

Article

Reproducible but Incoherent: A Reference Standard-Free Reliability Benchmark of LLM-Assisted Nutritional Estimation Across 20 Model Configurations

Jhony Fernando Lopez Torres ^{1,*} , Juan Gabriel Castañeda ²  and Edwin Rivas ³ ¹ Neptura, Libano 731040, Colombia² Facultad de Ciencias Humanas, Corporación Universitaria Minuto de Dios—UNIMINUTO, Bogotá 111021, Colombia; jucastaneda@uniminuto.edu³ Facultad de Ingeniería, Universidad Distrital Francisco José de Caldas, Bogotá 111711, Colombia; erivas@udistrital.edu.co

* Correspondence: neptura@41gu5.com

Abstract

Nutritional estimation pipelines built on large language models are entering biomarker research, but free-text meal descriptions have no ground truth, so outputs cannot be validated against a reference standard. What can be established without one is whether the same input returns the same number. We submitted 191 de-identified Spanish meal descriptions, byte-identically, three times each, at temperature 0, to 20 model configurations across two inference platforms (12,119 calls), and measured test–retest reliability by ICC(1,1), cross-model agreement, internal Atwater coherence between reported energy and macronutrients, and the reliability gradient across input specificity. All 20 models exceeded the audited cascade’s energy ICC under both comparators, but reproducibility did not imply internal coherence: the most reproducible model (ICC 1.0000) departed from its own macronutrients by more than 10% under the simplified 4/4/9 Atwater approximation in 65.58% of responses, against 3.80% for the cascade. The two axes are close to orthogonal, and the dissociation needs no threshold to state: that configuration ranks first of 20 in reliability and first of 20 in incoherence at every bound between 5% and 20%. A binary acceptance gate is not robust to that bound—3 to 17 of 20 configurations pass across the range—whereas the ordering is (Kendall tau ≥ 0.74 between any two), so coherence is reported as a ranked screening dimension with a decision curve, not as a validated criterion. Test–retest reliability is therefore necessary but not sufficient to authorise an LLM-assisted nutritional pipeline, and acceptance requires at least one reference standard-free coherence check alongside the reliability estimate.

Keywords: reproducibility of results; large language models; diet records; nutrition assessment; observer variation



Academic Editor: Guoqing Chao

Received: 19 August 2026

Revised: 30 September 2026

Accepted: 7 October 2026

Published: 15 October 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Dietary exposure is among the hardest variables to measure in metabolic epidemiology. Self-reported intake carries well-characterised misreporting that varies with body composition and assessment mode [1,2], and the resulting measurement error attenuates observed diet–outcome associations toward the null in ways that require explicit correction [3,4]—a problem the nutrition epidemiology literature has debated at length without resolving it in favour of abandoning self-report [5]. Free-text meal logging with automated

nutrient estimation is an attractive way to collect intake at scale, and a substantial body of literature now evaluates image- and app-based dietary assessment against conventional instruments [6–8]; a more recent strand evaluates large language models directly as nutrient estimators from meal photographs [9] and from text-only descriptions [10]. These component-level evaluations sit inside a broader, system-level integration of artificial intelligence across food and nutrition science, recently reviewed as a set of computational paradigms rather than as isolated tools [11]; the reliability of any single component is a precondition for that integration, and it is what we address here.

Large language models have moved quickly into that role, and into clinical information tasks more broadly, where evaluation frameworks are still consolidating [12,13]. Two findings from the literature bear directly on nutrient estimation. First, model responses to clinically substantive questions are inconsistent across repetitions and across model versions [14–16], and the variability is large enough that authors evaluating medical answers now report it as a primary property rather than a nuisance [17–19]. Second, models produce confidently stated content that is factually wrong [20,21], which is why prompting strategies that sample repeatedly and aggregate—self-consistency and its relatives—improve accuracy on reasoning tasks [22–24]. Those strategies are, in effect, an admission that a single call is not a stable measurement.

The difficulty specific to nutrient estimation is the absence of a reference standard. A logged description such as *arroz con pollo* has no true energy content: portion size, preparation and composition are all unobserved. Validation against a food composition table answers a different question—whether the mapping is plausible—and cannot detect an estimator that is plausible but unstable. Reliability, by contrast, is estimable without ground truth: resubmit the same input and compare. That is standard practice for digital measurement instruments [25,26] and rests on well-established estimators [27,28] whose interval behaviour under unbalanced designs is itself a subject of methodological work [29].

Two prior articles in this programme established the framework. The first [30] audited a deployed cascade and found its energy reliability moderate, with a marked gradient across input specificity.

The second, a manuscript in preparation by the same authors, derived the operating characteristics of the acceptance protocol, showing that its sample-size floor is insufficient near the decision threshold and that the interval rule loses nominal error control when replicate counts are dispersed. Both left the comparative question unanswered: the audited cascade was measured against a threshold, never against alternatives.

This article closes that gap and tests an assumption implicit in the protocol—that reliability is the aspect that matters. We report a benchmark of 20 model configurations on the same corpus and protocol as the audited cascade, adding an internal physical coherence axis requiring no reference standard. Reporting follows ADEMP for the simulation-derived components; we note that reporting guidance specific to artificial intelligence studies in health is converging on the same demands for version pinning and traceable figures [31,32]. The central result is a dissociation: the most reproducible configuration in the study is among the least physically coherent, which makes a reliability-only verdict not merely incomplete but actively misleading.

2. Materials and Methods

2.1. Design and Reporting

This is a test–retest reliability study of nutrient estimation by large language models, reported under ADEMP for the simulation-derived components [33]; reporting also follows the version pinning and traceability demands common to artificial intelligence reporting checklists [31,32]. GRRAS [34] does not apply: there are no human observers and no

reference standard. The estimand is the reliability of a deployed configuration—model, prompt, temperature and serving endpoint jointly—not of a model in the abstract.

The design is deliberately asymmetric to the usual benchmark. We do not ask whether an estimate is correct, because no ground truth exists for free-text meal descriptions in the wild. We ask whether the same input, resubmitted byte-identically, returns the same number. That question is answerable without a reference standard, and it is the question a deployment must answer before its outputs enter a biomarker model.

2.2. Corpus

All strings are Spanish free text, a setting where clinical natural language processing performance is known to differ from English [35]. The input corpus is 191 normalised free-text meal descriptions in Spanish, drawn from the production logs of the deployed nutrient estimation cascade audited in a companion manuscript [30] and de-identified before release. The parent programme holds ethics approval from the Research Ethics Committee of the Universidad Distrital Francisco José de Caldas, granted at programme level on 17 June 2024 (see Institutional Review Board Statement); the present study is a secondary analysis of records that programme generated in routine operation. Four strata were pre-declared: 1–2 words, 3–5 words, 6 or more words, and a fourth stratum of cascade filler labels—strings such as *comida variada* (“mixed meal”) that the cascade writes when it cannot parse the input. The filler stratum is not a length class; it is the degenerate input condition, and it is reported separately because pooling it with genuine descriptions conflates two different phenomena. The eight filler strings carry the corpus’s largest replicate counts and are, on their own, what pushes the replicate count coefficient of variation from 0.6187 to 0.8527.

2.3. Protocol

Each string was submitted $k = 3$ times per model at temperature 0, byte-identical across replicates, with one fixed prompt. Fixing the prompt is a deliberate scope restriction: prompt formulation is itself a large source of performance variation [36–38], so we hold it constant and estimate reliability within a configuration rather than across prompt space. Every call was recorded whether or not it parsed: the raw response text is the datum, and discarding unparsed calls would silently condition the estimand on the model’s formatting compliance. The frozen analysis file contains 12,119 calls.

Estimand and serving configuration. Because the estimand is the deployed configuration rather than the weights, the serving path is part of what was measured and is reported per model (Table 1). The 20 configurations were reached through two inference platforms: 16 through NVIDIA NIM at integrate.api.nvidia.com and 4 through the Anthropic Messages API. Temperature was 0 for every call in the main protocol. The provider-returned version string was logged for 15 of the 20 configurations and is released with the call log; for the remaining 5, the provider did not return one, so those configurations are pinned only by their model identifier and the acquisition timestamp. Three properties of the design are declared here because they bound what a replication can expect to recover. First, the completion token budget was not uniform: it was set per acquisition track and ranges from 200 to 2500 tokens across the 12 configurations for which it was logged, so configurations were not all measured under an identical output allowance. We tested whether this biased the estimates and found no detectable association with parse success (Spearman -0.33 , $p = 0.30$), corpus coverage (-0.05 , $p = 0.88$) or reliability (-0.04 , $p = 0.91$), but the inhomogeneity is real and we do not claim it is immaterial. Second, the prompt was held byte-identical across every call by construction, but the template text itself was not written to the per-call log; it is therefore not recoverable from the released data, and this is a repro-

ducibility limitation of the present release rather than a property of the design. Third, both platforms are commercial services whose routing, batching and checkpoint composition are outside our observation, so an exact re-run is not available to us either: what the released material supports is re-analysis of the recorded responses and re-collection under a declared configuration, not bit-level replication. (Table 1—artI_tab_mapeo_endpoints.csv).

Table 1. Serving configuration by model: platform, provider version string where returned, temperature and completion token budget. Source data: artI_tab_mapeo_endpoints.csv.

Configuration	Platform	Provider Version String	Temperature	Completion Token Budget	Calls	Parse Success	Corpus Coverage	Energy ICC
claude-haiku-4-5-20251001	Anthropic Messages API	claude-haiku-4-5-20251001	0	1000	573	0.9564	0.9529	0.9818
claude-opus-4-5-20251101	Anthropic Messages API	claude-opus-4-5-20251101	0	1000	573	0.9581	0.9529	0.9760
claude-sonnet-4-5-20250929	Anthropic Messages API	claude-sonnet-4-5-20250929	0	1000	573	0.6754	0.5812	0.9974
claude-sonnet-4-6	Anthropic Messages API	claude-sonnet-4-6	0	1000	573	0.7382	0.6021	0.9489
deepseek-ai/deepseek-v4-flash-0731	NVIDIA NIM (integrate.api.nvidia.com)	deepseek-ai/deepseek-v4-flash-0731	0		465	1	0.7906	0.9387
google/diffusiongemma-26b-a4b-it	NVIDIA NIM (integrate.api.nvidia.com)	google/diffusiongemma-26b-a4b-it	0		497	0.9879	0.6702	0.9607
meta/llama-3.1-8b-instruct	NVIDIA NIM (integrate.api.nvidia.com)	meta/llama-3.1-8b-instruct	0		573	0.9197	0.8743	0.9041
meta/llama-3.2-3b-instruct	NVIDIA NIM (integrate.api.nvidia.com)	meta/llama-3.2-3b-instruct	0		573	0.9668	0.9634	0.9978
mistralai/mistral-nemotron	NVIDIA NIM (integrate.api.nvidia.com)		0		573	1	1	0.9992
nvidia/ising-calibration-1.5-31b	NVIDIA NIM (integrate.api.nvidia.com)	nvidia/ising-calibration-1.5-31b	0	200	573	1	1	0.9978
nvidia/llama-3.1-nemotron-nano-v1-8b-v1	NVIDIA NIM (integrate.api.nvidia.com)		0		573	1	1	0.9896
nvidia/llama-3.3-nemotron-super-49b-v1	NVIDIA NIM (integrate.api.nvidia.com)	nvidia/llama-3.3-nemotron-super-49b-v1	0	200	573	0.8202	0.7120	0.8367
nvidia/nemotron-3-nano-30b-a3b	NVIDIA NIM (integrate.api.nvidia.com)	nvidia/nemotron-3-nano-30b-a3b	0	900	573	0.9738	0.9476	0.8920
nvidia/nemotron-3-nano-omni-30b-a3b-reasoning	NVIDIA NIM (integrate.api.nvidia.com)	nvidia/nemotron-3-nano-omni-30b-a3b-reasoning	0	200	573	0.9808	0.9529	0.7716
nvidia/nemotron-3-super-120b-a12b	NVIDIA NIM (integrate.api.nvidia.com)	nvidia/nemotron-3-super-120b-a12b	0	900	573	0.9825	0.9738	0.8977
nvidia/nemotron-3-ultra-550b-a55b	NVIDIA NIM (integrate.api.nvidia.com)	nvidia/nemotron-3-ultra-550b-a55b	0	200	564	0.7161	0.5393	0.9790
nvidia/nemotron-3.5-lightning-30b-a3b	NVIDIA NIM (integrate.api.nvidia.com)		0	2500	573	0.9756	0.9267	0.7271
nvidia/nemotron-mini-4b-instruct	NVIDIA NIM (integrate.api.nvidia.com)		0		573	0.9948	0.9948	1
nvidia/nemotron-nano-12b-v2-v1	NVIDIA NIM (integrate.api.nvidia.com)		0		573	1	1	0.9879
poolside/laguna-xs-2.1	NVIDIA NIM (integrate.api.nvidia.com)	poolside/laguna-xs-2.1	0	200	530	0.9943	0.8429	0.9582

Of the models probed, 20 met the inclusion rule of at least 100 strings with three usable replicates; 4 did not and are reported with their reason. Reliability is ICC(1,1) in the Shrout–Fleiss sense [39], following the estimator conventions of McGraw and Wong [40] and the reporting guidance of Koo and Li [41], with the unbalanced design correction, and intervals are bootstrap percentile intervals resampling inputs, not rows—a choice that matters because interval behaviour degrades under unbalanced replicate counts [29,42]—resampling rows would treat replicates of the same string as independent and understate the interval. (Table 2—artI_marco_final_analysis.csv).

Table 2. Analysis frame: models included, models excluded and the reason for each. Source data: artI_marco_final_analysis.csv.

Model	Strings with $k \geq 3$	Usable Rows	In Analysis	Reason for Exclusion
mistralai/mistral-nemotron	191	573	Yes	—
nvidia/ising-calibration-1.5-31b	191	573	Yes	—
nvidia/llama-3.1-nemotron-nano-vl-8b-v1	191	573	Yes	—
nvidia/nemotron-nano-12b-v2-v1	191	573	Yes	—
nvidia/nemotron-mini-4b-instruct	190	570	Yes	—
nvidia/nemotron-3-super-120b-a12b	186	563	Yes	—
meta/llama-3.2-3b-instruct	184	554	Yes	—
claude-haiku-4-5-20251001	182	548	Yes	—
claude-opus-4-5-20251101	182	549	Yes	—
nvidia/nemotron-3-nano-omni-30b-a3b-reasoning	182	562	Yes	—
nvidia/nemotron-3-nano-30b-a3b	181	558	Yes	—
nvidia/nemotron-3.5-lightning-30b-a3b	177	559	Yes	—
meta/llama-3.1-8b-instruct	167	527	Yes	—
poolside/laguna-xs-2.1	161	527	Yes	—
deepseek-ai/deepseek-v4-flash-0731	151	465	Yes	—
nvidia/llama-3.3-nemotron-super-49b-v1	136	470	Yes	—
google/diffusiongemma-26b-a4b-it	128	491	Yes	—
claude-sonnet-4-6	115	423	Yes	—
claude-sonnet-4-5-20250929	111	387	Yes	—
nvidia/nemotron-3-ultra-550b-a55b	103	396	Yes	—
z-ai/glm-5.2	3	13	No	only 3 strings with usable $k \geq 3$ (threshold 100)
meta/llama-3.1-70b-instruct	0	321	No	only 0 strings with usable $k \geq 3$ (threshold 100)
meta/llama-3.2-11b-vision-instruct	0	297	No	only 0 strings with usable $k \geq 3$ (threshold 100)
meta/llama-3.3-70b-instruct	0	248	No	only 0 strings with usable $k \geq 3$ (threshold 100)

Usable replicates per cell. Three calls were issued for every model-by-string cell and no call was retried, so a cell holds at most three usable rows; it holds fewer when a call returned no parseable energy value. Across the 20 analysed configurations, 11,213 calls were logged and 10,441 (93.1%) yielded a parseable value; the losses are refusals, responses truncated at the token budget, and responses whose JSON could not be parsed, all of which are retained verbatim in the released log with their finish reason. Of the 3642 model-by-string cells, 3300 (90.6%) retain all three replicates, 199 (5.5%) retain two and 143 (3.9%) retain one. Only cells with at least three replicates enter the reliability estimates, which is the pre-declared inclusion rule; cells with one or two are reported in the coverage table rather than dropped silently. Because the retained cells therefore all have $k = 3$, the one-way ICC(1,1) is computed on balanced groups and the unequal group correction does not arise; where unequal group sizes do occur in the sensitivity analyses, the estimator uses the harmonic mean group size $k_0 = (N - \sum(n_i^2)/N)/(a - 1)$ in place of k , which is the standard one-way treatment. Deep probe. To separate within-cell dispersion from between-input variance, six strings were re-submitted to six models with 30 replicates requested per cell, giving 36 model-by-string cells. The strings were chosen to span all four pre-declared

strata—huevo and ensalada (1–2 words), arroz con pollo and arepa y huevos revueltos (3–5 words), dos panes y un vaso con cafe sin azucar (6 or more words), and comida variada (filler)—and the models to span both platforms, both providers and the size bands of the sampling frame: google/diffusiongemma-26b-a4b-it, meta/llama-3.2-3b-instruct, mistralai/mistral-nemotron, nvidia/llama-3.3-nemotron-super-49b-v1, nvidia/nemotron-3-ultra-550b-a55b and z-ai/glm-5.2. Requests were issued sequentially per cell with the same fixed prompt, temperature 0 and the token budget of the main protocol; 1008 responses were logged at temperature 0 and a further 154 at temperature 0.7 as a contrast (1162 calls in total). Of the temperature 0 responses, 944 yielded a parseable energy value, between 13 and 30 per cell; cells below 30 reflect refusals and truncations, not sampling.

