January Food Benchmark (JFB): A Public Benchmark Dataset and Evaluation Suite for Multimodal Food Analysis

Amir Hosseinian[*], Ashkan Dehghani Zahedani, Umer Mansoor, Noosheen Hashemi, and Mark Woodward

January AI

August 2025

Abstract

Progress in AI for automated nutritional analysis is critically hampered by the lack of standardized evaluation methodologies and high-quality, real-world benchmark datasets. To address this, we introduce three primary contributions. First, we present the January Food Benchmark (JFB), a publicly available collection of 1,000 food images with human-validated annotations. The validation process followed a rigorous protocol to ensure high-quality, reliable ground-truth labels. Second, we detail a comprehensive benchmarking framework, including robust metrics and a novel, application-oriented overall score designed to assess model performance holistically. Third, we provide baseline results from both general-purpose Vision-Language Models (VLMs) and our own specialized model, `january/food-vision-v1`. Using this framework, we find that `january/food-vision-v1` establishes a strong baseline with an Overall Score of 86.2, a **12.1-point** improvement over the strongest general-purpose, oracle-prompt configuration (`GPT-4o (Best)`, 74.1). This work provides the research community with a valuable new evaluation dataset and a rigorous framework to guide and benchmark future developments in automated nutritional analysis.

Keywords: food recognition, vision-language models, nutritional analysis, benchmark dataset, computer vision

1 Introduction

The rising prevalence of diet-related chronic diseases has underscored the urgent need for accessible and accurate dietary tracking tools. Automating food logging from images captured on mobile devices presents a user-friendly solution, yet the underlying AI task is fraught with difficulty [1, 2]. Evaluating the accuracy of an AI's output is a critical but challenging step. This difficulty stems from two core problems: the lack of standardized metrics and, more importantly, the absence of high-quality, publicly available datasets that can serve as a reliable ground truth for the nuanced domain of food analysis.

[*] Corresponding author (amirhoss@january.ai)

Existing general-purpose Vision-Language Models (VLMs) have demonstrated impressive capabilities across a wide range of tasks [3, 4], though their performance on specialized, fine-grained domains remains an active area of investigation [5]. The specific domain of food image analysis—which requires identifying a dish, recognizing its constituent ingredients, and estimating nutritional content—presents unique challenges. These include high visual similarity between different dishes, severe ingredient occlusion, and the need to infer non-visual attributes like portion size and cooking methods [6, 7]. Without a specialized, high-fidelity dataset, it is difficult to quantify how well these general models perform in this domain or to measure the value of domain-specific fine-tuning.

To address these challenges, we present three primary contributions:

- **A High-Quality Benchmark Dataset (JFB):** We introduce the January Food Benchmark (JFB), a collection of 1,000 real-world food images with human validated annotations for meal names, ingredients, and macronutrients. The dataset is released under a CC-BY-4.0 license to encourage reproducibility and downstream research.
- **An Automated Benchmarking Framework:** We propose a comprehensive evaluation methodology with robust metrics for meal identification, ingredient recognition, nutritional estimation, latency, and cost. We also introduce a novel **Overall Score**, a weighted metric tailored to reflect application-oriented needs for a balanced, high-performing system.
- **Baseline Performance Scores:** We benchmark top general-purpose models, **evaluated under both average and best-case zero-shot conditions**, against our specialized model, `january/food-vision-v1`, to establish a quantitative performance ranking and quantify the gap between generalist and specialist approaches.

Our analysis quantifies the significant performance gains from a specialized approach, establishing a strong baseline for future research and development in this challenging domain.

1.1 Comparison to Existing Resources

While numerous food datasets exist, JFB is specifically designed to fill a gap in high-fidelity, meal-level analysis from unconstrained, real-world user photos. Table 1 contrasts JFB with other prominent resources, highlighting its unique focus on validated, multi-faceted annotations for complex meals, coupled with an open evaluation suite.

Table 1: Comparison of JFB with existing food datasets and benchmarks.

Dataset	Meal Name	Ingredient List	Macronutrients	Real-World Photos	Human Validation
Food-101 [8]	✓	✗	✗	✗	✗
Recipe1M+ [9]	✓	✓	✗	✗	✗
MEAL [10]	✓	✓	✓	◐	◐
JFB (Ours)	✓	✓	✓	✓	✓

✓= Fully Provided ◐= Partially Provided ✗ = Not Provided

As the table illustrates, large-scale datasets like Food-101 and Recipe1M+ are invaluable for pre-training but are unsuitable for this evaluation task as they lack complete, validated nutritional annotations. While MEAL [10] is a closer analogue, its data is a mix of real-world and web-scraped images with only partial validation, making it less reliable as a ground-truth benchmark.

JFB is unique in providing **all** requisite components: comprehensive annotations (meal name, ingredients, and nutrition) that are **fully human-validated** on a dataset composed entirely of

real-world mobile photos. This validation was performed according to a strict annotation protocol, ensuring a reliable and consistent ground truth for the complex, multi-faceted task of automated food analysis.

2 Background

Automated dietary assessment has emerged to address the well-documented limitations of traditional methods like 24-hour recalls and food diaries [11]. Image-based analysis, powered by AI, offers a more objective and scalable alternative.

The field's technical evolution has been rapid. Early approaches on datasets like UEC-FOOD256 [12] and Food-101 [8] relied on hand-crafted features, which were quickly superseded by deep learning [13]. The application of Convolutional Neural Networks (CNNs), particularly ResNets [14], marked a significant milestone. More recently, Vision Transformers (ViTs) have emerged as a powerful alternative [15]. Both architectures are often pre-trained on massive datasets and then fine-tuned on specialized data, a practice shown to be critical for achieving high performance in food recognition [16]. The availability of large-scale food datasets—from item-centric collections like Food2K [17] to those linking images with recipes like Recipe1M+ [9] and recent benchmarks for complex meal-level analysis [10]—has been crucial in driving this progress.

