 285–288. https://doi.org/10.1016/j.clnesp.2024.02.022

Arslan, S., Dal, N., Atan, R. M., Sahin, K., & Tari Selcuk, K. (2025). Relationship between food insecurity and ultra-processed food consumption in adults with overweight and obesity. *Clinical Nutrition ESPEN*, *70*, 660–665. https://doi.org/10.1016/j.clnesp.2025.11.021

Astrup, A., & Monteiro, C. A. (2022). Does the concept of 'ultra-processed foods' help inform dietary guidelines, beyond conventional classification systems? NO. *The American Journal of Clinical Nutrition*, *116*, 1482–1488. https://doi.org/10.1093/ajcn/nqac123

Baker, P., Slater, S., White, M., Wood, B., Contreras, A., Corvalán, C., Gupta, A., Hofman, K., Kruger, P., Laar, A., Lawrence, M., Mafuyeka, M., Mialon, M., Monteiro, C. A., Nanema, S., Phulkerd, S., Popkin, B. M., Serodio, P., Shats, K. *et al.* (2025). Towards unified global action on ultra-processed foods: Understanding commercial determinants, countering corporate power, and mobilising a public health response. *The Lancet*, *406*, 2703–2726. https://doi.org/10.1016/S0140-6736(25)01567-3

Bayram, H. M., & Ozturkcan, A. (2025). Applications of generative and predictive AI in nutrition and dietetics: A narrative review. *Informatics for Health and Social Care*, *50*, 133–156. https://doi.org/10.1080/17538157.2025.2560834

Bekar, D., Azmani Matar, A., Tari Selçuk, K., Öngün Yılmaz, H., & Arslan, S. (2025). Ultra-processed foods and premenstrual syndrome symptoms: Does consumption amount increase symptom severity? *Nutrition and Health*, Advance online publication, *32*, 905–912. https://doi.org/10.1177/02601060251338498

Durmaz, D., Gezgiç, B., Tarı Selçuk, K., Günşen, U., Arslan, S., & Öngün Yılmaz, H. (2025). Does social media addiction increase ultra-processed food consumption and influence functional food attitudes in adults? *Unika Sağlık Bilimleri Dergisi*, *5*, 329–343.

Grilo, M. F., Nunes, B. S., Duran, A. C., Zancheta Ricardo, C., Baraldi, L. G., Martinez Steele, E., & Borges, C. A. (2025). Applying the Nova food classification to food product databases using discriminative ingredients: A methodological proposal. *Frontiers in Public Health*, *13*, 1575136. https://doi.org/10.3389/fpubh.2025.1575136

Hall, K. D., Ayuketah, A., Brychta, R., Cai, H., Cassimatis, T., Chen, K. Y., Chung, S. T., Costa, E., Courville, A., Darcey, V., Fletcher, L. A., Forde, C. G., Gharib, A. M., Guo, J., Howard, R., Joseph, P. V., McGehee, S., Ouwerkerk, R., Raisinger, K. *et al.* (2019). Ultra-processed diets cause excess calorie intake and weight gain: An inpatient randomized controlled trial of ad libitum food intake. *Cell Metabolism*, *30*, 67–77.e3. https://doi.org/10.1016/j.cmet.2019.05.008

Hu, G., Flexner, N., Tiscornia, M. V., & L'Abbé, M. R. (2023). Accelerating the classification of NOVA food processing levels using a fine-tuned language model: A multi-country study. *Nutrients*, *15*, 4167. https://doi.org/10.3390/nu15194167

Karabiyik, D., Aslan, H., Tari Selçuk, K., Tiğli, A., Arslan, S., & Öngün Yılmaz, H. (2025). Association of ultra-processed food consumption with menopausal symptoms in postmenopausal women. *Women & Health*, *65*, 429–441. https://doi.org/10.1080/03630242.2025.2499175

Lane, M. M., Gamage, E., Du, S., Ashtree, D. N., McGuinness, A. J., Gauci, S., Baker, P., Lawrence, M., Rebholz, C. M., Srour, B., Touvier, M., Jacka, F. N., O'Neil, A., Segasby, T., & Marx, W. (2024). Ultra-processed food exposure and adverse health outcomes: Umbrella review of epidemiological meta-analyses. *BMJ*, *384*, e077310. https://doi.org/10.1136/bmj-2023-077310