Two quantities are reported from this probe, and they are not the same. A distinct value is a distinct parsed energy number within a cell: 17 of the 36 cells returned more than one, up to five. A distinct serialisation is a distinct raw response string that parses to a number already seen—the same figure written differently: 23 of the 36 cells produced at least one. The first measures numerical irreproducibility; the second measures formatting instability with the number held constant, and a pipeline with a stricter parser would be exposed to it.

2.4. The Four Axes

1. Reproducibility. ICC(1,1) per model and nutrient over byte-identical resubmissions.
2. Cross-model concordance. ICC of agreement and of consistency across models on the same strings, reported as two distinct quantities. Neither is “the” ICC: agreement penalises systematic offsets between models, consistency does not, and the gap between them is itself the result.
3. Atwater coherence. The relative discrepancy between reported energy and the energy implied by the reported macronutrients under the Atwater factors of 4/4/9 kcal per gram [43]. This is an internal consistency check, not a reference standard: it asks only whether a row agrees with itself. The 4/4/9 factors are conventional averages of metabolisable energy, so the relation is an approximation and not an identity. Specific Atwater factors differ by food (protein in eggs 4.36 kcal/g, in whole maize 2.73); alcohol contributes about 7 kcal/g and appears in none of the four fields; fibre sits inside total carbohydrate in most schemas yet yields roughly 2 kcal/g when fermented; and reported values are rounded. A small residual is therefore expected even from a correct estimator, which is why we report the whole distribution and a sensitivity curve over the bound rather than a single pass/fail verdict. Zero or absent energy is treated as undefined, never as coherent: a model returning all zeros would otherwise score perfectly.
4. Specificity gradient. Reliability by input stratum, testing whether more specific descriptions yield more reproducible estimates.

An acceptance verdict applies a threshold fixed in advance ($\text{ICC} \geq 0.75$ on every nutrient), and we report it both alone and with a second coherence gate, to quantify what a reliability-only protocol would authorise.

2.5. Baseline, and What It Can Fairly Be Compared to

The comparator is the production nutrient estimation cascade audited in a companion manuscript [30], whose operating characteristics—the imbalance correction and the sample size guidance applied here—were derived in a second companion manuscript by the same authors. Two properties of that system govern what the comparison can support, and both work against a naive ranking.

First, the audited object is not a model. It is a two-stage cascade of commercial large language models from two different providers—a vision-capable stage that renders photographs as Spanish text, and a text stage that maps descriptions to nutrients—reached through provider APIs. Its prompts, sampling temperature, model checkpoints and any composition table anchoring are set provider-side and revised without notice and were neither inspectable nor pinnable by the investigators [30] (§2.1 and limitation 9). Whatever reliability that cascade exhibits is therefore a property of third-party model behaviour observed through an integration, not of the integrating platform’s own engineering. The models benchmarked here are the same class of component that cascade consumes.

Second, the two designs are not equivalent, and the asymmetry favours the models. The cascade’s replicates are naturally occurring repetitions of identical strings, separated in time across a 179-day window, spanning unannounced provider-side version changes, and 48.3% of them carry an unstandardised photograph that traverses the vision stage as well, so the pooled cascade figure mixes two input modalities [30]. The models here were submitted $k = 3$ times within a single session, against a pinned model identifier, text-only, at temperature 0, with one fixed prompt. The benchmark therefore measures short-run sampling reproducibility under laboratory conditions, whereas the cascade figure additionally absorbs version drift and a second processing stage. Any gap between them is an upper bound on model-attributable difference, not an estimate of it.

We therefore use two baseline values and label them explicitly. The modality-matched comparator—the cascade’s text-only stratum, the only stratum whose input type matches the benchmark—is energy ICC(1,1) 0.7061 (95% CI 0.5419–0.8217; 85 strings, 352 meals). The pooled comparator, mixing text-only and photograph-accompanied entries, is 0.5905 at $k \geq 3$. The pooled value is the lower of the two precisely because it includes the photograph pathway, so comparing single-session text-only model calls against it would overstate the models’ advantage. The modality-matched value is the primary comparator here; the pooled value appears only where the companion manuscript’s own headline figure is referenced. We record here that the audit trail contains two defensible cuts of that deployment—a frozen cut of 191 repeated inputs and a reproducible cut of 246—which reproduce exactly from their own source but are not the same corpus. All baseline figures in this article are from the frozen cut; the reproducible cut is the pre-declared robustness specification. (Table 3—cascade_reliability_by_modality.csv).

Table 3. Reliability of the deployed cascade by input modality, the source of the modality-matched comparator. Source data: cascade_reliability_by_modality.csv, from the companion manuscript [30].

Input Modality	Nutrient	Strings	Meals	ICC(1,1)	95% CI
Pooled (all replicates)	Energy (kcal)	191	944	0.5905	0.4962–0.6866
Pooled (all replicates)	Protein (g)	191	944	0.5712	0.4849–0.6687
Pooled (all replicates)	Carbohydrate (g)	191	944	0.5876	0.4930–0.6834
Pooled (all replicates)	Fat (g)	191	944	0.6272	0.5116–0.7213
Pooled (all replicates)	Fibre (g)	191	944	0.6604	0.5676–0.7377
Pooled (all replicates)	Sugar (g)	191	944	0.5504	0.3483–0.7101
Text only	Energy (kcal)	85	352	0.7061	0.5419–0.8217
Text only	Protein (g)	85	352	0.7127	0.5432–0.8492
Text only	Carbohydrate (g)	85	352	0.7106	0.5650–0.8257
Text only	Fat (g)	85	352	0.7486	0.5119–0.8561
Text only	Fibre (g)	85	352	0.6985	0.5480–0.8331
Text only	Sugar (g)	85	352	0.6836	0.1013–0.8846
Photograph only	Energy (kcal)	82	427	0.5258	0.3778–0.6513

Table 3. *Cont.*

Input Modality	Nutrient	Strings	Meals	ICC(1,1)	95% CI
Photograph only	Protein (g)	82	427	0.4579	0.3336–0.5982
Photograph only	Carbohydrate (g)	82	427	0.5227	0.3778–0.6733
Photograph only	Fat (g)	82	427	0.4887	0.3378–0.6314
Photograph only	Fibre (g)	82	427	0.6289	0.4544–0.7417
Photograph only	Sugar (g)	82	427	0.5830	0.3021–0.7921

2.6. Verification

Every figure in this article is recalculated from its source artefact by script, with the artefact identifier, column and value recorded in a verification table (Table 4—artI_verificacion_cifras.csv). Figures that disagreed with an earlier draft were corrected in favour of the source and logged. Central findings each carry a pre-declared alternative specification. The verification script, the source artefacts named in Table 4 and the recomputation log are all contained in the replication package provided as Supplementary Materials (artI-repra-bench-v1.zip).

Table 4. Figure verification: every published figure with its source artefact, column and recalculated value. Source data: artI_verificacion_cifras.csv. Column expressions are reproduced verbatim as the verification script evaluates them.

ID	Section	Quantity	Published Value	Source Artefact	Column	Recalculated	Match
H1.1	Results	ICC(1,1) energy of the deployed cascade (baseline, $k \geq 3$, frozen cutoff)	0.5905	artI_linea_base_delfos.csv	icc [corte = congelado_13073, nutriente = energia, k_min = 3]	0.5905	Yes
H1.2	Results	ICC minimum energy across the 20 models	0.7271	artI_tab_icc_por_modelo.csv	icc [nutriente = kcal, min]	0.7271	Yes
H1.3	Results	ICC maximum energy across the 20 models	1	artI_tab_icc_por_modelo.csv	icc [nutriente = kcal, max]	1	Yes
H1.4	Results	ICC median energy across the 20 models	0.9684	artI_tab_icc_por_modelo.csv	icc [nutriente = kcal, median]	0.9684	Yes
H1.5	Results	Models with ICC energy above the baseline	20	artI_tab_icc_por_modelo.csv	count(icc > 0.5905)	20	Yes
H1.6	Results	Lower limit of the CI of the least reliable model	0.5812	artI_tab_icc_por_modelo.csv	ic_lo [argmin icc]	0.5812	Yes
H2.1	Results	ICC energy of nemotron-mini-4b-instruct	1	artI_tab_icc_por_modelo.csv	icc [modelo = nvidia/nemotron-mini-4b-instruct, nutriente = kcal]	1	Yes
H2.2	Results	Fraction of nemotron-mini-4b-instruct responses with Atwater incoherence >0.10	0.6558	artI_tab_atwater_sensibilidad.csv	frac_gt_010 [modelo = nvidia/nemotron-mini-4b-instruct]	0.6558	Yes
H2.3	Results	Median Atwater incoherence of the deployed cascade	0.0211	artI_atwater_delfos.csv	mediana [fila 1]	0.0211	Yes
H2.4	Results	Models rejected by the coherence gate despite passing reliability	8	artI_tab_veredicto_aceptacion.csv	count(veredicto_doble)	8	Yes
H2.5	Results	Models that pass both gates	11	artI_tab_veredicto_aceptacion.csv	count(veredicto_doble = PASA)	11	Yes
H3.1	Results	ICC agreement between models	0.6865	artI_tab_concordancia_modelos.csv	acuerdo	0.6865	Yes
H3.2	Results	Lower limit of the 95% CI of cross-model agreement	0.5990	artI_tab_concordancia_modelos.csv	acuerdo_lo	0.5990	Yes
H3.3	Results	Upper limit of the 95% CI of cross-model agreement	0.7563	artI_tab_concordancia_modelos.csv	acuerdo_hi	0.7563	Yes
H3.4	Results	ICC consistency between models	0.7056	artI_tab_concordancia_modelos.csv	consistencia	0.7056	Yes

Table 4. Cont.

ID	Section	Quantity	Published Value	Source Artefact	Column	Recalculated	Match
H3.5	Results	Strings used in the concordance matrix (listwise)	82	artI_tab_concordancia_modelos.csv	n_cadenas	82	Yes
H4.1	Results	Deployed cascade gradient: ICC energy in 1–2 words	0.5705	artI_linea_base_estratos.csv	icc [estrato = 1–2, energia, k_min = 3]	0.5705	Yes
H4.2	Results	Gradient of the deployed cascade: ICC energy in ≥ 6 words	0.9181	artI_linea_base_estratos.csv	icc [estrato = ≥ 6 , energia, k_min = 3]	0.9181	Yes
H4.3	Results	Delta of the deployed cascade gradient between extreme strata	0.3476	artI_linea_base_estratos.csv	icc [≥ 6] – icc [1–2]	0.3476	Yes
H4.4	Results	Models with a positive gradient, like the deployed cascade	0.0000	artI_robustez_gradiente.csv	count(clasificacion)	0.0000	Yes
H4.5	Results	Models with flat gradient	15	artI_robustez_gradiente.csv	count(clasificacion = plano)	15	Yes
H4.6	Results	Models with negative gradient	5	artI_robustez_gradiente.csv	count(clasificacion = negativo)	5	Yes
H4.7	Results	ICC of the filler stratum in the deployed cascade	0.1028	artI_linea_base_estratos.csv	icc [estrato = relleno]	0.1028	Yes
H5.1	Results	Fraction of cells (model $\times$ string) irreproducible at temperature 0	0.4722	artI_sondeo_profundo_resumen.csv	frac_celdas_irreproducibles	0.4722	Yes
H5.2	Results	Maximum of distinct values in a cell at temperature 0	5	artI_sondeo_profundo_resumen.csv	max_valores_distintos	5	Yes
H5.3	Results	Cells with the same value returned in different format	23	artI_sondeo_profundo_resumen.csv	celdas_mismo_valor_otro_formato	23	Yes
H5.4	Results	Mean number of distinct values at T = 0 (cells with both temperatures)	2	artI_sondeo_profundo_resumen.csv	media_distintos_t0	2	Yes
H5.5	Results	Mean number of distinct values at T = 0.7 (same cells)	5	artI_sondeo_profundo_resumen.csv	media_distintos_t07	5	Yes
M1	Methods	Recorded calls in the consolidated benchmark	12,119	artI_benchmark_consolidado.csv	nrow	12,119	Yes
M2	Methods	Models in the main analysis	20	artI_marco_final_analisis.csv	sum(en_analisis)	20	Yes
M3	Methods	Models excluded due to insufficient coverage	4	artI_marco_final_analisis.csv	sum(~en_analisis)	4	Yes
M4	Methods	Corpus strings admitted	191	artI_corpus_aprobado_191.csv	nrow	191	Yes
H1.7	Results	ICC median energy across the 20 models (table by nutrient)	0.9607	artI_tab_icc_por_nutriente.csv	median(kcal)	0.9607	Yes
H1.8	Results	ICC median protein	0.9783	artI_tab_icc_por_nutriente.csv	median(prot)	0.9783	Yes
H1.9	Results	ICC median carbohydrate	0.9587	artI_tab_icc_por_nutriente.csv	median(carb)	0.9587	Yes
H1.10	Results	ICC median fat	0.9618	artI_tab_icc_por_nutriente.csv	median(fat)	0.9618	Yes
H2.6	Results	Fraction of deployed cascade records with Atwater incoherence >0.10	0.0380	artI_atwater_delfos.csv	frac_gt_010 [fila 1]	0.0380	Yes
H2.7	Results	Strings with zero within-string variance in the model with perfect ICC	190	artI_tab_icc_por_modelo.csv	cadenas [modelo = nvidia/nemotron-mini-4b-instruct, nutriente = kcal]	190	Yes
H2.8	Results	Models that pass the reliability gate alone	19	artI_tab_veredicto_aceptacion.csv	count(veredicto = PASA)	19	Yes
M5	Methods	CV of full corpus replicates	0.8527	artI_distribucion_replicas.csv	cv_k [corte = congelado_13073, corpus = completo]	0.8527	Yes
M6	Methods	Replicate CV excluding filler labels	0.6187	artI_distribucion_replicas.csv	cv_k [corte = congelado_13073, corpus = sin_relleno]	0.6187	Yes
H2.9	Results	Models with better Atwater coherence than the deployed cascade	0.0000	artI_tab_atwater_sensibilidad.csv	count(frac_gt_010 < 0.037992)	0.0000	Yes

Table 4. Cont.