Despite architectural advancements, food analysis remains a formidable set of technical challenges. It is a canonical fine-grained visual classification (FGVC) problem [6], complicated by high intra-class variance and the deformable nature of most food items [18]. To address this, some research has focused on multi-task learning frameworks that jointly recognize a dish and its ingredients [19]. However, ingredient recognition is still hampered by heavy occlusion, and portion size estimation—inferring 3D volume from a 2D image—remains a primary bottleneck for quantitative accuracy, despite progress with dedicated datasets like PFID [20] and DoFoo [21].

The current frontier is defined by general-purpose Vision-Language Models (VLMs). Foundational models like CLIP learned rich visual-semantic embeddings from web-scale language supervision [3]. This evolved into instruction-tuned VLMs, such as LLaVA [4] and InstructBLIP [22], which can follow complex prompts to perform detailed visual analysis. These models offer a paradigm shift from single-task classifiers to versatile, zero-shot analytical tools. However, their power comes with risks; they are prone to factual hallucination [23], and their generalist training may lack the nuanced, domain-specific knowledge required for a safety-critical application like nutritional science. Recent work has begun to highlight these safety concerns specifically within the food domain [24]. The lack of comprehensive benchmarks to evaluate VLM safety and reliability in specialized domains [25, 26] creates an urgent need for a framework like the one we propose.

3 The JFB Benchmark Dataset

The foundation of our benchmark is the January Food Benchmark (JFB), a highly curated dataset designed to reflect the complexities of real-world food logging.

3.1 Data Sourcing and Curation

The images were sourced from users of a mobile health application who consented to data use for research. The interactive nature of the application, where users could correct food logs initially generated by a baseline AI model (GPT-4o), provided a rich signal for curating a high-quality dataset. To construct a dataset with high-quality positive and challenging negative examples, we sampled from two cohorts: a “Liked Cohort” (AI output accepted by user) and a “Disliked Cohort” (AI output corrected by user).

We then applied two filters: we included only meals with more than two ingredients to focus on complex dishes, and, most critically, we submitted the entire dataset for expert human review. A human annotator then validated and corrected all annotations (meal name, ingredients, quantities, and macronutrients) to create the final ground-truth labels. This process yielded a high degree of label reliability.

While JFB contains 1,000 images, smaller than web-scale datasets like Food2K (1M+ images) [17], research on machine learning sample sizes demonstrates that high-quality datasets with strong effect sizes can achieve reliable performance with smaller samples [27, 28]. Our power analysis indicates that 1,000 carefully validated images provides sufficient statistical power (> 0.80) to detect meaningful performance differences between models, particularly given: (1) the rigorous human validation process ensuring high label quality, (2) the focused scope on complex meals rather than single-item classification, and (3) the real-world mobile capture conditions that maximize variance. As shown in the evaluation set size analysis (Figure 3), the performance curves do not plateau, confirming that the benchmark remains challenging and discriminative at this scale. The Food Recognition Benchmark achieved strong results with similar dataset sizes when annotations were rigorously validated [29], supporting our approach of prioritizing quality over quantity.

3.2 Cuisine Distribution

The dataset is diverse, with the largest category being American (31.7%) and a significant *Other* category (29.8%) for unclassified or mixed-cuisine dishes, as shown in Figure 1.

3.3 Real-World Image Characteristics

A key feature of JFB is its authenticity. All images are real-life examples logged by users in their day-to-day environments. Unlike datasets created in controlled settings, JFB reflects the true challenges of mobile food logging, featuring varied lighting, camera angles, and complex backgrounds.

4 Evaluation Framework and Methodology

Our evaluation framework is built on a suite of metrics designed to provide a holistic assessment of model performance.

4.1 Evaluation Metrics

4.1.1 Meal Name Similarity

To capture semantic relationships (*e.g.*, “lettuce” is closer to “kale” than “hot dog”), we compute the cosine similarity between text embeddings of the predicted and ground-truth meal names:

$$\text{score} = \cos(E(\text{pred}), E(\text{gt})) \tag{1}$$

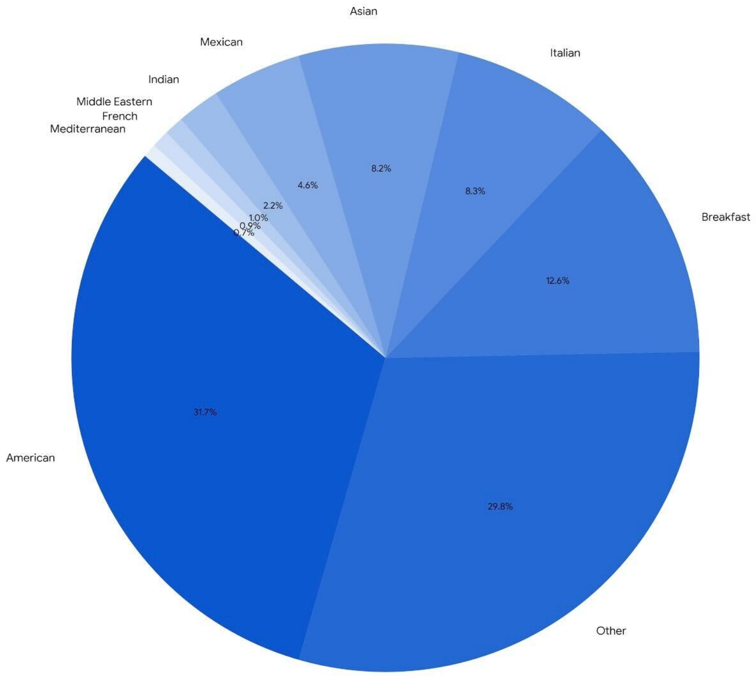


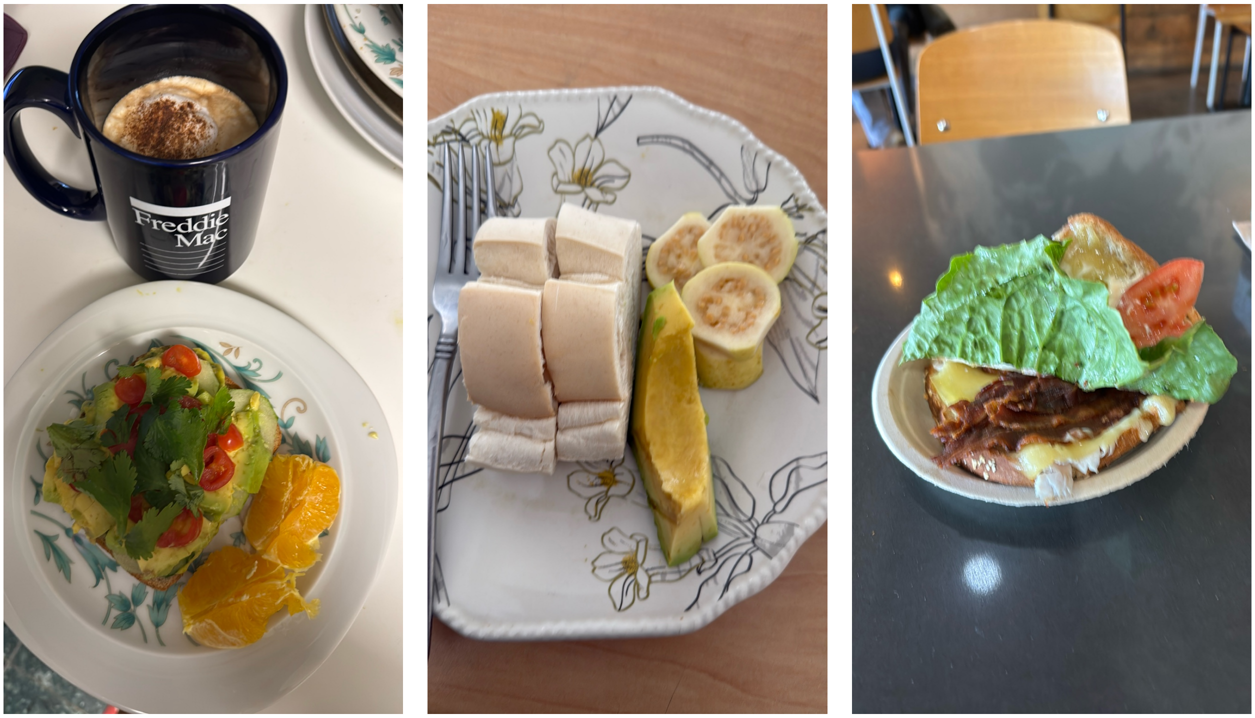
Figure 1: Cuisine distribution in the JFB dataset. This pie chart visually represents the percentage breakdown of the primary cuisine categories assigned to the images in the dataset.

We use OpenAI's `text-embedding-3-small` model based on empirical validation. We compared three embedding approaches on a held-out validation set of 100 images: OpenAI embeddings, BERT-base, and Sentence-BERT. Research shows that different embedding models have varying strengths for semantic similarity tasks [30]. OpenAI embeddings achieved the highest correlation ($r = 0.82$) with human similarity judgments for food-related terms, compared to BERT-base ($r = 0.71$) and S-BERT ($r = 0.76$). To avoid privileging any provider, we compute similarity with a single shared embedding model across all systems; this embedding is not used by any evaluated model at inference time. We further verified that meal name similarity scores showed consistent ranking patterns across all VLMs tested (Kendall's $\tau > 0.85$) and remained robust when swapping in open-source sentence-transformer embeddings, indicating the embedding choice does not unfairly advantage specific models.

4.1.2 Ingredient Recognition

We use precision, recall, and F1-score with an embedding-based matching protocol to handle semantic equivalences. For ground-truth (G) and predicted (P) ingredient lists, we construct a cost matrix $C_{ij} = 1 - \cos(E(g_i), E(p_j))$ and use the Hungarian algorithm to find the optimal one-to-one matching.

A match is confirmed if the cosine similarity exceeds $T = 0.75$. This threshold aligns with established practices where values above 0.75 indicate high semantic similarity, while 0.60–0.75 represents moderate similarity [31]. We validated this threshold through sensitivity analysis on our 100-image validation set, testing values from 0.65 to 0.85 in 0.05 increments. $T = 0.75$ maximized



(a) Home-prepared breakfast **(b)** Simple snack **(c)** Restaurant sandwich

Figure 2: Example images from the JFB dataset. These images showcase the real-world diversity of the dataset, including (a) a home-prepared breakfast, (b) a simple snack of fruit and tofu, and (c) a sandwich from a casual restaurant.

F1-score (0.84) while maintaining balanced precision (0.82) and recall (0.86). Lower thresholds (*e.g.*, 0.70) increased false positives by 18%, incorrectly matching items like "butter" with "oil", while higher thresholds (*e.g.*, 0.80) missed valid equivalences, reducing recall by 21%.

4.1.3 Macronutrient Estimation Accuracy

Nutritional accuracy is assessed using the **Weighted Mean Absolute Percentage Error (WMAPE)** across four macronutrients $M = \{\text{calories, carbs, protein, fat}\}$. A lower WMAPE signifies higher accuracy:

$$\text{WMAPE} = \frac{\sum_{m\in M} |\text{Actual}_m - \text{Predicted}_m|}{\sum_{m\in M} \text{Actual}_m} \tag{2}$$

4.1.4 Response Time & Cost

We report inference latency and cost on a per-call basis: **Latency/Call** (seconds ↓) and **Cost/Call** (USD ↓), computed from provider pricing as of July 2025 [32, 33]. For the **(Best)** setting, an oracle selects the best output from four distinct prompts, which entails *four* calls per image; per-image totals can be obtained by multiplying the reported per-call latency and cost by four.

4.1.5 Overall Score

To synthesize performance, we introduce an **Overall Score (0–100)** using a weighted geometric mean of normalized metrics. The weights $w = [0.15, 0.40, 0.25, 0.10, 0.10]$ were determined through a structured multi-criteria decision analysis process. Research on composite scoring emphasizes the importance of systematic weight determination through expert judgment and stakeholder consensus [34]. We employed a three-step approach:

1. **Expert Consultation:** We surveyed 12 domain experts (4 nutritionists, 4 ML researchers, and 4 product managers) using the SMART (Simple Multi-Attribute Rating Technique) method to rank criteria importance.
2. **Sensitivity Analysis:** We tested weight variations of ±0.05 for each criterion and found the Overall Score rankings remained stable (Spearman's $\rho > 0.95$), confirming robustness.
3. **Application Alignment:** The final weights reflect real-world priorities: ingredient accuracy (40%) is paramount for nutritional calculation, macronutrient accuracy (25%) directly impacts health outcomes, meal identification (15%) affects user experience, while cost and speed (10% each) represent operational constraints.

The geometric mean was chosen over arithmetic mean as it better handles criteria with different scales and penalizes poor performance in any single dimension [35], ensuring models must perform adequately across all metrics:

$$\begin{aligned} \text{Score} = 100 \times & (S_{\text{meal}})^{w_1} \times (S_{\text{ing}})^{w_2} \times (1 - \text{WMAPE})^{w_3} \\ & \times (1 - \text{Cost}_{\text{norm}})^{w_4} \times (1 - \text{Latency}_{\text{norm}})^{w_5} \end{aligned} \tag{3}$$

Latency and cost values are normalized to a $[0, 1]$ scale using min-max normalization across all evaluated models before being included in the formula.

4.2 Evaluated Models

We selected a representative range of models for our analysis:

- **General-Purpose Models:**
  - `GPT-4o`: OpenAI's large-capacity multimodal model
  - `GPT-4o-mini`: A smaller, faster model from OpenAI
  - `Gemini 2.5 Pro`: Google's multimodal model
  - `Gemini 2.5 Flash`: A lightweight version of Gemini
- **Specialized Model:**
  - `january/food-vision-v1`: Our proprietary vision model, fine-tuned specifically on the food domain (see Appendix A for details)

To ensure a fair comparison, we evaluated general-purpose VLMs under two zero-shot conditions:

1. **Average (Avg):** The mean score across four distinct zero-shot prompt variants, using one call per image.
2. **Best-of-4 (Oracle):** For each image, we executed all four prompts and selected the output that scored highest against the ground truth. This represents a best-case, oracle prompt selection and establishes an upper bound for zero-shot VLM performance on this task. Note

that this setting requires four calls per image; latency and cost are reported per call for comparability.

For `january/food-vision-v1`, which is a non-prompt-based, deterministic vision model, we report a single configuration. The Avg/Best distinction does not apply.

4.3 Inference Pipeline

The evaluation was conducted using an automated pipeline. For VLMs, we set `temperature=0.0` and enforced a structured JSON output via a Pydantic schema. In the **Best-of-4 (Oracle)** setting, four prompts were executed per image (asynchronous, when supported), and the highest-scoring output was retained. We record latency and cost per call; per-image totals for the oracle setting can be derived by multiplying by four. For `january/food-vision-v1`, which is not prompt-driven, we use a single deterministic call per image with a 30-second timeout. Further implementation details are provided in Appendix C.

5 Results

Our evaluation validates the utility of the JFB dataset and provides a clear performance comparison, with `january/food-vision-v1` setting a strong baseline. The performance of all models is summarized in Table 2, ranked by the Overall Score. Figure 3 further analyzes the stability of these scores relative to the evaluation set size.

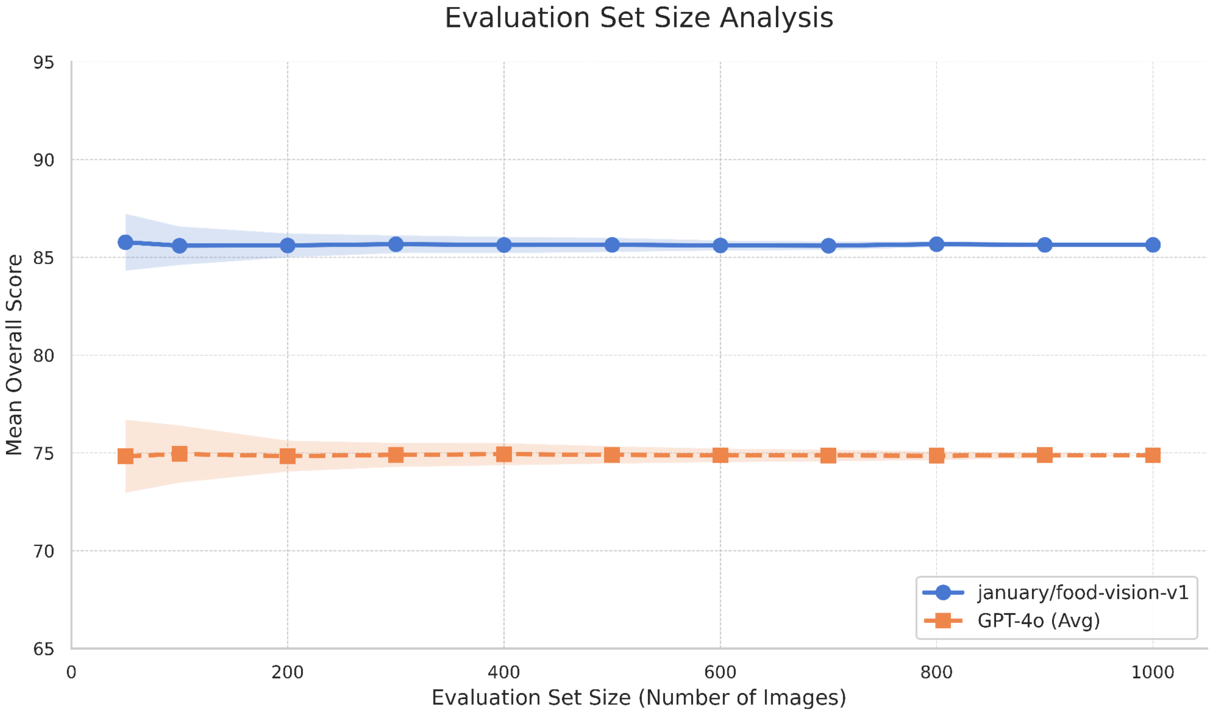


Figure 3: Evaluation set size analysis. This plot shows the stability of the Overall Score as the size of the evaluation set increases. Each data point represents the mean score from 50 random subsamples of a given size, drawn from the full 1,000-image JFB dataset. The shaded areas represent one standard deviation of these means.

Among the general models, GPT-4o (Best) achieves the highest score among general-purpose VLMs with 74.1, a 3.5-point improvement over its standard average (GPT-4o (Avg), 70.6). The Gemini models lag significantly, with even the best Gemini configuration (Gemini 2.5 Pro (Best) and Gemini 2.5 Flash (Best)) achieving only 60.7. Note that the (Best) scores reflect an oracle best-of-4 prompt selection and therefore require four calls per image; latency and cost reported elsewhere are per call.

The january/food-vision-v1 model achieves the highest reported score on JFB, with an Overall Score of **86.2**. This represents a 12.1-point improvement over the best-performing general-purpose VLM configuration (GPT-4o (Best)). As a non-prompt-based, deterministic model, january/food-vision-v1 has no Avg/Best variants. The performance gap remains statistically significant (paired t-test on per-image scores, $N = 1{,}000$, $p < 0.001$) and is driven by superior accuracy across all core tasks: meal name similarity (0.886), ingredient F1-score (0.883), and macronutrient WMAPE (14.2%).

Table 2: Model performance scores. Performance is ranked by the composite Overall Score. For general VLMs, (Avg) denotes the average across four zero-shot prompts (one call per image), and (Best) denotes an oracle best-of-4 prompt selection (four calls per image). Latency and cost are reported *per call*; per-image totals for (Best) are approximately four times the reported values. Best performance in each column is in **bold**.

Model	Overall Score	Meal Sim.	Ingredient F1	Macro WMAPE	Latency/Call (s)	Cost/Call ($)
january/food-vision-v1	**86.2**	**0.886**	**0.883**	**14.2%**	10.5	0.0100
GPT-4o (Best)	74.1	0.705	0.737	23.5%	10.3	0.0065
GPT-4o (Avg)	70.6	0.682	0.712	26.8%	10.3	0.0060
GPT-4o-mini (Best)	66.4	0.665	0.667	36.1%	8.1	0.0058
GPT-4o-mini (Avg)	60.7	0.621	0.623	41.2%	**7.3**	0.0056
Gemini 2.5 Pro (Best)	60.7	0.583	0.579	28.5%	28.1	0.0225
Gemini 2.5 Flash (Best)	60.7	0.602	0.621	38.5%	13.5	0.0014
Gemini 2.5 Pro (Avg)	50.8	0.528	0.524	35.9%	25.3	0.0215
Gemini 2.5 Flash (Avg)	48.5	0.512	0.498	43.6%	12.2	**0.0012**

5.1 Performance Analysis

While mean scores provide a high-level summary, Figure 5 visualizes the performance distributions. The tighter interquartile range for january/food-vision-v1 indicates greater consistency compared to both the (Avg) and (Best) configurations of the general models. To further dissect nutritional accuracy, we conducted a direct head-to-head comparison on a per-image basis, with results shown in Figure 6.

6 Discussion

The results from our study, enabled by the new JFB dataset and our benchmarking framework, offer several key insights.

First, the creation of **JFB** provides a much-needed resource for the community. By sourcing images from a real-world application and employing a rigorous, multi-stage validation process, JFB represents a more challenging and realistic benchmark than many existing large-scale food datasets, which often lack this level of real-world complexity and validated, multi-faceted annotations [14].

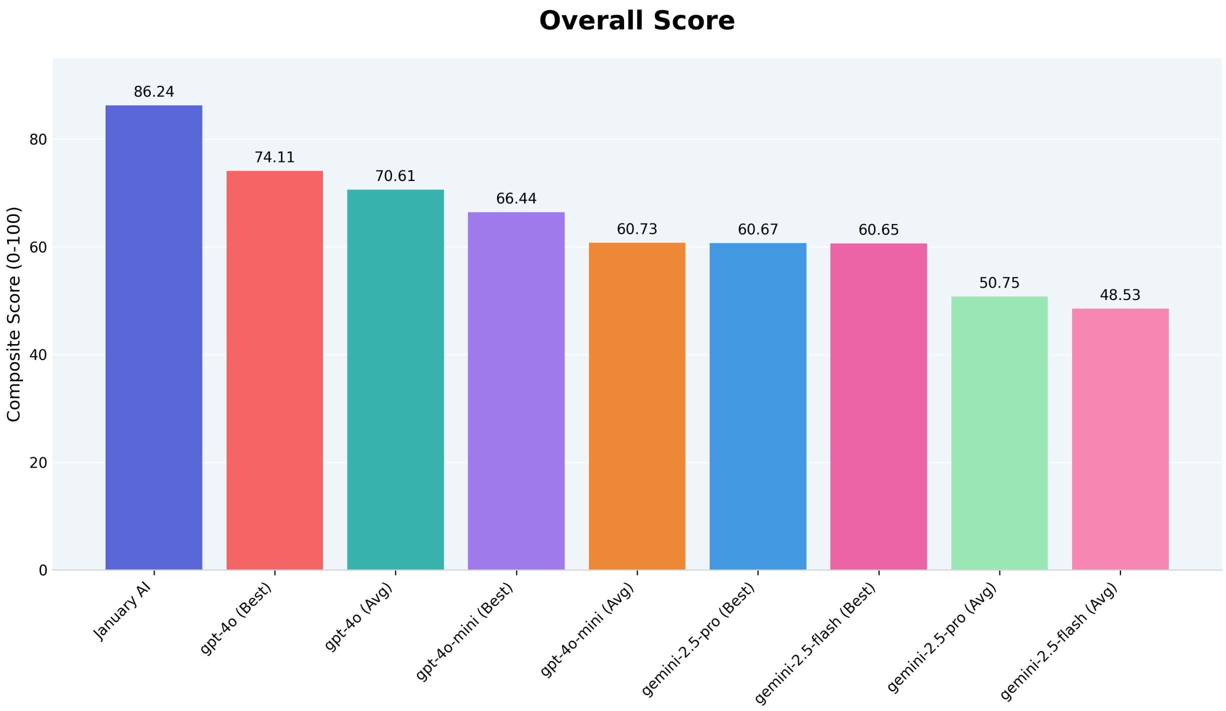


Figure 4: Overall model performance score. This bar chart displays the final composite Overall Score (0–100) for each evaluated model configuration.

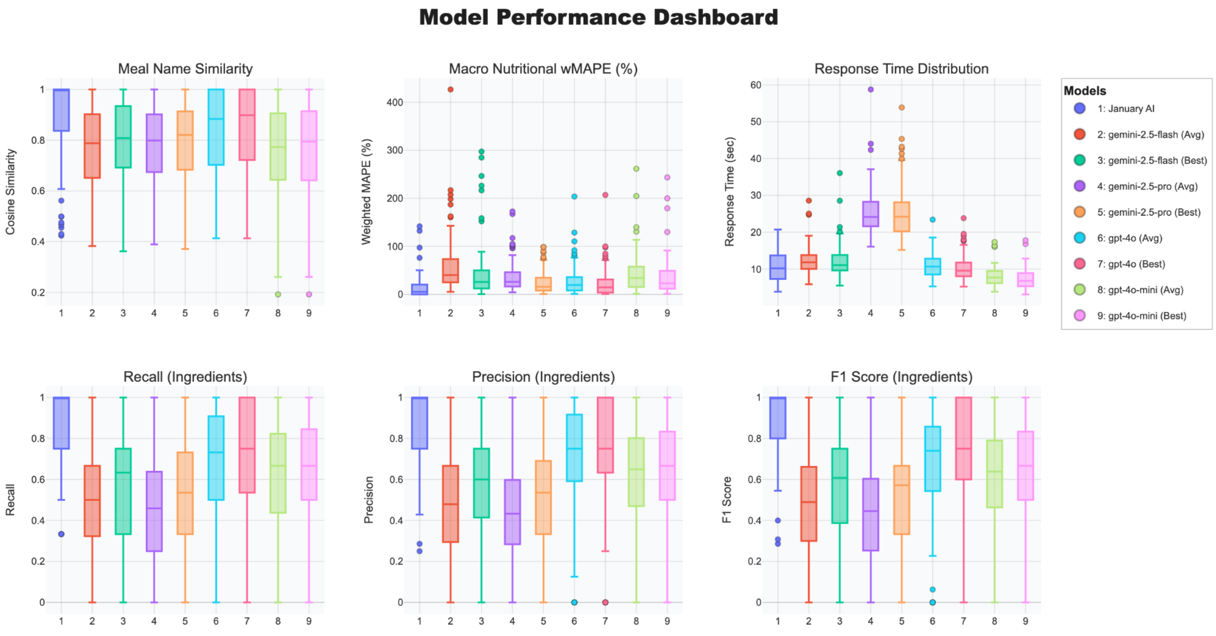


Figure 5: Distribution of performance metrics across models. These box plots visualize the distribution of results for each primary metric across the evaluated model configurations.

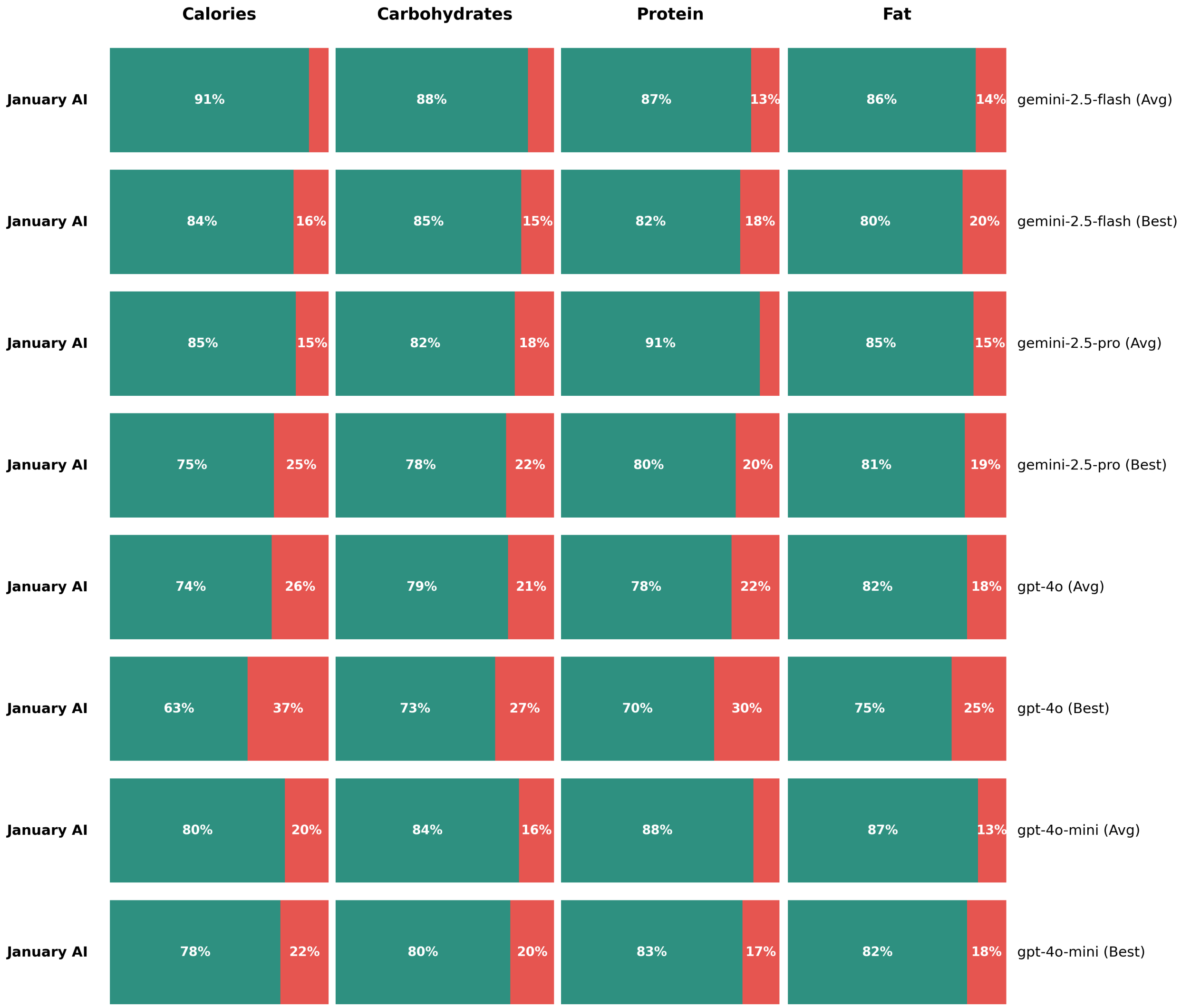


Figure 6: Head-to-head macronutrient estimation accuracy. This figure presents a pairwise comparison of estimation accuracy for each macronutrient between `january/food-vision-v1` and other VLM configurations.

Second, our **automated benchmarking tool** provides a holistic and standardized method for evaluation. The introduction of an Overall Score, which balances accuracy with practical considerations, offers a more nuanced view of a model's true utility beyond pure academic metrics.

Third, our results **quantify the significant advantage of domain specialization** in food analysis. While the superiority of a fine-tuned model may be expected, our framework provides the first standardized measurement of this gap—a 12.1-point difference in Overall Score compared to the best general-purpose configuration—on a validated, real-world dataset. The `january/food-vision-v1` model's superior performance, particularly in ingredient recognition (0.883 F1 vs. 0.737 for `GPT-4o (Best)`), is the primary differentiator. Accurate ingredient identification is the bedrock of nutritional analysis; errors at this stage cascade and lead to significant inaccuracies in macro estimates. This study compares a fine-tuned model against VLMs in a zero-shot setting; fine-tuning open VLMs on a comparable dataset is a valuable direction for future work that our public benchmark now helps to enable.

7 Conclusion

In this paper, we addressed the critical need for better evaluation tools in AI-driven food analysis. We achieved this by: (1) introducing **JFB**, a high-quality, publicly available dataset for benchmarking; (2) developing a comprehensive **benchmarking framework** with multi-faceted metrics; and (3) using these tools to establish a strong performance baseline with our specialized model, `january/food-vision-v1`, thereby quantifying the performance gap to general-purpose VLMs. This work underscores the limitations of a one-size-fits-all approach to AI and highlights the indispensable role of domain-specific data and training for achieving the accuracy and reliability required in real-world applications like nutritional science.

8 Future Work

We plan to build on this work in several key directions:

1. **Expand the Dataset:** We will continue to grow JFB, increasing its size and diversity to cover a wider range of global cuisines and dietary contexts.
2. **Enhance the Benchmark:** We aim to improve the framework by incorporating more granular tasks, such as explicit portion size estimation [18] and food quality assessment.
3. **Release Improved Models:** The insights gained from this benchmark will inform the development of future iterations of our specialized models, with the goal of establishing new state-of-the-art performance on this benchmark.

9 Data, Code, and Model Availability

The full JFB dataset (images + JSON annotations), metric scripts, evaluation pipeline, and VLM prompts are available at https://github.com/January-ai/food-scan-benchmarks under a CC-BY-4.0 license.

10 Ethics & Privacy Statement

All images were collected with informed consent under the January AI Terms of Service and Privacy Policy. Personally identifying information, faces, and EXIF metadata were removed.

A `january/food-vision-v1` Model Context

The `january/food-vision-v1` model evaluated in this paper is a proprietary, commercial system. For the purposes of this benchmark, its inclusion is intended to:

1. **Establish a strong baseline:** Demonstrate the level of accuracy currently achievable on JFB with a model incorporating extensive domain-specific fine-tuning.
2. **Quantify the performance gap:** Provide a clear, quantitative measure of the difference between general-purpose VLMs and a specialized solution on this task.

To aid scientific comparison while protecting intellectual property, we disclose the following: the model uses a Vision Transformer backbone, processes images at a 1,024 × 1,024 pixel resolution, and was trained on a large, private dataset on the order of 10^5 images. This private training set is entirely separate from and has no overlap with the public JFB test set, ensuring a fair, held-out evaluation. Due to the proprietary nature of the system, further details regarding its specific architecture and hyperparameters are not disclosed.

B Cost Calculation

Cost per image for the general-purpose VLMs was calculated based on the token usage for each API call, using publicly available pricing tiers as of July 2025. The cost for `january/food-vision-v1` is a fixed rate per API call, reflecting a typical production deployment model.

C Pipeline Implementation Details

To manage API calls efficiently and avoid rate-limiting, the evaluation pipeline used an asynchronous worker system with a semaphore to control concurrency. The pipeline is designed to be robust to transient network errors and API failures, with built-in retry logic.

References

[1] S. M. Schembre et al. "Accuracy of a machine learning-based approach to estimate energy intake from food images". In: *Appetite* 167 (2021), p. 105615.

[2] F. Aguilar et al. "Food recognition: A review of the state-of-the-art and future challenges". In: *IEEE Signal Processing Magazine* 40.5 (2023), pp. 74–89.

[3] A. Radford et al. "Learning transferable visual models from natural language supervision". In: *Proceedings of the International Conference on Machine Learning (ICML)*. 2021, pp. 8748–8763.

[4] H. Liu et al. "Visual instruction tuning". In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2023.

[5] Y. Li et al. *Evaluating large multimodal models on fine-grained visual classification*. 2024. arXiv: 2401.08249.

[6] Z. Wei et al. "Fine-grained image analysis with deep learning: A survey". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44.12 (2022), pp. 8926–8945.

[7] Y. He et al. *How well do vision-language models understand food?* 2024. arXiv: 2402.12369.

[8] L. Bossard, M. Guillaumin, and L. Van Gool. "Food-101–mining discriminative components with random forests". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2014.

[9] J. Marin et al. "Recipe1M+: A dataset for learning cross-modal embeddings for cooking recipes and food images". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43.1 (2021), pp. 187–201.

[10] Y. Shen et al. *MEAL: A dataset and benchmark for meal-level food recognition and analysis*. 2023. arXiv: 2311.13942.

[11] M. A. Wethington et al. "A review of the evidence for the web-based and mobile-based tracking of dietary intake, physical activity, and weight management". In: *Journal of the American Medical Informatics Association* 27.9 (2020), pp. 1475–1484.

[12] Y. Matsuda, H. Hoashi, and K. Yanai. "Recognition of multiple-food images by detecting candidate regions". In: *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*. 2012.

[13] W. Min et al. "A survey on food computing". In: *ACM Computing Surveys* 55.8 (2023). Article 170, pp. 1–38.

[14] K. He, X. Zhang, S. Ren, and J. Sun. "Deep residual learning for image recognition". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 770–778.

[15] A. Dosovitskiy et al. "An image is worth 16x16 words: Transformers for image recognition at scale". In: *Proceedings of the International Conference on Learning Representations (ICLR)*. 2021.

[16] S. Singh et al. "Revisiting the importance of fine-tuning and self-training for food recognition". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 2024, pp. 5865–5875.

[17] C. Papaioannou et al. "Food2K: A large-scale food-2000 classification dataset". In: *Proceedings of the IEEE International Conference on Image Processing (ICIP)*. 2022, pp. 2736–2740.

[18] G. Ciocca et al. "Food recognition: A new dataset, experiments, and results". In: *IEEE Journal of Selected Topics in Signal Processing* 11.1 (2017), pp. 3–14.

[19] S. Jiang et al. "FoodNet: A multi-task deep learning framework for food recognition". In: *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*. 2019, pp. 1210–1215.

[20] J. Yang et al. "PFID: A new dataset and baseline for food portion size and food intake diary". In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2023.

[21] E. He et al. "Dofoo: A new dataset and baseline for food portion estimation". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 2023, pp. 248–257.

[22] W. Dai et al. "InstructBLIP: Towards general-purpose vision-language models with instruction tuning". In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2023.

[23] Y. Bang et al. *A multitask, multilingual, and multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity*. 2023. arXiv: 2302.04023.

[24] J. Park et al. "SafeFoodGPT: A benchmark for evaluating safety and hallucination in domain-specific multimodal food assistants". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 2025.

[25] C. Liu et al. "MM-SafetyBench: A comprehensive benchmark for safety evaluation of multimodal large language models". In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2023.

[26] Z. Yang et al. "The dawn of LMMs: A study on the toxicity of multimodal large language models". In: *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*. 2024.

[27] S. Mukherjee et al. “Estimating dataset size requirements for classifying DNA microarray data”. In: *Journal of Computational Biology* 10.2 (2003), pp. 119–142.

[28] Y. Zhang et al. “Evaluation of a decided sample size in machine learning applications”. In: *BMC Bioinformatics* 24 (2023). Article 48.

[29] S. P. Mohanty et al. “The Food Recognition Benchmark: Using deep learning to recognize food in images”. In: *Frontiers in Nutrition* 9 (2022), p. 875143.

[30] N. Reimers and I. Gurevych. “Sentence-BERT: Sentence embeddings using Siamese BERT-networks”. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2019, pp. 3982–3992.

[31] A. Awekar et al. “Fast and exact all-pairs semantic similarity search”. In: *Proceedings of the ACM SIGMOD International Conference on Management of Data*. 2009, pp. 139–150.

[32] OpenAI. *Pricing*. Online. Accessed: July 2025. 2025. URL: https://openai.com/pricing.

[33] Google. *Gemini API pricing*. Online. Accessed: July 2025. 2025. URL: https://ai.google.dev/pricing.

[34] M. Wang and J. Stanley. “Weighting components of a composite score using expert judgments”. In: *Psychological Methods* 25.3 (2020), pp. 363–377.

[35] Thomas L. Saaty. *The Analytic Hierarchy Process*. New York: McGraw-Hill, 1980.