Martinez-Steele, E., Khandpur, N., Batis, C., Bes-Rastrollo, M., Bonaccio, M., Cediel, G., Huybrechts, I., Juul, F., Levy, R. B., da Costa Louzada, M. L., Machado, P. P., Moubarac, J. C., Nansel, T., Rauber, F., Srour, B., Touvier, M., & Monteiro, C. A. (2023). Best practices for applying the Nova food classification system. *Nature Food*, *4*, 445–448. https://doi.org/10.1038/s43016-023-00779-w

Menichetti, G., Ravandi, B., Mozaffarian, D., & Barabási, A. L. (2023). Machine learning prediction of the degree of food processing. *Nature Communications*, *14*, 2312. https://doi.org/10.1038/s41467-023-37457-1

Monteiro, C. A. (2009). Nutrition and health. The issue is not food, nor nutrients, so much as processing. *Public Health Nutrition*, *12*, 729–731. https://doi.org/10.1017/S1368980009005291

Monteiro, C. A., Cannon, G., Levy, R. B., Moubarac, J. C., Louzada, M. L., Rauber, F., Khandpur, N., Cediel, G., Neri, D., Martinez-Steele, E., Baraldi, L. G., & Jaime, P. C. (2019). Ultra-processed foods: What they are and how to identify them. *Public Health Nutrition*, *22*, 936–941. https://doi.org/10.1017/S1368980018003762

Monteiro, C. A., Cannon, G., Moubarac, J. C., Levy, R. B., Louzada, M. L. C., & Jaime, P. C. (2018). The UN decade of nutrition, the NOVA food classification and the trouble with ultra-processing. *Public Health Nutrition*, *21*, 5–17. https://doi.org/10.1017/S1368980017000234

Monteiro, C. A., Levy, R. B., Claro, R. M., Castro, I. R. R. D., & Cannon, G. (2010). A new classification of foods based on the extent and purpose of their processing. *Cadernos de Saúde Pública*, *26*, 2039–2049. https://doi.org/10.1590/s0102-311x2010001100005

Monteiro, C. A., Louzada, M. L., Martinez-Steele, E., Cannon, G., Andrade, G. C., Baker, P., Bes-Rastrollo, M., Bonaccio, M., Gearhardt, A. N., Khandpur, N., Kolby, M., Levy, R. B., Machado, P. P., Moubarac, J.-C., Rezende, L. F. M., Rivera, J. A., Scrinis, G., Srour, B., Swinburn, B., & Touvier, M. (2025). Ultra-processed foods and human health: The main thesis and the evidence. *The Lancet*, *406*, 2667–2684. https://doi.org/10.1016/S0140-6736(25)01565-X

Ozkan, I., Ozlu Karahan, T., & Seven Avuk, H. (2025). Processed food consumption based on the NOVA classification is associated with elevated cardiometabolic risk in Turkish adults. *Food Science & Nutrition*, *13*, e71014. https://doi.org/10.1002/fsn3.71014

Petrus, R. R., do Amaral Sobral, P. J., Tadini, C. C., & Gonçalves, C. B. (2021). The NOVA classification system: A critical perspective in food science. *Trends in Food Science & Technology*, *116*, 603–608. https://doi.org/10.1016/j.tifs.2021.08.010

Scrinis, G., Popkin, B. M., Corvalan, C., Duran, A. C., Nestle, M., Lawrence, M., Baker, P., Monteiro, C. A., Millett, C., Moubarac, J. C., Jaime, P., & Khandpur, N. (2025). Policies to halt and reverse the rise in ultra-processed food production, marketing, and consumption. *The Lancet*, *406*, 2685–2702. https://doi.org/10.1016/S0140-6736(25)01566-1

Shyr, C., Ren, B., Hsu, C. Y., Yan, C., Tinker, R. J., Cassini, T. A., Hamid, R., Wright, A., Bastarache, L., Peterson, J. F., Malin, B. A., & Xu, H. (2025). A statistical framework for evaluating the repeatability and reproducibility of large language models. *medRxiv*, preprint. https://doi.org/10.1101/2025.08.06.25333170

Şimşek, H., Rajabi, A., Öztürk, B., & Uçar, A. (2025). Ultra-processed foods marketed in Türkiye: An analysis of nutritional quality and packaging sustainability. *Nutrition Bulletin*, *50*, 302–310. https://doi.org/10.1111/nbu.70005

International Journal of Food Science and Technology, 2026, **61(2)**, vvag142
https://doi.org/10.1093/ijfood/vvag142
Original Article