ID	Section	Quantity	Published Value	Source Artefact	Column	Recalculated	Match
H6.1	Methods and Results	ICC energy of the deployed cascade, text-only stratum (matched modality comparator)	0.7061	cascade_reliability_by_modality.csv	ICC [estrato = Solo texto, nutriente = Energía (kcal)]	0.7061	Yes
H6.2	Results	Lower limit of the 95% CI of the text-only comparator	0.5419	cascade_reliability_by_modality.csv	IC95_lo [estrato = Solo texto, nutriente = Energía (kcal)]	0.5419	Yes
H6.3	Results	Upper limit of the 95% CI of the text-only comparator	0.8217	cascade_reliability_by_modality.csv	IC95_hi [estrato = Solo texto, nutriente = Energía (kcal)]	0.8217	Yes
H6.4	Results and Limitations	Models whose CI overlaps that of the text-only comparator	5	artI_tab_icc_por_modelo.csv	count(ic_lo $\leq$ 0.8217 & ic_hi $\geq$ 0.5419)	5	Yes
H6.5	Discussion	Median Atwater incoherence across models (fraction > 0.10)	0.1818	artI_tab_atwater_sensibilidad.csv	median(frac_gt_010)	0.1818	Yes
H6.6	Methods and Limitations	Fraction of deployed cascade records with an attached photograph (6,320 of 13,073)	0.4830	cascade_corpus_characteristics.csv	Records with an attached photograph	0.4830	Yes
H6.7	Results	Models whose CI overlaps the matched modality comparator	5	artI_tab_icc_por_modelo.csv	names(ic_lo $\leq$ 0.8217 & ic_hi $\geq$ 0.5419)	5	Yes
R2.1	Methods and Results	Deep-probe cells with more than one distinct energy value	17	artI_sondeo_t0_por_celda.csv	count(distintos > 1)	17	Yes
R2.2	Methods and Results	Cells with more distinct serialisations than distinct values (same number, different format)	23	artI_sondeo_t0_por_celda.csv	count(crudos > distintos)	23	Yes
R2.3	Methods	Deep-probe cells = 6 models $\times$ 6 strings	36	artI_sondeo_t0_por_celda.csv	nrow	36	Yes
R2.4	Methods	Parsed energy observations in the temperature-0 probe	944	artI_sondeo_t0_por_celda.csv	sum(n)	944	Yes
R2.5	Discussion	Gap between cross-model ICC of agreement and ICC of consistency	0.0191	artI_tab_concordancia_modelos.csv	brecha_acuerdo_consistencia	0.0191	Yes
R2.6	Results, Section 3.1	Models passing the dual gate at a 5% coherence bound	3	artI_tab_puerta_sensibilidad.csv	sum(pasa_doble_u05)	3	Yes
R2.7	Results, Section 3.1	Models passing the dual gate at a 20% coherence bound	17	artI_tab_puerta_sensibilidad.csv	sum(pasa_doble_u20)	17	Yes
R2.9	Results	Overlap of the five least coherent models between the 5% and 20% bounds	5	artI_tab_atwater_sensibilidad.csv	top5(frac_gt_005) $\cap$ top5(frac_gt_020)	5	Yes
R2.10	Results and Discussion	Rank of nemotron-mini-4b by incoherence at the 20% bound (rank 1 in reliability)	1	artI_tab_atwater_sensibilidad.csv	rank(frac_gt_020, desc) [nvidia/nemotron-mini-4b-instruct]	1	Yes
R2.11	Results	Spearman correlation between ICC and the fraction of strings with identical replicates	0.8710	artI_tab_repetibilidad_absoluta.csv	spearman(icc, frac_identicas)	0.8710	Yes
R2.13	Results	Maximum ICC change when restricting to strings present in at least 16 models (174 of 191)	0.0431	artI_tab_entradas_comunes_graduado.csv	delta_max [min_modelos = 16]	0.0431	Yes
R2.14	Results, Section 3.1	Minimum parse rate across the 20 analysed models	0.6754	artI_tab_cobertura_y_entradas_comunes.csv	min(tasa_parseo)	0.6754	Yes

Table 4. *Cont.*

ID	Section	Quantity	Published Value	Source Artefact	Column	Recalculated	Match
R2.15	Results, Section 3.1	Models passing the reliability gate when the lower CI bound is required instead of the point estimate	17	artI_tab_puerta_sensibilidad.csv	sum(pasa_icc_ic_lo)	17	Yes
R2.16	Results, Section 3.1	Fraction of nemotron-3-ultra responses incoherent above 10% (corrected row)	0.1094	artI_tab_atwater_sensibilidad.csv	frac_gt_010 [modelo = nvidia/nemotron-3-ultra-550b-a55b]	0.1094	Yes

2.7. Declaration of Generative AI Use

A separate disclosure in the Acknowledgments records that generative AI assistants were used to draft and debug the analysis code and to edit prose. The two statements are not in conflict, and we state the boundary explicitly: assistants were used to write code, and that code—released in full—was then executed deterministically over the logged raw responses. No assistant was placed in the path that produced, parsed or aggregated any reported number, and no reported value depends on an assistant’s output at run time. Every figure in this article is recomputed from the released call log by the verification script, which is what makes the distinction checkable rather than declarative.

3. Results

Every model’s point estimate exceeds the cascade’s under the study conditions (Figure 1). Across the 20 models, energy ICC(1,1) ranges from 0.7271 (nemotron-3.5-lightning-30b-a3b) to 1.0000 (nemotron-mini-4b-instruct), with a median of 0.9684. All 20 point estimates exceed both cascade comparators—the modality-matched text-only stratum (0.7061) and the pooled figure (0.5905). The separation is not established for every model, however. Taking the modality-matched interval (0.5419–0.8217) as the reference, 5 of 20 model intervals overlap it: nemotron-3.5-lightning-30b-a3b (0.7271, CI 0.5812–0.8770), nemotron-3-nano-omni-30b-a3b-reasoning (0.7716, CI 0.7148–0.8194), llama-3.3-nemotron-super-49b-v1 (0.8367, CI 0.6430–0.9823), nemotron-3-nano-30b-a3b (0.8920, CI 0.7532–0.9528) and claude-sonnet-4-6 (0.9489, CI 0.8191–0.9804). For those five, the difference from the deployed cascade is not resolved at this sample size. For the remaining 15, the intervals do not overlap, which is not a formal test of difference and should not be read as one; no hypothesis test between the cascade and any model is reported here, because the two were not measured under comparable conditions. The cascade sits at the low end of the distribution under a design that grants the models a pinned version, a single session and text-only input, none of which the cascade had. It is a contextual comparator, not a competitor measured on equal terms.

Reliability is homogeneous across nutrients—median ICC 0.9607 (energy), 0.9783 (protein), 0.9587 (carbohydrate), 0.9618 (fat)—unlike the cascade, where energy was the critical field. (Table 5—artI_tab_icc_por_modelo.csv).

Table 5. Reliability by model and nutrient: ICC(1,1), bootstrap 95% interval, strings and observations. Source data: artI_tab_icc_por_modelo.csv. ICC(1,1) with a 95% interval from 2000 bootstrap resamples over input strings.

Model	Nutrient	ICC(1,1)	95% CI	Strings	Observations
nvidia/llama-3.3-nemotron-super-49b-v1	Energy (kcal)	0.8367	0.6430–0.9823	136	408
nvidia/llama-3.3-nemotron-super-49b-v1	Protein (g)	0.9545	0.9130–0.9858	136	408
nvidia/llama-3.3-nemotron-super-49b-v1	Carbohydrate (g)	0.8392	0.6462–0.9831	136	408
nvidia/llama-3.3-nemotron-super-49b-v1	Fat (g)	0.8823	0.7574–0.9664	136	408

Table 5. Cont.

Model	Nutrient	ICC(1,1)	95% CI	Strings	Observations
meta/llama-3.1-8b-instruct	Energy (kcal)	0.9041	0.8316–0.9695	167	501
meta/llama-3.1-8b-instruct	Protein (g)	0.9363	0.8888–0.9789	167	501
meta/llama-3.1-8b-instruct	Carbohydrate (g)	0.9267	0.8849–0.9685	167	501
meta/llama-3.1-8b-instruct	Fat (g)	0.8305	0.6296–0.9651	167	501
nvidia/nemotron-nano-12b-v2-v1	Energy (kcal)	0.9879	0.9778–0.9949	191	573
nvidia/nemotron-nano-12b-v2-v1	Protein (g)	0.9821	0.9684–0.9916	191	573
nvidia/nemotron-nano-12b-v2-v1	Carbohydrate (g)	0.9883	0.9781–0.9954	191	573
nvidia/nemotron-nano-12b-v2-v1	Fat (g)	0.9653	0.9442–0.9811	191	573
meta/llama-3.2-3b-instruct	Energy (kcal)	0.9978	0.9915–1.0000	184	552
meta/llama-3.2-3b-instruct	Protein (g)	0.9963	0.9891–0.9999	184	552
meta/llama-3.2-3b-instruct	Carbohydrate (g)	0.9987	0.9968–0.9997	184	552
meta/llama-3.2-3b-instruct	Fat (g)	0.9928	0.9773–0.9994	184	552
nvidia/nemotron-mini-4b-instruct	Energy (kcal)	1.0000	1.0000–1.0000	190	570
nvidia/nemotron-mini-4b-instruct	Protein (g)	0.9991	0.9968–1.0000	190	570
nvidia/nemotron-mini-4b-instruct	Carbohydrate (g)	0.9999	0.9995–1.0000	190	570
nvidia/nemotron-mini-4b-instruct	Fat (g)	0.9986	0.9947–0.9999	190	570
mistralai/mistral-nemotron	Energy (kcal)	0.9992	0.9977–0.9999	191	573
mistralai/mistral-nemotron	Protein (g)	0.9994	0.9987–0.9999	191	573
mistralai/mistral-nemotron	Carbohydrate (g)	0.9986	0.9964–0.9997	191	573
mistralai/mistral-nemotron	Fat (g)	0.9980	0.9951–0.9995	191	573
nvidia/ising-calibration-1.5-31b	Energy (kcal)	0.9978	0.9944–0.9995	191	573
nvidia/ising-calibration-1.5-31b	Protein (g)	0.9949	0.9902–0.9979	191	573
nvidia/ising-calibration-1.5-31b	Carbohydrate (g)	0.9952	0.9899–0.9981	191	573
nvidia/ising-calibration-1.5-31b	Fat (g)	0.9970	0.9931–0.9987	191	573
nvidia/llama-3.1-nemotron-nano-v1-8b-v1	Energy (kcal)	0.9896	0.9671–0.9984	191	573
nvidia/llama-3.1-nemotron-nano-v1-8b-v1	Protein (g)	0.9788	0.9442–0.9965	191	573
nvidia/llama-3.1-nemotron-nano-v1-8b-v1	Carbohydrate (g)	0.9902	0.9752–0.9981	191	573
nvidia/llama-3.1-nemotron-nano-v1-8b-v1	Fat (g)	0.9773	0.9406–0.9961	191	573
nvidia/nemotron-3.5-lightning-30b-a3b	Energy (kcal)	0.7271	0.5812–0.8770	177	531
nvidia/nemotron-3.5-lightning-30b-a3b	Protein (g)	0.8178	0.7224–0.8947	177	531
nvidia/nemotron-3.5-lightning-30b-a3b	Carbohydrate (g)	0.7451	0.6322–0.8881	177	531
nvidia/nemotron-3.5-lightning-30b-a3b	Fat (g)	0.8152	0.7006–0.8965	177	531
nvidia/nemotron-3-nano-30b-a3b	Energy (kcal)	0.8920	0.7532–0.9528	181	543
nvidia/nemotron-3-nano-30b-a3b	Protein (g)	0.8342	0.7033–0.9266	181	543
nvidia/nemotron-3-nano-30b-a3b	Carbohydrate (g)	0.8933	0.8005–0.9476	181	543
nvidia/nemotron-3-nano-30b-a3b	Fat (g)	0.8577	0.6561–0.9550	181	543
nvidia/nemotron-3-nano-omni-30b-a3b-reasoning	Energy (kcal)	0.7716	0.7148–0.8194	182	546
nvidia/nemotron-3-nano-omni-30b-a3b-reasoning	Protein (g)	0.8258	0.7605–0.8825	182	546
nvidia/nemotron-3-nano-omni-30b-a3b-reasoning	Carbohydrate (g)	0.8000	0.7464–0.8528	182	546
nvidia/nemotron-3-nano-omni-30b-a3b-reasoning	Fat (g)	0.7898	0.7245–0.8334	182	546
nvidia/nemotron-3-super-120b-a12b	Energy (kcal)	0.8977	0.8519–0.9331	186	558
nvidia/nemotron-3-super-120b-a12b	Protein (g)	0.9418	0.8949–0.9710	186	558
nvidia/nemotron-3-super-120b-a12b	Carbohydrate (g)	0.9102	0.8644–0.9484	186	558
nvidia/nemotron-3-super-120b-a12b	Fat (g)	0.8460	0.7357–0.9201	186	558
poolside/laguna-xs-2.1	Energy (kcal)	0.9582	0.9163–0.9826	161	483
poolside/laguna-xs-2.1	Protein (g)	0.9746	0.9552–0.9879	161	483
poolside/laguna-xs-2.1	Carbohydrate (g)	0.9502	0.9144–0.9813	161	483

Table 5. Cont.

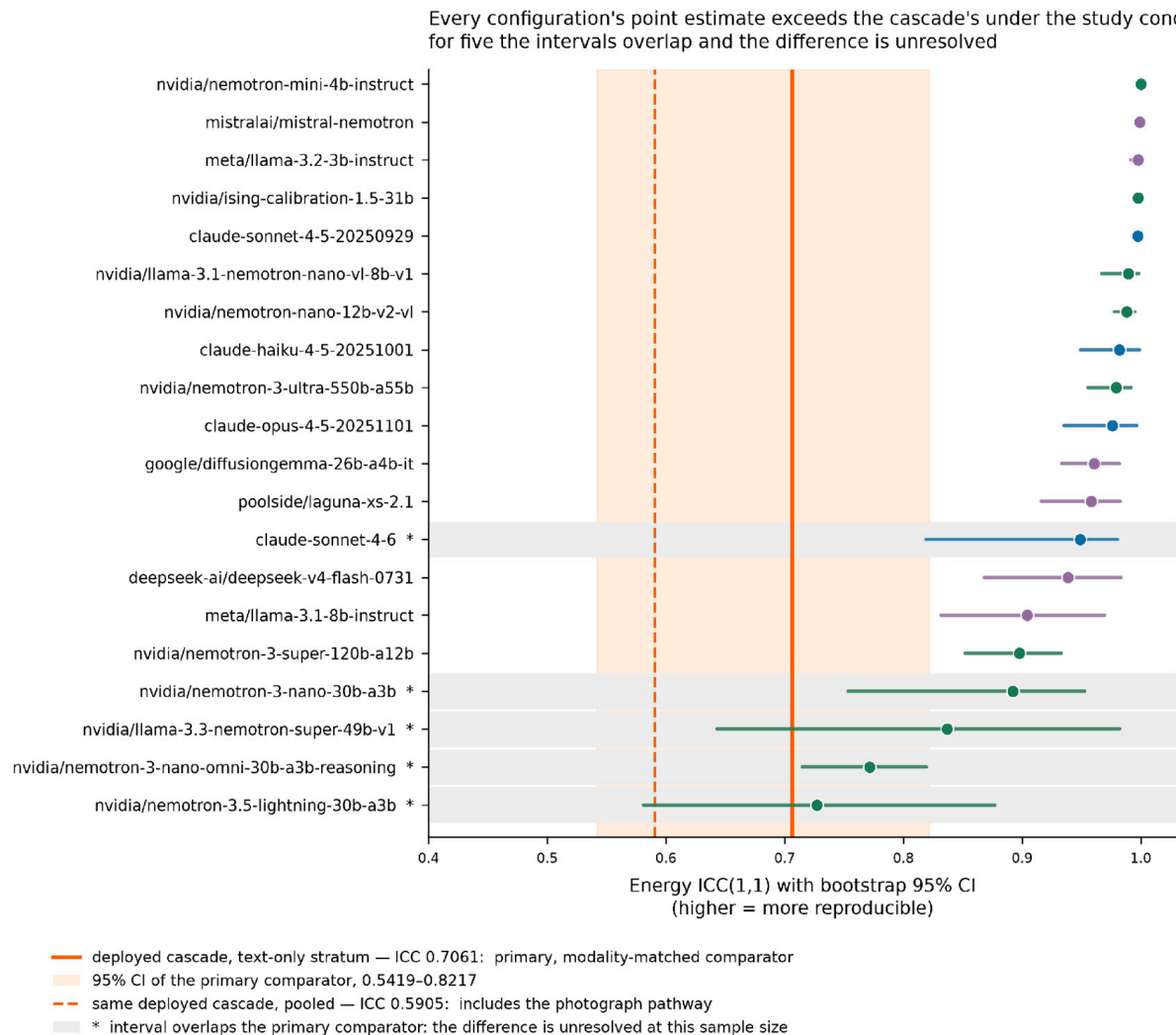
Model	Nutrient	ICC(1,1)	95% CI	Strings	Observations
poolside/laguna-xs-2.1	Fat (g)	0.9358	0.8942–0.9647	161	483
deepseek-ai/deepseek-v4-flash-0731	Energy (kcal)	0.9387	0.8682–0.9833	151	453
deepseek-ai/deepseek-v4-flash-0731	Protein (g)	0.9715	0.9333–0.9903	151	453
deepseek-ai/deepseek-v4-flash-0731	Carbohydrate (g)	0.9431	0.8869–0.9837	151	453
deepseek-ai/deepseek-v4-flash-0731	Fat (g)	0.9596	0.9270–0.9809	151	453
google/diffusionngemma-26b-a4b-it	Energy (kcal)	0.9607	0.9333–0.9819	128	384
google/diffusionngemma-26b-a4b-it	Protein (g)	0.9855	0.9757–0.9925	128	384
google/diffusionngemma-26b-a4b-it	Carbohydrate (g)	0.9587	0.9207–0.9881	128	384
google/diffusionngemma-26b-a4b-it	Fat (g)	0.9618	0.9382–0.9786	128	384
claude-haiku-4-5-20251001	Energy (kcal)	0.9818	0.9493–0.9991	182	546
claude-haiku-4-5-20251001	Protein (g)	0.9783	0.9412–0.9993	182	546
claude-haiku-4-5-20251001	Carbohydrate (g)	0.9895	0.9716–0.9989	182	546
claude-haiku-4-5-20251001	Fat (g)	0.9888	0.9743–0.9972	182	546
claude-opus-4-5-20251101	Energy (kcal)	0.9760	0.9351–0.9968	182	546
claude-opus-4-5-20251101	Protein (g)	0.9960	0.9900–0.9991	182	546
claude-opus-4-5-20251101	Carbohydrate (g)	0.9659	0.9283–0.9910	182	546
claude-opus-4-5-20251101	Fat (g)	0.9845	0.9586–0.9950	182	546
claude-sonnet-4-5-20250929	Energy (kcal)	0.9974	0.9928–0.9998	111	333
claude-sonnet-4-5-20250929	Protein (g)	0.9979	0.9942–0.9999	111	333
claude-sonnet-4-5-20250929	Carbohydrate (g)	0.9958	0.9898–0.9992	111	333
claude-sonnet-4-5-20250929	Fat (g)	0.9981	0.9962–0.9994	111	333
claude-sonnet-4-6	Energy (kcal)	0.9489	0.8191–0.9804	115	345
claude-sonnet-4-6	Protein (g)	0.9598	0.9050–0.9886	114	342
claude-sonnet-4-6	Carbohydrate (g)	0.9392	0.8385–0.9767	114	342
claude-sonnet-4-6	Fat (g)	0.9382	0.8414–0.9813	113	339
nvidia/nemotron-3-ultra-550b-a55b	Energy (kcal)	0.9790	0.9554–0.9917	103	309
nvidia/nemotron-3-ultra-550b-a55b	Protein (g)	0.9826	0.9574–0.9949	103	309
nvidia/nemotron-3-ultra-550b-a55b	Carbohydrate (g)	0.9637	0.9208–0.9908	103	309
nvidia/nemotron-3-ultra-550b-a55b	Fat (g)	0.9808	0.9604–0.9913	103	309

3.1. Reproducibility Does Not Imply Internal Coherence

The headline result is a dissociation. nemotron-mini-4b-instruct attains an energy ICC of 1.0000—variance between replicates of the same string is exactly zero across 190 strings—while 0.6558 of its responses depart from their own macronutrients by more than 10% under the simplified Atwater approximation. It reproduces the same internally inconsistent number every time. The cascade, by contrast, has a median Atwater discrepancy of 0.0211 and violates the 10% bound in 3.80% of records: less reproducible, more coherent.

This is not an isolated case (Figure 2), but the verdict depends on where the coherence bound is set and we did not pre-register that bound. Applying the pre-declared reliability threshold alone, 19 of 20 configurations pass on the point estimate and 17 pass if the lower confidence bound is required instead. Adding a coherence screen at 20% of responses above a 10% departure, 11 pass and eight fail on coherence alone—including the two most reproducible configurations in the study. Sweeping that departure bound from 5% to 20% moves the count from three to 17 of 20 (Figure 3a), so the binary verdict is not a validated criterion, and we do not present it as one. What is stable across the same range is the ordering: Kendall tau is at least 0.74 between any two bounds, the median configuration shifts three rank positions of 20, and the five least coherent configurations are the same five at both ends (Figure 3b). We therefore report coherence as a ranked screening dimension

with an explicit decision curve, leaving the bound to the deploying team, and note that a reliability-only protocol would authorise a configuration whose macronutrient arithmetic contradicts its own reported energy in two thirds of its answers at any bound in the range. (Table 6—artI_tab_atwater_sensibilidad.csv), (Table 7—artI_tab_veredicto_aceptacion.csv), (Table 8—artI_tab_puerta_sensibilidad.csv).



Non-comparable conditions: the 20 configurations were measured in a single session, pinned version, text-only; the deployed cascade over 179 days, unpinned versions, 48.3% photographs. Interval overlap is not a formal test.

Figure 1. Test–retest reproducibility of energy estimation for the 20 model configurations, against both comparators from the deployed two-stage cascade. Points are ICC(1,1) for energy and horizontal bars are bootstrap 95% CI resampled over input strings; rows are ordered by decreasing ICC. The solid orange line is the primary comparator, the deployed cascade’s text-only stratum (ICC 0.7061), and the tinted band is its 95% CI (0.5419–0.8217); this comparator is modality-matched to the models, which were evaluated on text descriptions only. The dashed orange line is the same cascade pooled across all replicates (ICC 0.5905); it is lower because it includes the photographic pathway, so using it alone would overstate the models’ advantage, and both are therefore shown. Grey row bands and an asterisk mark the 5 configurations whose 95% CI overlaps that of the primary comparator ($ic_{lo} \leq 0.8217$ and $ic_{hi} \geq 0.5419$); for these the difference is not resolved at this sample size. Model identifiers are printed in full, ‘provider/model’, with no abbreviation. The two measurement regimes are not comparable, and the difference is not adjusted for: the 20 configurations were measured within a single

session against a pinned model version on text-only inputs, whereas the deployed cascade was measured over 179 days with unpinned versions and 48.3% photographic inputs. Higher ICC means more reproducible; reproducibility here is agreement among repeated responses to the same byte-identical input, not agreement with any nutritional reference standard, of which none was used. $n = 20$ configurations; the unit of replication is the input string. Five configurations are marked with an asterisk and shaded: their interval overlaps the primary comparator's. Interval overlap is a descriptive statement about precision, not a formal test of difference, and no hypothesis test between the cascade and any configuration is reported; the two were not measured under comparable conditions, as the footer records.

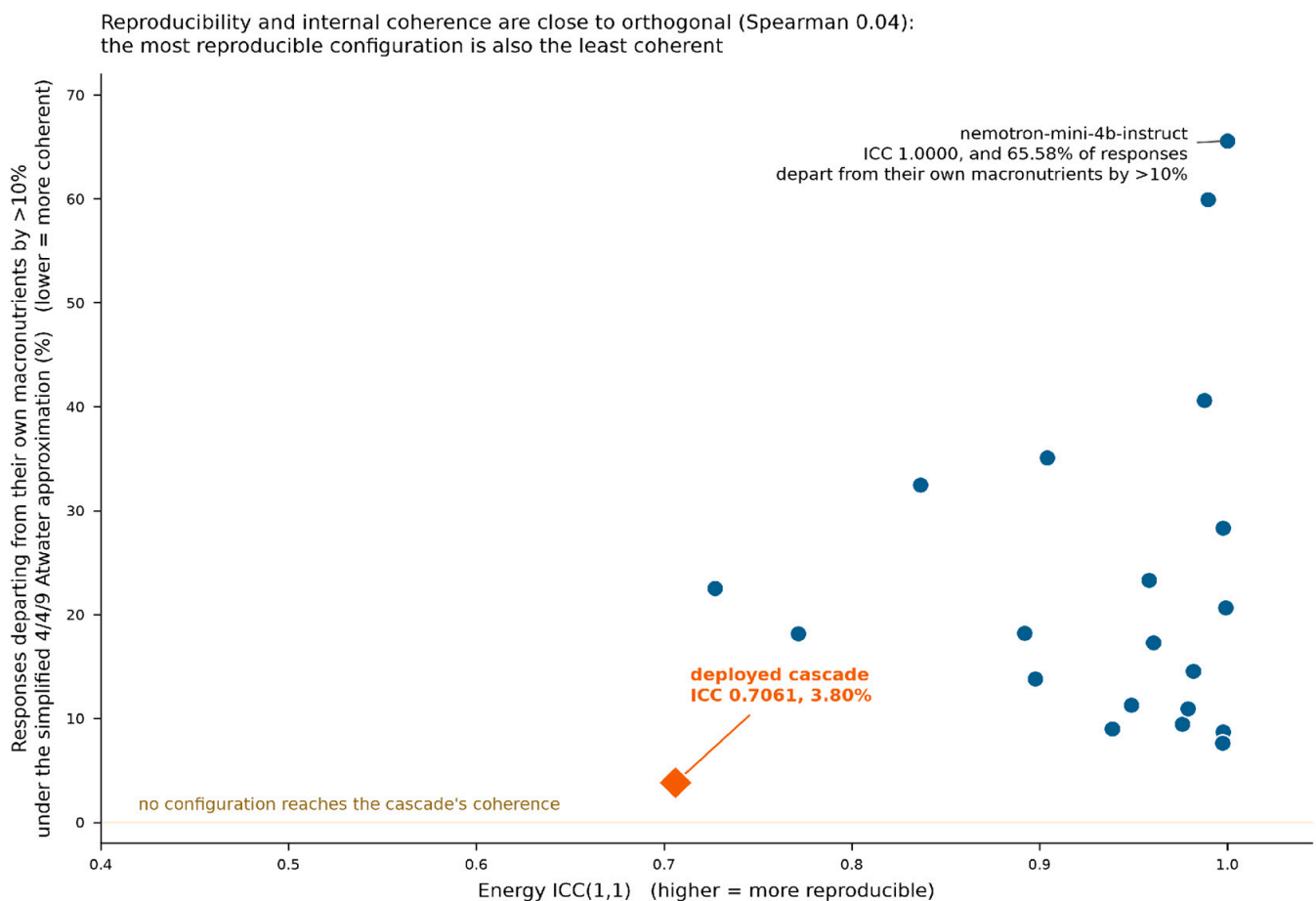


Figure 2. Reproducibility and internal coherence are close to orthogonal: reproducibility on the x -axis against departure from internal energy–macronutrient consistency on the y -axis, under the simplified 4/4/9 Atwater approximation. The most reproducible configuration is also the least coherent, and the Spearman correlation between the two axes is 0.04. The deployed cascade is the opposite case: the most coherent of the 21 systems compared, and the least reproducible under a design that was unfavourable to it. The 10% bound shown here is one point on the sensitivity curve of Figure 3, not a validated criterion. Source data: artl_tab_icc_por_modelo.csv and artl_tab_atwater_sensibilidad.csv.

Reliability alone also understates within-input divergence. ICC is computed against between-meal variance, which is large in this corpus, so a configuration can clear the reliability threshold while its repeated answers to the same string differ by clinically material amounts. Across the 20 configurations, the fraction of strings whose three replicates are exactly identical ranges from 0.1469 to 1.0000 and tracks the ICC ordering (Spearman 0.871), but the agreement breaks down in the middle of the range: nemotron-3-nano-omni-30b-a3b-reasoning passes the reliability threshold with an ICC of 0.7716 while 80.22% of its strings return more than one value and 37.91% differ by at least 100 kcal, and nemotron-3-

nano-30b-a3b passes with 0.8920 while 78.45% differ at all and 27.07% by at least 100 kcal. We therefore report the identical-response fraction and the within-string range alongside every ICC. (Table 9—artI_tab_repetibilidad_absoluta.csv).

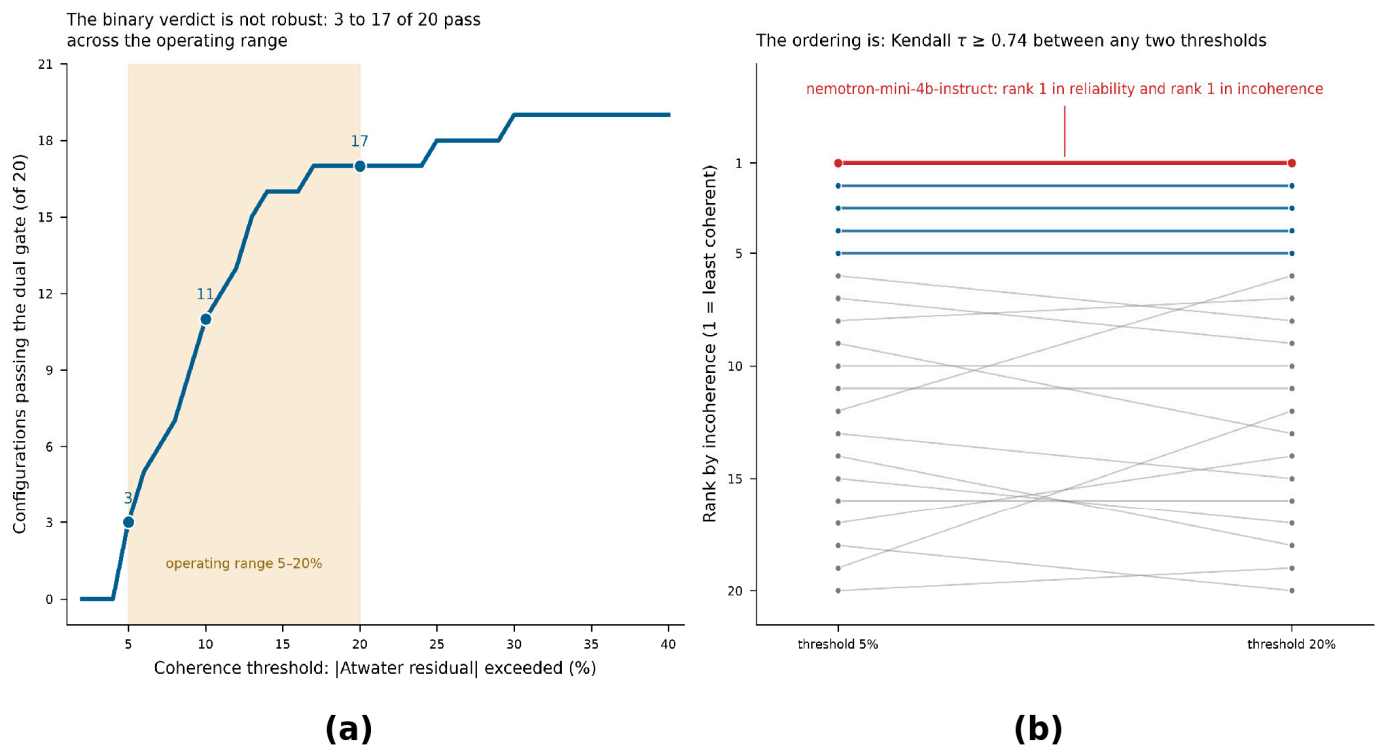


Figure 3. The acceptance verdict is not robust to the coherence bound, but the ordering is. (a) Configurations passing the dual gate as the bound on the Atwater departure is swept from 2% to 40%; the shaded band is the 5–20% operating range, over which the count moves from 3 to 17 of 20. (b) Rank by incoherence at the two ends of that range; Kendall tau is at least 0.74 between any two bounds in the range, the median configuration shifts 3 rank positions of 20, and the five least coherent configurations are the same five at both ends. The most reproducible configuration (nemotron-mini-4b-instruct) ranks first of 20 on both axes at every bound, which is the dissociation stated without reference to any threshold. Source data: artI_tab_atwater_sensibilidad.csv and artI_tab_puerta_sensibilidad.csv.

Table 6. Internal energy–macronutrient consistency by model under the simplified 4/4/9 Atwater approximation: median departure, p95, fraction above the 0.10 bound, fraction undefined. The bound is not a validated criterion; Table 8 and Figure 3 give the sensitivity across bounds. Source data: artI_tab_atwater_sensibilidad.csv.

Model	Median Discrepancy	p95	Fraction > 0.10	Fraction Undefined	Estimates
claude-sonnet-4-5-20250929	0.0160	0.1192	7.61%	1.55%	381
claude-opus-4-5-20251101	0.0200	0.1378	9.44%	1.64%	540
nvidia/nemotron-3-ultra-550b-a55b	0.0222	0.1346	10.94%	3.03%	384
claude-sonnet-4-6	0.0231	0.1620	11.27%	3.55%	408
deepseek-ai/deepseek-v4-flash-0731	0.0270	0.1307	8.97%	4.09%	446
nvidia/nemotron-3-super-120b-a12b	0.0316	0.1532	13.79%	3.37%	544
nvidia/ising-calibration-1.5-31b	0.0333	0.1300	8.69%	1.57%	564
claude-haiku-4-5-20251001	0.0338	0.1667	14.55%	0.91%	543
nvidia/nemotron-3-nano-omni-30b-a3b-reasoning	0.0374	0.2771	18.15%	1.96%	551
nvidia/nemotron-3-nano-30b-a3b	0.0379	0.2000	18.21%	1.61%	549
poolside/laguna-xs-2.1	0.0421	0.1977	23.30%	2.28%	515

Table 6. Cont.

Model	Median Discrepancy	p95	Fraction > 0.10	Fraction Undefined	Estimates
google/diffusionngemma-26b-a4b-it	0.0457	0.1726	17.28%	1.02%	486
nvidia/nemotron-3.5-lightning-30b-a3b	0.0462	0.2598	22.51%	3.04%	542
mistralai/mistral-nemotron	0.0500	0.1963	20.65%	3.66%	552
meta/llama-3.2-3b-instruct	0.0556	0.2250	28.33%	3.79%	533
nvidia/llama-3.3-nemotron-super-49b-v1	0.0688	0.2469	32.45%	3.62%	453
meta/llama-3.1-8b-instruct	0.0703	0.2389	35.09%	2.66%	513
nvidia/nemotron-nano-12b-v2-v1	0.0817	0.3000	40.60%	1.57%	564
nvidia/llama-3.1-nemotron-nano-v1-8b-v1	0.1235	0.4500	59.93%	4.19%	549
nvidia/nemotron-mini-4b-instruct	0.1667	0.5487	65.58%	3.16%	552

Table 7. Acceptance verdict under the reliability gate alone and under the dual gate. Source data: artl_tab_veredicto_aceptacion.csv.

Model	Reliability Gate	Weakest Field	Weakest ICC	Atwater Median	Atwater > 0.10	Coherence Gate	Dual Gate
nvidia/nemotron-mini-4b-instruct	Pass	Fat (g)	0.9986	0.1667	65.58%	Fail	Fail (coherence)
mistralai/mistral-nemotron	Pass	Fat (g)	0.9980	0.0500	20.65%	Fail	Fail (coherence)
meta/llama-3.2-3b-instruct	Pass	Fat (g)	0.9928	0.0556	28.33%	Fail	Fail (coherence)
nvidia/llama-3.1-nemotron-nano-v1-8b-v1	Pass	Fat (g)	0.9773	0.1235	59.93%	Fail	Fail (coherence)
nvidia/nemotron-nano-12b-v2-v1	Pass	Fat (g)	0.9653	0.0817	40.60%	Fail	Fail (coherence)
poolside/laguna-xs-2.1	Pass	Fat (g)	0.9358	0.0421	23.30%	Fail	Fail (coherence)
nvidia/llama-3.3-nemotron-super-49b-v1	Pass	Energy (kcal)	0.8367	0.0688	32.45%	Fail	Fail (coherence)
meta/llama-3.1-8b-instruct	Pass	Fat (g)	0.8305	0.0703	35.09%	Fail	Fail (coherence)
nvidia/nemotron-3.5-lightning-30b-a3b	Fail	Energy (kcal)	0.7271	0.0462	22.51%	Fail	Fail (reliability)
claude-sonnet-4-5-20250929	Pass	Carbohydrate (g)	0.9958	0.0160	7.61%	Pass	Pass
nvidia/ising-calibration-1.5-31b	Pass	Protein (g)	0.9949	0.0333	8.69%	Pass	Pass
claude-haiku-4-5-20251001	Pass	Protein (g)	0.9783	0.0338	14.55%	Pass	Pass
claude-opus-4-5-20251101	Pass	Carbohydrate (g)	0.9659	0.0200	9.44%	Pass	Pass
nvidia/nemotron-3-ultra-550b-a55b	Pass	Carbohydrate (g)	0.9637	0.0222	10.94%	Pass	Pass
google/diffusionngemma-26b-a4b-it	Pass	Carbohydrate (g)	0.9587	0.0457	17.28%	Pass	Pass
deepseek-ai/deepseek-v4-flash-0731	Pass	Energy (kcal)	0.9387	0.0270	8.97%	Pass	Pass
claude-sonnet-4-6	Pass	Fat (g)	0.9382	0.0231	11.27%	Pass	Pass
nvidia/nemotron-3-super-120b-a12b	Pass	Fat (g)	0.8460	0.0316	13.79%	Pass	Pass
nvidia/nemotron-3-nano-30b-a3b	Pass	Protein (g)	0.8342	0.0379	18.21%	Pass	Pass
nvidia/nemotron-3-nano-omni-30b-a3b-reasoning	Pass	Energy (kcal)	0.7716	0.0374	18.15%	Pass	Pass

Table 8. Acceptance verdict as a function of the coherence bound, with both reliability gate variants. Source data: artl_tab_puerta_sensibilidad.csv.

Configuration	Incoherent > 5%	Incoherent > 10%	Incoherent > 15%	Incoherent > 20%	Dual Gate, Bound 5%	Dual Gate, Bound 10%	Dual Gate, Bound 15%	Dual Gate, Bound 20%	Energy ICC	ICC Lower Bound	Reliability Gate, Point Estimate	Reliability Gate, Lower Bound	Verdict Changes vs. 10%
claude-sonnet-4-5-20250929	0.1654	0.0761	0.0420	0.0262	1	1	1	1	0.9974	0.9928	1	1	0
nvidia/ising-calibration-1.5-31b	0.2713	0.0869	0.0372	0.0284	0	1	1	1	0.9978	0.9944	1	1	1
deepseek-ai/deepseek-v4-flash-0731	0.2354	0.0897	0.0336	0.0291	0	1	1	1	0.9387	0.8682	1	1	1
claude-opus-4-5-20251101	0.1944	0.0944	0.0370	0.0222	1	1	1	1	0.9760	0.9351	1	1	0
nvidia/nemotron-3-ultra-550b-a55b	0.1771	0.1094	0.0495	0.0417	1	1	1	1	0.9790	0.9554	1	1	0
claude-sonnet-4-6	0.2034	0.1127	0.0588	0.0343	0	1	1	1	0.9489	0.8191	1	1	1
nvidia/nemotron-3-super-120b-a12b	0.3180	0.1379	0.0551	0.0276	0	1	1	1	0.8977	0.8519	1	1	1
claude-haiku-4-5-20251001	0.3204	0.1455	0.0589	0.0313	0	1	1	1	0.9818	0.9493	1	1	1
google/diffusion-gemma-26b-a4b-it	0.4630	0.1728	0.0679	0.0391	0	1	1	1	0.9607	0.9333	1	1	1
nvidia/nemotron-3-nano-omni-30b-a3b-reasoning	0.3739	0.1815	0.1053	0.0744	0	1	1	1	0.7716	0.7148	1	0	1
nvidia/nemotron-3-nano-30b-a3b	0.4007	0.1821	0.1038	0.0437	0	1	1	1	0.8920	0.7532	1	1	1
mistralai/mistral-nemotron	0.4783	0.2065	0.0996	0.0453	0	0	1	1	0.9992	0.9977	1	1	2
nvidia/nemotron-3.5-lightning-30b-a3b	0.4705	0.2251	0.1015	0.0738	0	0	0	0	0.7271	0.5812	0	0	0
poolside/laguna-xs-2.1	0.4117	0.2330	0.1146	0.0447	0	0	1	1	0.9582	0.9163	1	1	2
meta/llama-3.2-3b-instruct	0.5253	0.2833	0.1445	0.0657	0	0	1	1	0.9978	0.9915	1	1	2
nvidia/llama-3.3-nemotron-super-49b-v1	0.5938	0.3245	0.1391	0.0817	0	0	1	1	0.8367	0.6430	1	0	2
meta/llama-3.1-8b-instruct	0.6179	0.3509	0.1442	0.0994	0	0	1	1	0.9041	0.8316	1	1	2
nvidia/nemotron-nano-12b-v2-v1	0.6365	0.4060	0.2305	0.1348	0	0	0	1	0.9879	0.9778	1	1	1
nvidia/llama-3.1-nemotron-nano-v1-8b-v1	0.8160	0.5993	0.4007	0.2842	0	0	0	0	0.9896	0.9671	1	1	0
nvidia/nemotron-mini-4b-instruct	0.8370	0.6558	0.5344	0.3786	0	0	0	0	1	1	1	1	0

Coverage differs across configurations and is itself a finding, not only an exclusion rule. Parse success ranges from 0.6754 to 1.0000 of logged calls and usable corpus coverage from 53.9% to 100% of the 191 strings, and coverage correlates with reliability (Spearman 0.429), so configurations evaluated on fewer strings are also the less reproducible ones. To bound the resulting selection bias, we recomputed every ICC on progressively more common input sets. Only nine strings carry three replicates in all 20 configurations, too few to estimate on, and restricting to them raises every ICC, because the strings that

every configuration resolves are the tractable ones. The direction of that effect is worth stating carefully, because it is the opposite of the one a selection bias objection would predict. Equalising inputs would flatter the low-coverage configurations, not penalise them; since we do not equalise, the estimates we publish are the lower of the two for exactly those configurations. Unequal coverage therefore cannot be inflating the weak configurations in our results—if anything it holds them down. On the strings present in at least 16 configurations, which retains 174 of 191, no ICC moves by more than 0.043, and the ordering is preserved at Spearman 0.984. That analysis does not evaluate all 20 configurations on the same 174 strings: the set is defined by coverage in at least 16 configurations, and each configuration is then estimated on the subset of those strings it actually covers with three replicates, which ranges from 100 to 174 strings (median 167). The strictly balanced set—the nine strings covered by all 20—is too small to estimate on, which is why the graded bound is reported instead. (Table 10—artI_tab_cobertura_y_entradas_comunes.csv), (Table 11—artI_tab_entradas_comunes_graduado.csv).

Table 9. Absolute within-input repeatability by configuration: identical response fraction, within-string range and coefficient of variation, alongside the ICC. Source data: artI_tab_repetibilidad_absoluta.csv.

Configuration	Strings	Observations	Identical Replicates (Fraction)	Within-String Range, Median (kcal)	Within-String Range, p95 (kcal)	Within-String Range, Max (kcal)	Relative Range, Median	Within-String CV, Median	Within-String CV, p95	Energy ICC
nvidia/nemotron-mini-4b-instruct	190	570	1	0	0	0	0	0	0	1
mistralai/mistral-nemotron	191	573	0.9529	0	0	100	0	0	0	0.9992
nvidia/ising-calibration-1.5-31b	191	573	0.9319	0	10.5000	100	0	0	0.1060	0.9978
meta/llama-3.2-3b-instruct	184	552	0.9837	0	0	200	0	0	0	0.9978
claude-sonnet-4-5-20250929	111	333	0.9550	0	0	130	0	0	0	0.9974
nvidia/llama-3.1-nemotron-nano-v1-8b-v1	191	573	0.9005	0	55	450	0	0	0.1068	0.9896
nvidia/nemotron-nano-12b-v2-v1	191	573	0.8848	0	85	167	0	0	0.1496	0.9879
claude-haiku-4-5-20251001	182	546	0.9396	0	5	491	0	0	0.0117	0.9818
nvidia/nemotron-3-ultra-550b-a55b	103	309	0.6311	0	114.5000	235	0	0	0.3270	0.9790
claude-opus-4-5-20251101	182	546	0.9011	0	30	482	0	0	0.0663	0.9760
google/diffusionge-mma-26b-a4b-it	128	384	0.5625	0	100	302	0	0	0.2072	0.9607
poolside/laguna-xs-2.1	161	483	0.7578	0	100	321	0	0	0.3514	0.9582
claude-sonnet-4-6	115	345	0.6000	0	249.5000	320	0	0	0.7236	0.9489
deepseek-ai/deepseek-v4-flash-0731	151	453	0.7219	0	100	500	0	0	0.1823	0.9387
meta/llama-3.1-8b-instruct	167	501	0.6946	0	167.3000	1000	0	0	0.4001	0.9041
nvidia/nemotron-3-super-120b-a12b	186	558	0.3011	17.5000	207.2500	421	0.0887	0.0495	0.4288	0.8977
nvidia/nemotron-3-nano-30b-a3b	181	543	0.2155	38	290	820	0.1875	0.1039	0.5407	0.8920
nvidia/llama-3.3-nemotron-super-49b-v1	136	408	0.7941	0	138.7500	1450	0	0	0.3046	0.8367
nvidia/nemotron-3-nano-omni-30b-a3b-reasoning	182	546	0.1978	63	380	1200	0.2863	0.1469	0.7059	0.7716
nvidia/nemotron-3.5-lightning-30b-a3b	177	531	0.1469	50	254.6000	1870	0.1960	0.1038	0.7646	0.7271

Table 10. Coverage, parse success and reliability on full coverage against common inputs, by configuration. Source data: artl_tab_cobertura_y_entradas_comunes.csv.

Configuration	Calls Logged	Responses Parsed	Usable Rows	Strings with $k \geq 3$	Parse Success	Corpus Coverage	ICC, Full Coverage	ICC, Common Inputs	Difference
nvidia/nemotron-3-ultra-550b-a55b	553	396	309	103	0.7161	0.5393	0.9790	0.9947	0.0157
claude-sonnet-4-5-20250929	573	387	333	111	0.6754	0.5812	0.9974	1	0.0026
claude-sonnet-4-6	573	423	345	115	0.7382	0.6021	0.9489	0.9648	0.0159
google/diffusiongemma-26b-a4b-it	497	491	384	128	0.9879	0.6702	0.9607	0.9994	0.0387
nvidia/llama-3.3-nemotron-super-49b-v1	573	470	408	136	0.8202	0.7120	0.8367	0.9822	0.1455
deepseek-ai/deepseek-v4-flash-0731	465	465	453	151	1	0.7906	0.9387	0.9991	0.0604
poolside/laguna-xs-2.1	530	527	483	161	0.9943	0.8429	0.9582	1	0.0418
meta/llama-3.1-8b-instruct	573	527	501	167	0.9197	0.8743	0.9041	0.9996	0.0955
nvidia/nemotron-3.5-lightning-30b-a3b	573	559	531	177	0.9756	0.9267	0.7271	0.9812	0.2541
nvidia/nemotron-3-nano-30b-a3b	573	558	543	181	0.9738	0.9476	0.8920	0.9827	0.0907
nvidia/nemotron-3-nano-omni-30b-a3b-reasoning	573	562	546	182	0.9808	0.9529	0.7716	0.9521	0.1805
claude-haiku-4-5-20251001	573	548	546	182	0.9564	0.9529	0.9818	1	0.0182
claude-opus-4-5-20251101	573	549	546	182	0.9581	0.9529	0.9760	1	0.0240
meta/llama-3.2-3b-instruct	573	554	552	184	0.9668	0.9634	0.9978	1	0.0022
nvidia/nemotron-3-super-120b-a12b	573	563	558	186	0.9825	0.9738	0.8977	0.9761	0.0784
nvidia/nemotron-mini-4b-instruct	573	570	570	190	0.9948	0.9948	1	1	0
nvidia/nemotron-nano-12b-v2-v1	573	573	573	191	1	1	0.9879	0.9997	0.0118
mistralai/mistral-nemotron	573	573	573	191	1	1	0.9992	1	0.0008
nvidia/llama-3.1-nemotron-nano-v1-8b-v1	573	573	573	191	1	1	0.9896	1	0.0104
nvidia/ising-calibration-1.5-31b	573	573	573	191	1	1	0.9978	1	0.0022

Table 11. Sensitivity of reliability to input equalisation: ICC recomputed on strings present in at least N configurations. Source data: artl_tab_entradas_comunes_graduado.csv.

Present in at Least N Configurations	Strings Retained	Median ICC	Spearman vs. Full Coverage	Largest ICC Change
20	9			
18	96	0.9846	0.8469	0.1498
16	174	0.9721	0.9838	0.0431
14	185	0.9699	0.9861	0.0529
12	188	0.9681	0.9891	0.0525
10	190	0.9684	0.9996	0.0012
8	191	0.9684	0.9996	0

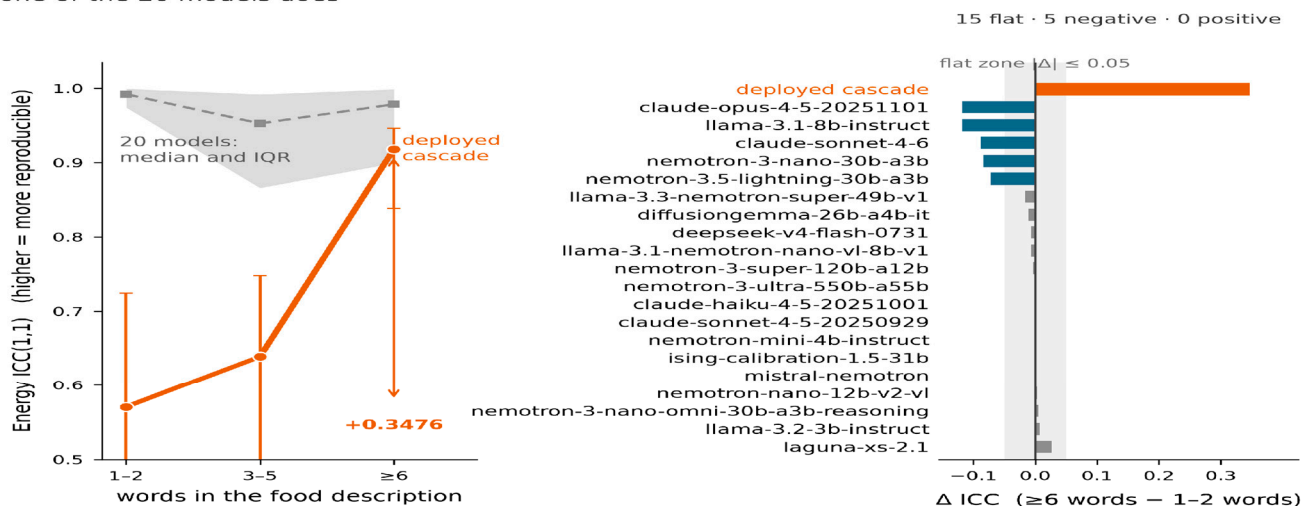
3.2. Cross-Model Agreement

Cross-model agreement is ICC 0.6865 (95% CI 0.5990 to 0.7563) on the 82 strings with complete coverage, against individual reproducibility of 0.7271 to 1.0000. Consistency is 0.7056, so the agreement-consistency gap is small: the disagreement is not a matter of additive offsets between models. Each configuration is internally stable and mutually incompatible, which places the disagreement in the model, not in the serving infrastructure.

3.3. The Cascade's Specificity Gradient Does Not Transfer

The cascade gains reliability as descriptions get more specific: energy ICC 0.5705 on 1–2 word inputs rising to 0.9181 on inputs of six words or more, a change of +0.3476. That gradient is the operational intuition the deployment was built on, and it does not survive (Figure 4). Of the 20 models, 0 show a comparable positive gradient, 15 are flat and five are negative. The finding holds under both the stratified specification and a continuous one regressing within-string dispersion on word count.

The deployed cascade gains reliability as descriptions get more specific; none of the 20 models does



Filler stratum excluded. Non-comparable conditions: models measured in one session, pinned version, text-only; the deployed cascade over 179 days, unpinned versions, 48.3% photographs. Blue = negative ($\Delta < -0.05$), grey = flat.

Figure 4. The specificity gradient is a property of the deployed cascade and is not reproduced by any of the 20 model configurations. Left: energy ICC(1,1) by description-length stratum (1–2, 3–5 and ≥ 6 words). The deployed cascade (orange, with bootstrap 95% CI) rises from 0.5705 in the 1–2 word stratum to 0.9181 in the ≥ 6 word stratum, a gain of +0.3476 spanned by the arrow. The 20 configurations are summarised by their median (dashed) and interquartile range (grey ribbon) rather than as individual traces, because at this width, twenty overlapping lines are not separable; they sit near the reproducibility ceiling in every stratum and do not rise. Right: the per-configuration change in reliability between the extreme strata, $\Delta = \text{ICC}(\geq 6 \text{ words}) - \text{ICC}(1-2 \text{ words})$, sorted, with the deployed cascade shown as the top bar on the same axis. The shaded band is the ± 0.05 zone within which a configuration is classified as flat; 15 configurations are flat, 5 are negative ($\Delta < -0.05$, blue) and 0 are positive. The largest Δ observed for any configuration is +0.0266, inside the flat zone. Abbreviation rule, right panel only: model names are shortened by removing the provider prefix that precedes the slash and nothing else, so ‘nvidia/nemotron-3-nano-omni-30b-a3b-reasoning’ is printed as ‘nemotron-3-nano-omni-30b-a3b-reasoning’; all 20 shortened names remain unique, and Figure 1 gives the unabbreviated identifiers. The low-information filler stratum is excluded from both panels. As in Figure 1, the two regimes are not comparable and no adjustment is made: the configurations were measured in a single session with a pinned version on text-only inputs, the deployed cascade over 179 days with unpinned versions and 48.3% photographic inputs. Per-model, per-stratum ICC values are given in Table 12. Twenty models. Right: change between extreme strata per model, against the cascade’s +0.3476.

The filler stratum separates the two regimes sharply. The cascade is essentially random on degenerate input (ICC 0.1028, interval spanning zero), whereas models split: several are perfectly reproducible on comida variada and others collapse. (Table 12—artI_tab_gradiente_especificidad.csv).

Table 12. Specificity gradient: ICC by stratum and model, including the filler stratum. Source data: artI_tab_gradiente_especificidad.csv.

Model	Stratum	ICC(1,1)	Strings	Observations
nvidia/llama-3.3-nemotron-super-49b-v1	1–2 words	0.9886	51	153
nvidia/llama-3.3-nemotron-super-49b-v1	3–5 words	0.9331	65	195
nvidia/llama-3.3-nemotron-super-49b-v1	≥6 words	0.9729	18	54
nvidia/llama-3.3-nemotron-super-49b-v1	Filler label	0.0560	2	6
meta/llama-3.1-8b-instruct	1–2 words	0.9916	47	141
meta/llama-3.1-8b-instruct	3–5 words	0.8787	97	291
meta/llama-3.1-8b-instruct	≥6 words	0.8745	18	54
meta/llama-3.1-8b-instruct	Filler label	0.7750	5	15
nvidia/nemotron-nano-12b-v2-v1	1–2 words	0.9950	52	156
nvidia/nemotron-nano-12b-v2-v1	3–5 words	0.9798	107	321
nvidia/nemotron-nano-12b-v2-v1	≥6 words	0.9982	24	72
nvidia/nemotron-nano-12b-v2-v1	Filler label	0.9724	8	24
meta/llama-3.2-3b-instruct	1–2 words	0.9928	52	156
meta/llama-3.2-3b-instruct	3–5 words	1.0000	107	321
meta/llama-3.2-3b-instruct	≥6 words	1.0000	24	72
nvidia/nemotron-mini-4b-instruct	1–2 words	1.0000	52	156
nvidia/nemotron-mini-4b-instruct	3–5 words	1.0000	107	321
nvidia/nemotron-mini-4b-instruct	≥6 words	1.0000	24	72
nvidia/nemotron-mini-4b-instruct	Filler label	1.0000	7	21
mistralai/mistral-nemotron	1–2 words	0.9978	52	156
mistralai/mistral-nemotron	3–5 words	0.9992	107	321
mistralai/mistral-nemotron	≥6 words	0.9991	24	72
mistralai/mistral-nemotron	Filler label	1.0000	8	24
nvidia/ising-calibration-1.5-31b	1–2 words	0.9998	52	156
nvidia/ising-calibration-1.5-31b	3–5 words	0.9933	107	321
nvidia/ising-calibration-1.5-31b	≥6 words	1.0000	24	72
nvidia/ising-calibration-1.5-31b	Filler label	1.0000	8	24
nvidia/llama-3.1-nemotron-nano-v1-8b-v1	1–2 words	1.0000	52	156
nvidia/llama-3.1-nemotron-nano-v1-8b-v1	3–5 words	0.9913	107	321
nvidia/llama-3.1-nemotron-nano-v1-8b-v1	≥6 words	0.9932	24	72
nvidia/llama-3.1-nemotron-nano-v1-8b-v1	Filler label	0.9799	8	24
nvidia/nemotron-3.5-lightning-30b-a3b	1–2 words	0.9155	51	153
nvidia/nemotron-3.5-lightning-30b-a3b	3–5 words	0.7607	97	291
nvidia/nemotron-3.5-lightning-30b-a3b	≥6 words	0.8440	22	66
nvidia/nemotron-3.5-lightning-30b-a3b	Filler label	0.4726	7	21
nvidia/nemotron-3-nano-30b-a3b	1–2 words	0.8749	52	156
nvidia/nemotron-3-nano-30b-a3b	3–5 words	0.7637	99	297
nvidia/nemotron-3-nano-30b-a3b	≥6 words	0.7905	22	66
nvidia/nemotron-3-nano-30b-a3b	Filler label	0.9870	8	24
nvidia/nemotron-3-nano-omni-30b-a3b-reasoning	1–2 words	0.8195	51	153
nvidia/nemotron-3-nano-omni-30b-a3b-reasoning	3–5 words	0.7163	103	309
nvidia/nemotron-3-nano-omni-30b-a3b-reasoning	≥6 words	0.8244	24	72
nvidia/nemotron-3-nano-omni-30b-a3b-reasoning	Filler label	0.7767	4	12
nvidia/nemotron-3-super-120b-a12b	1–2 words	0.9617	52	156
nvidia/nemotron-3-super-120b-a12b	3–5 words	0.8276	106	318

Table 12. Cont.

Model	Stratum	ICC(1,1)	Strings	Observations
nvidia/nemotron-3-super-120b-a12b	≥6 words	0.9577	24	72
nvidia/nemotron-3-super-120b-a12b	Filler label	0.8550	4	12
nvidia/nemotron-3-ultra-550b-a55b	1–2 words	0.9816	39	117
nvidia/nemotron-3-ultra-550b-a55b	3–5 words	0.9767	42	126
nvidia/nemotron-3-ultra-550b-a55b	≥6 words	0.9801	13	39
nvidia/nemotron-3-ultra-550b-a55b	Filler label		2	6
poolside/laguna-xs-2.1	1–2 words	0.9714	40	120
poolside/laguna-xs-2.1	3–5 words	0.9334	95	285
poolside/laguna-xs-2.1	≥6 words	0.9980	20	60
poolside/laguna-xs-2.1	Filler label	0.9478	6	18
deepseek-ai/deepseek-v4-flash-0731	1–2 words	0.9840	44	132
deepseek-ai/deepseek-v4-flash-0731	3–5 words	0.9403	80	240
deepseek-ai/deepseek-v4-flash-0731	≥6 words	0.9771	21	63
deepseek-ai/deepseek-v4-flash-0731	Filler label	0.4207	6	18
google/diffusiongemma-26b-a4b-it	1–2 words	0.9753	38	114
google/diffusiongemma-26b-a4b-it	3–5 words	0.9457	73	219
google/diffusiongemma-26b-a4b-it	≥6 words	0.9645	13	39
google/diffusiongemma-26b-a4b-it	Filler label	0.7443	4	12
claude-haiku-4-5-20251001	1–2 words	1.0000	51	153
claude-haiku-4-5-20251001	3–5 words	0.9599	107	321
claude-haiku-4-5-20251001	≥6 words	0.9986	24	72
claude-opus-4-5-20251101	1–2 words	0.9998	51	153
claude-opus-4-5-20251101	3–5 words	0.9717	105	315
claude-opus-4-5-20251101	≥6 words	0.8820	22	66
claude-sonnet-4-5-20250929	1–2 words	0.9993	21	63
claude-sonnet-4-5-20250929	3–5 words	0.9958	70	210
claude-sonnet-4-5-20250929	≥6 words	0.9990	19	57
claude-sonnet-4-6	1–2 words	0.9930	47	141
claude-sonnet-4-6	3–5 words	0.7579	56	168
claude-sonnet-4-6	≥6 words	0.9053	6	18
claude-sonnet-4-6	Filler label	0.9985	6	18

3.4. Irreproducibility at Temperature 0 Is Not a Sampling Artefact

In a deep probe of six strings across six models (36 model-by-string cells) with 30 replicates each, 17 of the 36 cells (0.4722) returned more than one distinct energy value at temperature 0, up to five distinct values in 30 calls; separately, 23 of the 36 cells returned the same number in more than one serialisation format. Raising temperature to 0.7 amplifies dispersion (mean distinct values 2.0 to 5.0) but does not explain it: one model returns three distinct values at temperature 0 and two at 0.7. A temperature probe is consistent with the absence of response caching—byte-identical repetition at temperature 0 gives way to variation at temperature 1 for the same nominal cache key—but it does not rule caching out: provider cache policies may depend on decoding parameters, server load, request timestamps or endpoint, none of which we can observe from the client side. We therefore treat the temperature 0 identity as consistent with greedy determinism rather than as established proof of it.

4. Discussion

4.1. What the Dissociation Means for Acceptance Protocols

The result that should change practice is not the ranking. It is that eight of 20 models pass a reliability threshold fixed in advance and fail a coherence check applied afterwards, and that the two most reproducible configurations in the study are among them. Reliability and internal coherence are not two measurements of the same underlying quality; they are orthogonal, and a protocol that measures only the first will authorise a pipeline whose macronutrient arithmetic is impossible in a large fraction of its outputs.

This matters precisely because reliability is the quantity that is estimable without a reference standard. The literature evaluating language models on clinical questions has converged on repetition as the diagnostic of choice [14–16], and prompting strategies that sample repeatedly and aggregate are now standard for accuracy [22–24]—both implicitly concede that a single call is not a measurement. What none of that work supplies is a check that the aggregated answer is possible. The temptation is to treat it as a proxy for quality because it is the available measurement.

The Atwater check shows the temptation is avoidable: energy and macronutrients are linked by a stoichiometric relation that the conventional Atwater factors approximate, so their mutual consistency is checkable with no external truth. Any deployment reporting both can be audited on this axis at no additional data cost. We would expect analogous internal constraints to exist in other estimation domains, and the general lesson is that reference standard-free auditing should search for them rather than settle for reliability alone.

4.2. What This Study Does Not Establish About the Deployed Cascade

Three inferences a reader might draw from the ranking are not supported by this design, and we state them as prohibitions rather than caveats. A paired comparison between the configurations and the cascade on the same meal descriptions is not possible from these data, and we state the reason rather than leaving it implicit: the corpus is drawn from the cascade's production logs, but the cascade's own per-description nutrient outputs for those same records were not retained in a form we can align to the released strings, so there is no per-item cascade value to pair against. The cascade figures we report are therefore group-level comparators computed on its own records, not the other half of a matched pair, and no paired or within-item test is reported anywhere in this article.

It is not an assessment of the integrating platform's engineering. The audited object is a two-stage cascade of commercial models from two providers, reached through APIs whose prompts, temperature, checkpoints and composition anchoring are provider-controlled and revised without notice, so no version of the deployed cascade can be pinned retrospectively [30]. The reliability we compare against is a property of third-party model behaviour observed through an integration. The benchmarked models are the same class of component, so the finding concerns that component class, not the software that calls it. The central dissociation in fact cuts the other way for anyone building such a system: selecting the most reproducible available model would have chosen the configuration whose outputs are internally inconsistent in 65.58% of responses.

It is not evidence that switching models would improve the deployment. Every model here was measured text-only, in one session, against a pinned identifier. The cascade's figure comes from naturally occurring repetitions across 179 days, spanning unannounced version changes, with 48.3% of entries routed additionally through a vision stage. Placing any of these models into that setting would expose it to the same drift and the same second stage, both excluded here by design. The benchmark bounds the difference; it does not predict the substitution.

It is not a coherence failure of the deployed cascade—the opposite. On the axis that needs neither a reference standard nor version pinning, the cascade is the best of the 21 systems compared: 3.80% of its records depart from their own macronutrients by more than 10% under the simplified Atwater approximation, against a model median of 18.18% (4.8x) and a worst case of 65.58% (17.3x). Zero of 20 models match it. A reader who treats reliability as a summary verdict inverts the result: the deployed system is the most physically coherent estimator in the comparison, and the one whose reliability was measured under the least favourable design.

4.3. Models Agree with Themselves but Not with Each Other

Cross-model agreement (ICC 0.6865) sits below every individual model's own reproducibility, and the small agreement consistency gap (0.6865 against 0.7056, a gap of 0.0191) indicates that systematic additive offsets between models account for little of the disagreement: removing them would barely improve concordance. The disagreement therefore lies in the estimates themselves rather than in a constant shift per model. Comparative evaluations of several models on the same clinical items report the same divergence on identical inputs [44], and the expert-consensus reporting standard for studies of these systems now requires that the prompting and sampling conditions under which outputs were produced be documented explicitly [45]. For an exposure variable, the implication is procedural: the choice of configuration is itself a measurement decision, and switching model mid-cohort changes the instrument.

4.4. Why the Cascade's Specificity Gradient Did Not Transfer

The audited cascade becomes more reliable as descriptions become more specific (+0.3476 across the extreme strata), which is intuitive: a longer description constrains the lookup. That gradient did not appear in a single one of the 20 configurations—15 flat, five negative. The mechanism is plausibly that a table-anchored estimator's ambiguity falls with input specificity whereas a model's sampling variability does not: a longer description gives a language model more, not fewer, degrees of freedom in how it decomposes the meal. We observed that decomposition directly in the raw responses, where several models switched between per-portion and per-100 g bases across byte-identical replicates without declaring it.

The operational consequence is that a deployment's reliability profile cannot be inherited from a predecessor system, even one solving the same task on the same inputs. It has to be measured per configuration—and, because unreliable exposure attenuates associations in proportion to its reliability [2–4], the measurement has to precede any epidemiological use of the resulting variable.

4.5. Irreproducibility at Temperature 0

Greedy decoding is often assumed deterministic. In 47.22% of the probed cells, it was not, and raising the temperature amplified but did not explain the dispersion. Reviews of language model limitations in medicine flag output instability as an obstacle to deployment [12,13,20,21], though they generally attribute it to sampling rather than to the serving stack. Two hypotheses are consistent with this pattern, neither of which this study tests: non-associative floating point reduction under varying batch composition, and mixture-of-experts routing that depends on co-batched requests—both properties of the serving infrastructure rather than of the weights. For an audit, this distinction is consequential: the estimand is the deployed configuration, and the same weights behind a different endpoint are a different measurement instrument.

4.6. Limitations

Coverage. Four probed configurations did not reach the inclusion rule of 100 strings with three usable replicates and are excluded with their counts recorded: glm-5.2 (3 strings); llama-3.1-70b-instruct (0 strings); llama-3.2-11b-vision-instruct (0 strings); llama-3.3-70b-instruct (0 strings). Three of these have the full corpus at two replicates but none at three, so their exclusion reflects an interrupted acquisition and not a property of the models. The account-level rate limit on the inference platform, not study design, determined where acquisition stopped; this is an operational constraint on the benchmark's breadth, and we report it rather than presenting the achieved set as the intended one.

The baseline comparison is asymmetric by construction, and the asymmetry favours the models. This is the most important limitation on any ranking read out of this study. The models were measured in a single session, against a pinned model identifier, text-only, at temperature 0, with one fixed prompt.

The deployed cascade's figure comes from naturally occurring repetitions of identical strings separated in time across 179 days, spanning provider-side version changes that were neither announced nor controllable, with 48.3% of entries additionally traversing a vision stage, and over an interval in which provider-side configuration could change without notice [30].

The benchmark therefore estimates short-run sampling reproducibility, whereas the cascade figure absorbs version drift and a second processing stage as well. Every reliability gap reported here is an upper bound on model attributable difference, not an estimate of it, and five of 20 model intervals overlap the modality-matched cascade interval. The design that would make the comparison symmetric—one session, one pinned version, text-only input for both arms—is specified in the companion manuscript [30] (§4.4) and is inexpensive: resubmit the same inputs to the deployed cascade within one session, against pinned checkpoints, with the intermediate text frozen. Until that is run, no statement of the form “model X is more reliable than the deployment” is established for any individual model.

Single prompt, single temperature. The estimand is one deployed configuration per model. Prompt sensitivity is not measured here and could plausibly exceed the between-model differences we report, since prompt formulation alone shifts performance substantially on comparable tasks [36–38].

Parsing as an estimand-shaping decision. A substantial share of responses returned nutrients decomposed per food item rather than as a meal total. Discarding them is not neutral: they concentrate in the descriptive strata, so a plain-JSON specification silently conditions reliability on the simplest inputs and inflates it. We aggregated them and report the plain specification as a robustness analysis; the aggregated estimates are lower, which is the direction that matters.

Coherence threshold. The 20% coherence gate is a decision rule we propose, not one established in the literature. The dissociation itself does not depend on the cut point; the pass count does.

No external validity for accuracy. Nothing here licenses a claim that any model estimates nutrients correctly. Establishing that would require weighed food reference data, and the instruments that come closest—image- and app-based assessment validated against food diaries or doubly labelled water [6–8,46,47]—measure agreement with an imperfect comparator rather than truth. A model could be perfectly reproducible, perfectly Atwater-coherent, and systematically wrong by a factor of two. That question needs weighed-food reference data and is outside this design.

4.7. Relation to the Sibling Articles

This is the third article of the BIOMARCADORES-111598 programme. The first [30] audited the deployed cascade and supplies the baseline used throughout; the second, a manuscript in preparation by the same authors, derived the operating characteristics of the acceptance protocol and supplies the imbalance correction and the sample-size guidance. The three share a corpus lineage and are explicitly cross-declared. The tool used here is released as repra-bench, which reproduces every figure of the companion audit [30] from source as a regression test.

5. Conclusions

Test–retest reliability is measurable without a reference standard and is necessary for an LLM-assisted nutritional pipeline, but it is not sufficient. The most reproducible configuration we measured returns energy that contradicts its own macronutrients, under the conventional Atwater factors, in two thirds of its responses, so an acceptance protocol built on reliability alone would authorise it. Acceptance should require, at minimum, one internal coherence check alongside the reliability interval, and both should be reported per deployed configuration rather than inherited from a predecessor system.

The comparison should not be read as a ranking of deployments. The systems benchmarked here are the component class that a production cascade consumes, measured under conditions the cascade did not have: one session, a pinned model identifier, text-only input. On the one axis that requires neither version pinning nor a reference standard—internal energy–macronutrient consistency—the deployed cascade outperformed all 20 models, and none matched it. The actionable conclusion for teams integrating third-party nutrition models is therefore not “switch models” but “measure both axes on your own configuration, because reliability alone would have selected the least coherent option available.”

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/make8100331/s1>, The replication package described in the Data Availability Statement can be downloaded at the journal website as a single archive (artI-repra-bench-v1.zip). No table or figure is supplied as a separate supplementary item: all twelve tables and all four figures are typeset in the manuscript itself, and the archive holds the data and code behind them. It contains the benchmark corpus, the 12,119 raw model calls with their unparsed responses, the published tables and figures, and an offline script that recomputes all 63 published figures and exits non-zero if any one of them changes.

Author Contributions: Conceptualization, J.F.L.T.; methodology, J.F.L.T. and E.R.; software, J.F.L.T.; validation, J.F.L.T., J.G.C. and E.R.; formal analysis, J.F.L.T.; investigation, J.F.L.T. and J.G.C.; resources, J.G.C. and E.R.; data curation, J.F.L.T.; writing—original draft preparation, J.F.L.T.; writing—review and editing, J.F.L.T., J.G.C. and E.R.; visualization, J.F.L.T.; supervision, E.R.; project administration, J.G.C.; funding acquisition, E.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by MinCiencias (Ministerio de Ciencia, Tecnología e Innovación, Colombia), contract 1100 of 2024 (call 949 of 2024), BIOMARCADORES-111598 research programme. The APC was funded by the same programme.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by the Research Ethics Committee (Comité de Ética en Investigación) of the Universidad Distrital Francisco José de Caldas, Bogotá, Colombia. Approval was granted at programme level on 17 June 2024 for programme BIOMARCADORES-111598, under Resolution No. 8430 of 4 October 1993 of the Colombian Ministry of Health, the national regulation that sets the scientific, technical and administrative standards for health research. The present work is a secondary analysis of de-identified operational records generated by that programme, with no intervention, no

new data collection and no participant contacted for research purposes; it falls within the scope of the programme-level approval, and no separate protocol was required. Two points that the approval does not by itself settle are stated here explicitly. The first is the scope of the approval. The committee's approval covers the secondary analysis of the programme's operational records; it does not contain an explicit authorisation to publish the corpus, and we do not claim that it does. The decision to release the 191 meal description strings was taken by the authors, on the grounds set out next: the released file contains no personal data, so it falls outside the scope of the consent and data protection provisions that govern the underlying records. Readers who consider that release to require a separate institutional authorisation should treat the corpus as the authors' decision rather than the committee's, and we would withdraw it to restricted access on request. The second is how that claim about personal data was established. De-identification was performed on the free-text strings before release and was verified by reading, not only by rule: the strings were normalised and screened with a pattern-based detector for person names, numeric identifiers, dates and contact details, and then every one of the 191 strings was read individually. The detector raised two candidates, both of which were false positives on inspection—a commercial beverage brand name that the detector took for a surname, and one description flagged only by its length (71 characters) with nothing identifying in it. Both were retained. The released corpus carries no participant identifier of any kind and consists only of meal descriptions and their stratum labels, so there is no linkage key by which a record could be traced back to a person.

Informed Consent Statement: Informed consent was obtained from all participants at enrolment in the parent programme, as required by Article 6, literal g of Resolution No. 8430 of 1993 of the Colombian Ministry of Health, under which the programme's ethics approval was granted. No additional consent was sought for the present analysis, which uses only operational records already generated by the programme's routine functioning. The corpus released with this article contains no participant identifier of any kind—not a name, not a record number and not a pseudonym—only the normalised free-text meal descriptions and their stratum labels.

Data Availability Statement: All data and code needed to reproduce every number in this article are provided as a single replication package, supplied to the journal as Supplementary Material (artl-repra-bench-v1.zip, 2.6 MB, 62 files, CC BY 4.0 for data and MIT for code). The package contains: the 191 de-identified benchmark input strings with their replicate counts and strata (data/corpus/); the 12,119 model calls as returned, one row per call and including the unparsed response text, so that a reader can apply a different parsing rule and observe its effect (data/raw/); the consolidated analysis file (data/processed/); the twelve tables reported in this article and the artefacts behind each published figure (tables/); the four figures at 400 dpi (figures/); and the estimator package together with the consolidation script and an offline verifier (code/). Running "python code/verify_figures.py tables" reproduces all 63 published figures from the artefact and column named for each in tables/artl_verificacion_cifras.csv and returns a non-zero exit code if any single value disagrees; the run requires no network access and no credentials. Every file is listed with its SHA-256 in MANIFEST-SHA256.txt. The provider-side prompts, checkpoints and serving configuration of the evaluated commercial models are not ours to release and can be revised by their providers without notice, so re-collection reproduces the procedure rather than the exact responses; docs/REPRODUCE.md states the procedure and the two collection parameters that most affect the result.

Acknowledgments: During the preparation of this manuscript, the authors used three general-purpose generative AI assistants—ChatGPT (OpenAI), Claude (Anthropic), and Gemini (Google)—in the versions publicly available to them during the first half of 2026. The authors did not record specific model build identifiers at the time of use; therefore, they indicate the period of usage rather than specifying a version they cannot verify. Their use was limited to two support tasks. First, the drafting and debugging of analytical code: every snippet generated in this manner was read, tested, and verified by the authors against the frozen data extract prior to use. Second, the linguistic refinement of texts written by the authors: every edited sentence was reviewed and subsequently either accepted or rewritten by the authors, who assume responsibility for the content. No AI system was used to conceive the study; select or specify analyses; calculate or generate any reported values;

create figures or tables; retrieve, select, or format references; interpret data; or draft any statement presented as a finding. All statistical results presented here were calculated from the frozen data extract using the authors' own code and cross-checked against the source tables. The 47 references were located by the authors, and each Digital Object Identifier (DOI) was verified against the official record of the cited work; no reference was included based on an AI system's suggestion. No AI system is listed as an author or met any authorship criteria: all previously declared CRediT roles were performed by identified human authors. The authors have reviewed and edited the generated results and assume full responsibility for the content of this publication, including its accuracy, originality, and the integrity of the presented data and analyses.

Conflicts of Interest: The authors declare the following. J.F.L.T. is employed by Neptura (Líbano, Tolima, Colombia), J.G.C. by Corporación Universitaria Minuto de Dios—UNIMINUTO (Bogotá D.C., Colombia) and E.R. by Universidad Distrital Francisco José de Caldas (Bogotá, Colombia). Neptura developed and operates the platform whose production logs supplied the 191 benchmark strings—a subset of the corpus audited in the companion manuscript [30]—whose recorded outputs over those records provide the deployed-cascade comparator, and whose replicate counts determined which strings were eligible. The authors disclose this as a potential conflict of interest. Neptura provides that platform and not the language models it consumes: what the study evaluates is the estimation cascade the platform buys in—third-party commercial models from independent cloud providers whose prompts, checkpoints and configuration the investigators can neither inspect nor pin (Section 2.5). No author holds any interest in those providers, received support from them, or supplied, selected or altered any audited output. The aim of the study bears on how that conflict could operate: the benchmark was built to establish whether a more reliable configuration exists for this task. It reports two results about the deployed cascade—its test–retest reliability is the lowest of the 21 configurations examined and its internal energy–macronutrient consistency the highest—and publishes the comparator in both its modality-matched and pooled forms, so that the comparison does not turn on the choice of comparator; the design asymmetry favouring the models is declared as a limitation. The potential conflict thus attaches to the provenance of the data and to the use of the platform's recorded performance as comparator, not to the system under evaluation, and the benchmark applies to the authors' own platform on the same terms as to any other—which is how it is being used. No author is an inventor of any planned, pending or awarded patent covering this work, and no author reports honoraria, expert testimony or advisory board membership related to it. The funder had no role in the design of the study, in the collection, analysis or interpretation of data, in the writing of the manuscript, or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

APF	Article processing fee
CI	Confidence interval
CV	Coefficient of variation
ICC	Intraclass correlation coefficient
LLM	Large language model
GRRAS	Guidelines for Reporting Reliability and Agreement Studies

References

1. Howes, E.M.; Laskaridou, E.; Davy, K.P.; Hedrick, V.E. Representation of adults with class III obesity in studies assessing validity of self-reported energy intake using doubly labeled water: A systematic review. *Obes. Rev.* **2024**, *25*, e13713. [[CrossRef](#)] [[PubMed](#)]
2. Biloft-Jensen, A.; Ygil, K.H.; Knudsen, L.; Matthiessen, J.; Fagt, S.; Trolle, E.; Nielsen, T.H.; Hansen, D.M.; Licht, C.L.; Martens, M.; et al. Validation of the 2 × 24 h recall method and a 7-d web-based food diary against doubly labelled water in Danish adults. *Br. J. Nutr.* **2023**, *130*, 1444–1457. [[CrossRef](#)] [[PubMed](#)]

3. Kirkpatrick, S.I.; Troiano, R.P.; Barrett, B.; Cunningham, C.; Subar, A.F.; Park, Y.; Bowles, H.R.; Freedman, L.S.; Kipnis, V.; Rimm, E.B.; et al. Measurement Error Affecting Web- and Paper-Based Dietary Assessment Instruments: Insights From the Multi-Cohort Eating and Activity Study for Understanding Reporting Error. *Am. J. Epidemiol.* **2022**, *191*, 1125–1139. [[CrossRef](#)] [[PubMed](#)]
4. Prentice, R.L.; Pettinger, M.; Neuhouwer, M.L.; Raftery, D.; Zheng, C.; Gowda, G.A.N.; Huang, Y.; Tinker, L.F.; Howard, B.V.; Manson, J.E.; et al. Four-Day Food Record Macronutrient Intake, With and Without Biomarker Calibration, and Chronic Disease Risk in Postmenopausal Women. *Am. J. Epidemiol.* **2022**, *191*, 1061–1070. [[CrossRef](#)] [[PubMed](#)]
5. Bajunaid, R.; Niu, C.; Hambly, C.; Liu, Z.; Yamada, Y.; Aleman-Mateo, H.; Anderson, L.J.; Arab, L.; Baddou, I.; Bandini, L.; et al. Predictive equation derived from 6,497 doubly labelled water measurements enables the detection of erroneous self-reported energy intake. *Nat. Food* **2025**, *6*, 58–71. [[CrossRef](#)] [[PubMed](#)]
6. Ji, Y.; Plourde, H.; Bouzo, V.; Kilgour, R.D.; Cohen, T.R. Validity and Usability of a Smartphone Image-Based Dietary Assessment App Compared to 3-Day Food Diaries in Assessing Dietary Intake Among Canadian Adults: Randomized Controlled Trial. *JMIR mHealth uHealth* **2020**, *8*, e16953. [[CrossRef](#)] [[PubMed](#)]
7. Chotwanvirat, P.; Prachansuwan, A.; Sridonpai, P.; Kriengsinyos, W. Advancements in Using AI for Dietary Assessment Based on Food Images: Scoping Review. *J. Med. Internet Res.* **2024**, *26*, e51432. [[CrossRef](#)] [[PubMed](#)]
8. Nguyen, P.H.; Tran, L.M.; Hoang, N.T.; Truong, D.T.T.; Tran, T.H.T.; Huynh, P.N.; Koch, B.; McCloskey, P.; Gangupantulu, R.; Folson, G.; et al. Relative validity of a mobile AI-technology-assisted dietary assessment in adolescent females in Vietnam. *Am. J. Clin. Nutr.* **2022**, *116*, 992–1001. [[CrossRef](#)] [[PubMed](#)]
9. O'Hara, C.; Kent, G.; Flynn, A.C.; Gibney, E.R.; Timon, C.M. An Evaluation of ChatGPT for Nutrient Content Estimation from Meal Photographs. *Nutrients* **2025**, *17*, 607. [[CrossRef](#)] [[PubMed](#)]
10. Lawabni, R.; Solano, O.R.; Kampen, L.L.; Kurasch, J.; Alshurafa, N.; Chang, J.-S.; Evans, R.; Feng, D.; Lin, A.W. Evaluation of Energy and Nutrient Estimates from Large Language Models Using Text-Based Queries. *J. Nutr.* **2026**, *156*, 101712. [[CrossRef](#)] [[PubMed](#)]
11. Agrawal, S.; Kumar, A.; Singh, R.; Lu, Y. Artificial intelligence in food and nutrition science: A paradigm-centric review of computational frameworks and system-level integration. *Food Funct.* **2026**, *17*, 7908–7941. [[CrossRef](#)] [[PubMed](#)]
12. Hager, P.; Jungmann, F.; Holland, R.; Bhagat, K.; Hubrecht, I.; Knauer, M.; Vielhauer, J.; Makowski, M.; Braren, R.; Kaissis, G.; et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nat. Med.* **2024**, *30*, 2613–2622. [[CrossRef](#)] [[PubMed](#)]
13. Zakka, C.; Shad, R.; Chaurasia, A.; Dalal, A.R.; Kim, J.L.; Moor, M.; Fong, R.; Phillips, C.; Alexander, K.; Ashley, E.; et al. Almanac—Retrieval-Augmented Language Models for Clinical Medicine. *NEJM AI* **2024**, *1*, AIoa2300068. [[CrossRef](#)] [[PubMed](#)]
14. Walker, H.L.; Ghani, S.; Kuemmerli, C.; Nebiker, C.A.; Müller, B.P.; Raptis, D.A.; Staubli, S.M. Reliability of Medical Information Provided by ChatGPT: Assessment Against Clinical Guidelines and Patient Information Quality Instrument. *J. Med. Internet Res.* **2023**, *25*, e47479. [[CrossRef](#)] [[PubMed](#)]
15. Koga, S. Exploring the pitfalls of large language models: Inconsistency and inaccuracy in answering pathology board examination-style questions. *Pathol. Int.* **2023**, *73*, 618–620. [[CrossRef](#)] [[PubMed](#)]
16. Iqbal, U.; Lee, L.T.; Rahmanti, A.R.; Celi, L.A.; Li, Y.-C.J. Can large language models provide secondary reliable opinion on treatment options for dermatological diseases? *J. Am. Med. Inform. Assoc.* **2024**, *1*, 1341–1347. [[CrossRef](#)] [[PubMed](#)]
17. Giuffrè, M.; You, K.; Shung, D. Evaluating ChatGPT in Medical Contexts: The Imperative to Guard Against Hallucinations and Partial Accuracies. *Clin. Gastroenterol. Hepatol.* **2024**, *22*, 1145–1146. [[CrossRef](#)] [[PubMed](#)]
18. Lai, Y.; Liao, F.; Zhao, J.; Zhu, C.; Hu, Y.; Li, Z. Exploring the capacities of ChatGPT: A comprehensive evaluation of its accuracy and repeatability in addressing *Helicobacter pylori*-related queries. *Helicobacter* **2024**, *29*, e13078. [[CrossRef](#)] [[PubMed](#)]
19. Dhar, S.; Kothari, D.S.; Vasquez, M.; Clarke, T.; Maroda, A.; McClain, W.G.; Sheyn, A.; Tuliszewski, R.M.; Tang, D.M.; Rangarajan, S.V. The utility and accuracy of ChatGPT in providing post-operative instructions following tonsillectomy: A pilot study. *Int. J. Pediatr. Otorhinolaryngol.* **2024**, *179*, 111901. [[CrossRef](#)] [[PubMed](#)]
20. Herrmann-Werner, A.; Festl-Wietek, T.; Holderried, F.; Herschbach, L.; Griewatz, J.; Masters, K.; Zipfel, S.; Mahling, M. Assessing ChatGPT's Mastery of Bloom's Taxonomy Using Psychosomatic Medicine Exam Questions: Mixed-Methods Study. *J. Med. Internet Res.* **2024**, *26*, e52113. [[CrossRef](#)] [[PubMed](#)]
21. Bélisle-Pipon, J. Why we need to be careful with LLMs in medicine. *Front. Med.* **2024**, *11*, 1495582. [[CrossRef](#)] [[PubMed](#)]
22. Chen, B.; Zhang, Z.; Langrené, N.; Zhu, S. Unleashing the potential of prompt engineering for large language models. *Patterns* **2025**, *6*, 101260. [[CrossRef](#)] [[PubMed](#)]
23. Korthals, L.; Akrong, E.; Geller, G.; Rosenbusch, H.; Grasman, R.; Visser, I. Towards Reliable LLM Grading Through Self-Consistency and Selective Human Review: Higher Accuracy, Less Work. *Mach. Learn. Knowl. Extr.* **2026**, *8*, 74. [[CrossRef](#)]

24. Maharjan, J.; Garikipati, A.; Singh, N.P.; Cyrus, L.; Sharma, M.; Ciobanu, M.; Barnes, G.; Thapa, R.; Mao, Q.; Das, R. OpenMedLM: Prompt engineering can out-perform fine-tuning in medical question-answering with open-source large language models. *Sci. Rep.* **2024**, *14*, 14156. [[CrossRef](#)] [[PubMed](#)]
25. Lipsmeier, F.; Taylor, K.I.; Postuma, R.B.; Volkova-Volkmar, E.; Kilchenmann, T.; Mollenhauer, B.; Bamdadian, A.; Popp, W.L.; Cheng, W.-Y.; Zhang, Y.-P.; et al. Reliability and validity of the Roche PD Mobile Application for remote monitoring of early Parkinson's disease. *Sci. Rep.* **2022**, *12*, 12081. [[CrossRef](#)] [[PubMed](#)]
26. Chan, L.L.Y.; Choi, T.C.M.; Lord, S.R.; Brodie, M.A. Development and large-scale validation of the Watch Walk wrist-worn digital gait biomarkers. *Sci. Rep.* **2022**, *12*, 16211. [[CrossRef](#)] [[PubMed](#)]
27. Ohyama, T. A comparison of confidence interval methods for the intraclass correlation coefficient based on the one-way random effects model. *Jpn. J. Stat. Data Sci.* **2025**, *8*, 587–602. [[CrossRef](#)]
28. Haghayegh, S.; Kang, H.; Khoshnevis, S.; Smolensky, M.H.; Diller, K.R. A comprehensive guideline for Bland–Altman and intra class correlation calculations to properly compare two methods of measurement and interpret findings. *Physiol. Meas.* **2020**, *41*, 055012. [[CrossRef](#)] [[PubMed](#)]
29. Xu, M.; Reiss, P.T.; Cribben, I. Generalized reliability based on distances. *Biometrics* **2020**, *77*, 258–270. [[CrossRef](#)] [[PubMed](#)]
30. Lopez Torres, J.F.; Castañeda, J.G.; Rivas, E. Internal consistency does not validate an LLM-assisted nutrition pipeline: A reference-standard-free forensic reliability audit of 13,073 free-text meal records. *Nutrients* **2026**, *18*, 3233. [[CrossRef](#)]
31. Tejani, A.S.; Klontzas, M.E.; Gatti, A.A.; Mongan, J.T.; Moy, L.; Park, S.H.; Kahn, C.E.; Panel, F.T.C.2.U.; Abbata, S.; Afat, S.; et al. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): 2024 Update. *Radiol. Artif. Intell.* **2024**, *6*, e240300. [[CrossRef](#)] [[PubMed](#)]
32. Padula, W.V.; Kreif, N.; Vanness, D.J.; Adamson, B.; Rueda, J.-D.; Felizzi, F.; Jonsson, P.; Ijzerman, M.J.; Butte, A.; Crown, W. Machine Learning Methods in Health Economics and Outcomes Research—The PALISADE Checklist: A Good Practices Report of an ISPOR Task Force. *Value Health* **2022**, *25*, 1063–1080. [[CrossRef](#)] [[PubMed](#)]
33. Morris, T.P.; White, I.R.; Crowther, M.J. Using simulation studies to evaluate statistical methods. *Stat. Med.* **2019**, *38*, 2074–2102. [[CrossRef](#)] [[PubMed](#)]
34. Kottner, J.; Audigé, L.; Brorson, S.; Donner, A.; Gajewski, B.J.; Hróbjartsson, A.; Roberts, C.; Shoukri, M.; Streiner, D.L. Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. *J. Clin. Epidemiol.* **2011**, *64*, 96–106. [[CrossRef](#)] [[PubMed](#)]
35. Menezes, M.C.S.; Hoffmann, A.F.; Tan, A.L.M.; Nalbandyan, M.; Omenn, G.S.; Mazzotti, D.R.; Hernández-Arango, A.; Visweswaran, S.; Venkatesh, S.; Mandl, K.D.; et al. The potential of Generative Pre-trained Transformer 4 (GPT-4) to analyse medical notes in three different languages: A retrospective model-evaluation study. *Lancet Digit. Health* **2025**, *7*, e35–e43. [[CrossRef](#)] [[PubMed](#)]
36. Sanghera, R.; Thirunavukarasu, A.J.; El Khoury, M.; O'Logbon, J.; Chen, Y.; Watt, A.; Mahmood, M.; Butt, H.; Nishimura, G.; Soltan, A.A.S. High-performance automated abstract screening with large language model ensembles. *J. Am. Med. Inform. Assoc.* **2025**, *32*, 893–904. [[CrossRef](#)] [[PubMed](#)]
37. Sorin, V.; Collins, J.D.; Bratt, A.K.; Kusmirek, J.E.; Mugu, V.K.; Kline, T.L.; Butler, C.L.; Wood, N.G.; Cook, C.J.; Korfiatis, P. Evaluating prompt and data perturbation sensitivity in large language models for radiology reports classification. *JAMIA Open* **2025**, *8*, ooaf073. [[CrossRef](#)] [[PubMed](#)]
38. Li, D.; Wu, L.; Zhang, M.; Shpileva, S.; Lin, Y.-C.; Huang, H.-Y.; Li, T.; Xu, J. Assessing the performance of large language models in literature screening for pharmacovigilance: A comparative study. *Front. Drug Saf. Regul.* **2024**, *4*, 1379260. [[CrossRef](#)] [[PubMed](#)]
39. Shrout, P.E.; Fleiss, J.L. Intraclass correlations: Uses in assessing rater reliability. *Psychol. Bull.* **1979**, *86*, 420–428. [[CrossRef](#)] [[PubMed](#)]
40. McGraw, K.O.; Wong, S.P. Forming inferences about some intraclass correlation coefficients. *Psychol. Methods* **1996**, *1*, 30–46. [[CrossRef](#)]
41. Koo, T.K.; Li, M.Y. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J. Chiropr. Med.* **2016**, *15*, 155–163. [[CrossRef](#)] [[PubMed](#)]
42. Seydi, M.R.; Pini, A.; Pataky, T.C.; Schelin, L. Confidence sets for intraclass correlation coefficients in test–retest curve measurements. *J. Biomech.* **2024**, *173*, 112232. [[CrossRef](#)] [[PubMed](#)]
43. Merrill, A.L.; Watt, B.K. *Energy Value of Foods: Basis and Derivation*; Agriculture Handbook No. 74; US Department of Agriculture: Washington, DC, USA, 1973.
44. Dhanvijay, A.K.D.; Pinjar, M.J.; Dhokane, N.; Sorte, S.R.; Kumari, A.; Mondal, H.; Dhanvijay, A.K. Performance of Large Language Models (ChatGPT, Bing Search, and Google Bard) in Solving Case Vignettes in Physiology. *Cureus* **2023**, *15*, e42972. [[CrossRef](#)] [[PubMed](#)]
45. Gallifant, J.; Afshar, M.; Ameen, S.; Aphinyanaphongs, Y.; Chen, S.; Cacciamani, G.; Demner-Fushman, D.; Dligach, D.; Daneshjou, R.; Fernandes, C.; et al. The TRIPOD-LLM reporting guideline for studies using large language models. *Nat. Med.* **2025**, *31*, 60–69. [[CrossRef](#)] [[PubMed](#)]

46. Vega-Salas, M.J.; Curi-Quinto, K.; Hidalgo-Aréstegui, A.; Meza-Carbajal, K.; Lago-Berrocal, N.; Arias, L.; Favara, M.; