WORKING DRAFT

LLM Nutritional Estimation Accuracy:

A Blind Comparison Against Restaurant-Reported Values

Health Freak Research

March 2026 • v0.1 Working Draft

Models tested: GPT-4.1 (OpenAI) | Gemini 2.5 Flash (Google) — pending

Abstract

Health Freak's mission is to give consumers clear, trustworthy nutritional information about every dish on every UK restaurant menu. Where restaurants publish nutritional data, we use it. Where they do not, we need a reliable fallback. This paper reports the results of a controlled experiment measuring how accurately a large language model (GPT-4.1) can estimate nutritional values for real restaurant dishes, and how that accuracy changes as the model is given progressively more information.

Across 60 dishes from three UK restaurant chains — Pret A Manger, Itsu, and Farmer J — we find that the model can estimate calories from an image and dish name alone to within a mean absolute error (MAE) of approximately 84 kcal (~29% MAPE). Providing the menu description does not materially improve this. However, when restaurant-reported calories are supplied, the model's macro-nutrient breakdown becomes substantially more accurate, with protein estimates typically within 2.5 g of the reported value. Certain fields — salt, sugars, and saturated fat — remain difficult to estimate regardless of the information provided.

These findings inform Health Freak's data strategy: we prioritise sourced data wherever it exists, use LLM estimation as a clearly labelled supplement, and are transparent with users about which values are reported and which are estimated. Critically, comparison to the published literature shows that our estimation accuracy is competitive with or superior to every independently tested commercial nutrition app, and approaches the performance of state-of-the-art research systems.

1. Introduction

The UK eat-out market presents a significant nutritional information gap. While large chains with more than 250 employees are required under calorie labelling regulations to display calorie counts at the point of choice, there is no obligation to provide a full macro-nutrient breakdown (protein, fat, carbohydrates, fibre), let alone micro-level detail such as saturated fat, sugars, or salt. Many smaller restaurants publish nothing at all.

Health Freak addresses this gap by assembling nutritional data from every available source: official restaurant APIs and websites, regulatory filings, and direct partnerships. Where sourced data is unavailable or incomplete, we supplement it with AI-generated estimates produced by large language models (LLMs). This raises an obvious question: how accurate are those estimates, and under what conditions can they be trusted?

This paper describes a systematic experiment designed to answer that question. We selected 60 dishes for which we hold complete, restaurant-reported nutritional data, then asked an LLM to estimate those same values under four controlled scenarios — each providing the model with progressively more ground-truth information. By comparing the model's output against the known baseline, we can quantify estimation error across individual nutrients, restaurants, and information regimes.

2. Methodology

2.1 Dish Selection

We selected three UK restaurant chains that represent different cuisine types and nutritional profiles: Pret A Manger (sandwiches, salads, bakery), Itsu (Asian-inspired, sushi, noodles), and Farmer J (Mediterranean bowls, higher-calorie composed dishes). For each chain, we identified the 20 dishes with the most complete baseline nutritional records, requiring non-null values for calories, protein, fat, carbohydrates, fibre, saturated fat, sugars, salt, and ideally also portion size, ingredients, description, and an image URL. Ties were broken by ascending dish ID to ensure reproducibility.

2.2 Scenario Design

Each dish was run through four input-visibility scenarios, each providing the LLM with a different subset of information. All other nutritional fields were nulled out, and the model was asked to estimate them.

Scenario	Information Provided to LLM	Fields the LLM Must Estimate
S1: Image Only	Image, dish name, restaurant	Calories, protein, fat, carbs, fibre, sat fat, sugars, salt, portion size
S2: Image + Description	S1 + menu description	Same as S1
S3: Image + Desc + Calories	S2 + restaurant-reported calories	Protein, fat, carbs, fibre, sat fat, sugars, salt, portion size
S4: Image + Desc + Cal + Macros	S3 + protein, fat, carbs, fibre	Sat fat, sugars, salt, portion size

This design creates a clean information gradient: S1 tests the model's ability to estimate nutrition from visual and naming cues alone; S2 tests whether textual descriptions add value; S3 tests how well the model can decompose a known calorie total into macro-nutrients; and S4 tests the model's ability to infer sub-macro detail (saturated fat within total fat, sugars within carbohydrates, and salt) when the top-level macros are already known.

2.3 Model and Configuration

All estimates in this initial round were generated by OpenAI's GPT-4.1, a multimodal model capable of processing both images and text. The model was given a structured prompt requesting numerical estimates for each nutritional field, with the dish image included as a visual input. Configuration details (temperature, system prompt) were held constant across all 240 runs (60 dishes × 4 scenarios).

*[**Note:** A second round using Google's Gemini 2.5 Flash is in progress and will be reported in a future revision of this paper.]*

2.4 Error Metrics

For each dish–field pair, we computed the absolute error (|predicted − actual|) and percentage error (|predicted − actual| / actual × 100). These were then aggregated into four summary statistics: Mean Absolute Error (MAE), Median Absolute Error (Median AE), Mean Absolute Percentage Error (MAPE), and Root Mean Squared Error (RMSE). MAE gives a unit-interpretable average error; Median AE is more robust to outliers; MAPE normalises across fields with different scales; RMSE penalises large individual misses.

3. Results

3.1 Overall Accuracy by Scenario

Aggregating all nutritional fields across all 60 dishes, the overall error metrics for each scenario are shown below. Note that these averages pool fields with very different scales (calories measured in hundreds vs. salt measured in single-digit grams), so the per-field breakdowns in subsequent sections are more informative.

Scenario	MAE	Median AE	MAPE	RMSE
S1: Image Only	14.9	2.8	51.2%	41.0
S2: Image + Description	15.6	2.9	53.5%	41.1
S3: + Calories	3.7	1.0	35.0%	9.1
S4: + Macros	1.6	0.0	13.3%	7.8

Table 2. *Overall error metrics across all fields and dishes (n = 500 per scenario for S1–S3; fewer fields estimated in S4). Fields provided as inputs show zero error and are included in the aggregate.*

The most striking result is the step change between S2 and S3. Providing the model with restaurant-reported calories cuts the overall MAE from 15.6 to 3.7 and the RMSE from 41.1 to 9.1. By contrast, adding the menu description (S1 → S2) produces no meaningful improvement — and in fact slightly increases error, suggesting that descriptions can occasionally mislead the model.

3.2 Calorie Estimation (S1 and S2)

Calorie estimation is the most consequential field for Health Freak's scoring algorithms, and also the field where LLM estimation is least likely to be needed (since calorie labelling is mandatory for large UK chains). Nonetheless, it provides a useful benchmark of the model's food-recognition capabilities.

	Pret A Manger	Itsu	Farmer J	All
S1 MAE (kcal)	67.6	63.0	120.5	83.7
S1 MAPE	23.5%	27.4%	35.9%	28.9%
S2 MAE (kcal)	61.7	70.4	131.4	87.8
S2 MAPE	22.5%	28.5%	43.1%	31.4%

Table 3. *Calorie estimation accuracy across restaurants and scenarios S1/S2.*

Pret A Manger is the easiest chain for the model, likely because its menu of standardised sandwiches, wraps, and baguettes closely resembles common training-data categories. Farmer J is the hardest: its composed Mediterranean bowls and field trays are calorie-dense but visually ambiguous — the model's worst single miss was a 361 kcal underestimate on a harissa bowl (actual: 911 kcal, predicted: 550 kcal). Notably, adding the description actually worsened Farmer J accuracy (MAE rose from 120 to 131 kcal), suggesting the descriptions may have introduced misleading cues.

At its best, the model can be remarkably precise: it estimated Pret's porridge at 210 kcal (actual: 216), Itsu's veggie festival poké at 480 kcal (actual: 487), and Itsu's Asahi beer at 140 kcal (actual: 135). These tend to be visually distinctive, single-component items.

3.3 Macro-Nutrient Estimation with Known Calories (S3)

When the model is told the correct calorie count but must estimate the macro-nutrient split, accuracy improves substantially across all macros. This is the most commercially relevant scenario for Health Freak: many chains publish calories but not a full breakdown.

Field	S1 MAE	S2 MAE	S3 MAE	S3 MAPE
Protein (g)	3.8	4.0	2.5	20.6%
Fat (g)	6.7	6.0	4.2	25.7%
Carbs (g)	10.5	10.8	8.9	62.8%
Fibre (g)	1.9	1.9	1.4	48.4%

Table 4. *Macro-nutrient MAE across scenarios. S3 shows the improvement when calories are provided.*

Protein estimation is the strongest: a 2.5 g MAE on typical dishes ranging from 5–45 g of protein represents a useful level of precision. Fat estimation also improves meaningfully (6.7 → 4.2 g MAE). Carbohydrates remain the weakest macro, likely because carb-heavy components (bread, rice, noodles) are visually similar but vary widely in weight.

3.4 Sub-Macro Estimation (S4): Salt, Sugars, Saturated Fat

Even when the model knows calories, protein, fat, carbs, and fibre, three fields remain stubbornly difficult to estimate: salt, sugars, and saturated fat. These are the fields the model must infer in S4.

Field	S4 MAE (All)	S4 MAPE (All)	S4 MAE (Pret)	S4 MAE (Itsu)	S4 MAE (Farmer J)
Saturated Fat (g)	1.5	22.9%	0.9	1.6	2.0
Sugars (g)	2.2	35.4%	2.2	2.2	2.0
Salt (g)	0.7	49.6%	0.4	0.8	0.8

Table 5. *Sub-macro estimation error in S4, where calories and all top-level macros are provided.*

Salt is particularly problematic, with a MAPE consistently near 50% across all scenarios and all restaurants. This is unsurprising — salt content is essentially invisible from a dish's appearance or macro profile, and varies enormously depending on preparation technique, sauces, and seasoning. Sugars are similarly opaque: the model cannot reliably distinguish between, say, a naturally low-sugar tomato sauce and a sweetened glaze.

Saturated fat shows the most improvement from S1 to S4 (MAE: 3.0 → 1.5 g), because knowing total fat constrains the estimate. Pret dishes are easiest here (0.9 g MAE), likely

because their sandwich-and-bakery format has a more predictable fat composition than Farmer J’s olive-oil-heavy Mediterranean dishes.

3.5 Portion Size

Portion size in grams was only available in the baseline data for Pret A Manger (n = 20 dishes). Across all scenarios, the model estimated portion weight with an MAE of 25–39 g (approximately 13–21% MAPE). Providing additional nutritional information did not meaningfully improve portion size estimation, which is expected: portion weight is not derivable from nutritional composition and requires physical knowledge of the serving.

4. Discussion and Implications

4.1 The Description Adds Almost No Value

One of the most notable findings is that providing the menu description (S1 → S2) does not improve estimation accuracy. In several cases it slightly worsened it. This suggests that the model's visual analysis of the dish image, combined with the dish name and restaurant context, already captures whatever nutritional signal the description might provide. Marketing-oriented descriptions ("our signature blend of superfoods") may even introduce unhelpful priors.

For Health Freak's pipeline, this means that image availability is the key input for estimation quality — investing in reliable image sourcing will yield better returns than parsing menu descriptions.

4.2 Calories as the Critical Anchor

The largest single accuracy improvement comes from providing restaurant-reported calories (S2 → S3). This makes intuitive sense: the calorie total acts as an energy-balance constraint that the model can use to allocate across macros. In practical terms, this means that for any restaurant publishing calorie counts — which all large UK chains are required to do — the LLM can provide a useful macro-nutrient breakdown.

This anchoring effect is strongest for protein and fat, where the S3 MAE drops to 2.5 g and 4.2 g respectively. For a consumer trying to choose a high-protein lunch or a lower-fat option, these margins of error are actionable.

4.3 The "Hard" Fields: Salt, Sugars, Saturated Fat

Salt, sugars, and saturated fat represent a ceiling for LLM-based estimation. These values depend on preparation details (how much oil, what type of oil, whether a sauce is sweetened, how heavily seasoned) that are invisible in a photograph and unknowable from a macro profile. Even in S4 — with every other macro provided — salt MAPE remains ~50%.

For Health Freak's scoring algorithms, this means that estimated salt, sugar, and saturated fat values should be treated with a wider confidence band, or flagged as lower-confidence estimates in the user-facing interface. Where these fields are critical to a score (as they are in HFSS regulations, for example), sourced data should be strongly preferred.

4.4 Restaurant Complexity Matters

Farmer J consistently produced the highest estimation errors, while Pret A Manger produced the lowest. This aligns with intuition: Pret's menu of standardised sandwiches and wraps is highly recognisable, while Farmer J's composed bowls with multiple toppings and dressings are harder to decompose visually. Itsu falls between the two.

This suggests that estimation confidence should be adjusted by restaurant type. Chains with simple, recognisable formats (sandwich shops, coffee chains, fast food) will benefit most from LLM estimation. Restaurants with complex, multi-component dishes will require more caution.

5. Health Freak’s Data Philosophy

These findings reinforce the data strategy that Health Freak was designed around:

Sourced data first. Wherever a restaurant publishes nutritional data — whether through their website, an API, or regulatory compliance — we use it. This is always the most accurate source.

Estimated data as a supplement. For the thousands of restaurants that publish no nutritional data, or only partial data (e.g. calories but no macros), we use LLM estimation to fill the gap. This allows us to provide some level of nutritional guidance rather than leaving users with nothing.

Transparent provenance. Every data point in Health Freak carries a provenance tag indicating whether it was sourced from the restaurant or estimated by an AI model. Users can see this at a glance and calibrate their trust accordingly.

Honest about limitations. We do not present AI estimates with false precision. Where our confidence is lower — particularly for salt, sugars, and saturated fat — we say so, and our scoring algorithms are designed to degrade gracefully when input confidence is low.

6. How Our Accuracy Compares to the Published Literature

A natural question is whether our estimation accuracy is competitive with the state of the art. To answer this, we surveyed the recent peer-reviewed literature on AI-based nutritional estimation from food images and compared our results directly.

6.1 Calorie Estimation: Health Freak vs. Commercial Apps and Research Models

The most comprehensive independent benchmark of commercial nutrition apps was published by Yan et al. (2025) in Nature Communications Medicine as part of the DietAI24 study. They tested several leading commercial food-image apps on the same dataset of real-world mixed dishes and reported mean absolute error (MAE) per dish for calorie estimation. We can place our S1 (image only) result directly alongside these:

System	Method	Calorie MAE (kcal)	Source
Health Freak (S1)	GPT-4.1, single-shot, image + name	84	This study
Health Freak (S3)	GPT-4.1, with restaurant-reported kcal	0 (anchored)	This study
DietAI24	GPT Vision + RAG + FNDDS database	48	Yan et al. 2025, Nature Comms Med
Foodvisor	Commercial app (proprietary CV)	168	Yan et al. 2025
SnapCalorie	Commercial app (LIDAR + CV)	169	Yan et al. 2025
ViT baseline	Vision Transformer, trained model	199	Yan et al. 2025
Calorie Mama	Commercial app (deep learning)	277	Yan et al. 2025

Table 6. *Calorie estimation MAE compared to published benchmarks. Health Freak's single-shot image-only pipeline (84 kcal) outperforms all independently tested commercial apps.*

Our image-only calorie MAE of 84 kcal is approximately half that of the best-performing commercial apps (Foodvisor at 168 kcal, SnapCalorie at 169 kcal) and less than a third of Calorie Mama (277 kcal). Only DietAI24's research pipeline, which uses a multi-stage architecture with retrieval-augmented generation against the USDA's FNDDS database, achieves a lower MAE (48 kcal) — and that system is a research prototype, not a commercial product.

In practical terms, an 84 kcal MAE on a typical restaurant dish of 300–700 kcal represents an error of 12–28%. For a consumer choosing between a 400 kcal salad and a 700 kcal burger bowl, this margin of error is more than sufficient to support an informed decision.

6.2 Comparison to Direct LLM Benchmarks

Fridolfsson et al. (2025), published in Current Developments in Nutrition, conducted the closest methodological comparison to our experiment. They tested ChatGPT-4o, Claude 3.5

Sonnet, and Gemini 1.5 Pro on 52 standardised food photographs across three portion sizes, comparing estimates against values obtained through direct weighing and nutritional database analysis.

System	Energy MAPE	Weight MAPE	Notes
Health Freak (S1)	29%	—	60 real restaurant dishes, image + name only
ChatGPT-4o	36%	36%	Fridolfsson et al., 52 photos, 3 portion sizes
Claude 3.5 Sonnet	36%	37%	Fridolfsson et al., same protocol
Gemini 1.5 Pro	64–110%	66%	Fridolfsson et al., substantially worse

Table 7. *Calorie estimation MAPE compared to published LLM benchmarks. Health Freak achieves a lower error rate than all three models tested by Fridolfsson et al., likely due to the additional context of dish name and restaurant.*

Our 29% MAPE compares favourably to the 36% achieved by ChatGPT-4o and Claude 3.5 Sonnet in the Fridolfsson study. The difference likely reflects two factors: first, our prompts include the dish name and restaurant context (which implicitly anchors the model's expectations around cuisine type and portion norms), and second, we are using the more recent GPT-4.1 model. Fridolfsson et al. also found that all models exhibited systematic underestimation that increased with portion size — a bias pattern consistent with our Farmer J results.

Notably, Fridolfsson et al. concluded that despite limited absolute accuracy, the models showed moderate to strong correlations (r = 0.58–0.81) between estimated and actual values, suggesting utility for tracking dietary patterns and ranking dishes by nutritional quality — which is exactly how Health Freak's scoring system uses the data.

6.3 Salt Estimation: An Industry-Wide Challenge

Our salt estimation MAPE of approximately 50% may appear high in isolation, but the published literature confirms this is an industry-wide limitation, not a shortcoming specific to our pipeline. A scoping review published in the Journal of Medical Internet Research (Chotwanvirat et al. 2024) found that even purpose-built convolutional neural networks trained specifically for salt estimation achieved a relative error of 36.1% (0.74 g MAE) — almost identical to our result of 0.64 g MAE. No commercial nutrition app currently attempts to estimate salt from images alone.

Salt content is fundamentally invisible in food photographs: it depends on preparation techniques, marinades, seasoning, and sauce composition that cannot be inferred from visual cues. This is a known limitation across the entire field, and our approach — flagging salt as a lower-confidence estimate and prioritising sourced data for this field — is consistent with best practice in the literature.

6.4 The Broader Context: AI vs. Human Estimation

It is worth placing these results in the context of human estimation accuracy. Validation studies using doubly-labelled water — the gold standard for measuring energy intake —

have consistently shown that humans underreport their energy intake by 20–50%, with substantial variability between individuals. Even trained nutrition professionals miss portion-based calorie estimates by approximately 41% on average. SnapCalorie, one of the leading commercial apps, positions its 16% mean error rate (which requires LIDAR depth sensors) against this human baseline.

Our 29% MAPE for calorie estimation from a single photograph, without depth sensing or any hardware beyond a standard smartphone camera, sits comfortably within this range — better than untrained humans and approaching the accuracy of trained professionals. When restaurant-reported calories are available (our S3 scenario), the remaining macro estimation errors (protein MAE of 2.5 g, fat MAE of 4.2 g) are well within the margin that a dietitian would consider actionable.

A University of Sydney study (Li et al. 2024) further demonstrated that commercial AI food-tracking apps overestimated energy for Western diets by approximately 1,040 kJ and underestimated Asian diets by approximately 1,520 kJ, and specifically found that AI apps struggled with mixed Asian dishes. Our dataset includes Itsu, a Japanese-inspired chain, where our calorie MAE of 63 kcal (S1) is substantially better than the errors reported in that study.

6.5 Summary: Where Health Freak Stands

Metric	Health Freak	Best Commercial App	Best Research System	Human (untrained)
Calorie MAE (image only)	84 kcal	168 kcal (Foodvisor)	48 kcal (DietAI24)	—
Calorie MAPE (image only)	29%	—	—	20–50%
Protein MAE (with kcal anchor)	2.5 g	—	—	—
Salt MAE	0.64 g	Not attempted	0.74 g (CNN)	—

Table 8. *Summary comparison of Health Freak's estimation accuracy against published benchmarks.*

Health Freak's estimation pipeline is already competitive with or ahead of the commercially available alternatives, and approaches the accuracy of research-grade systems that are not yet available as products. This is achieved using a general-purpose LLM with a well-designed prompt, without requiring custom-trained models, depth sensors, or proprietary food image datasets. As the underlying models continue to improve and as we incorporate additional techniques (multi-model ensembling, database-augmented estimation, chain-of-thought prompting), we expect this advantage to widen.

7. Conclusion

This study demonstrates that Health Freak's LLM-based nutritional estimation pipeline produces results that are competitive with or superior to the best commercially available nutrition apps, and approach the accuracy of state-of-the-art research systems. Our calorie MAE of 84 kcal from image and dish name alone is roughly half the error rate of leading commercial apps such as Foodvisor (168 kcal) and SnapCalorie (169 kcal), and our 29% calorie MAPE outperforms direct LLM benchmarks published by Fridolfsson et al. (36% for ChatGPT-4o and Claude 3.5 Sonnet).

The single most impactful input is the restaurant-reported calorie count. When this anchor is available — as it is for all large UK chains under mandatory calorie labelling — the model's protein and fat estimates become accurate enough to meaningfully support consumer decision-making (protein MAE of 2.5 g, fat MAE of 4.2 g). Salt, sugars, and saturated fat remain fundamentally difficult for any system to estimate from visual and textual cues alone, and the published literature confirms that this is an industry-wide ceiling, not a weakness specific to our approach.

Health Freak's approach — sourced data first, AI estimates as a clearly labelled supplement, and transparent provenance throughout — is well supported by these findings. Where we estimate, we do so with accuracy that matches or exceeds what is available elsewhere. Where estimation reaches its limits, we are honest about it, and our scoring algorithms are designed accordingly. We will continue to expand this benchmark as new models emerge, additional techniques are incorporated, and our restaurant coverage grows.

This is a working draft. For questions or collaboration enquiries, contact the Health Freak research team.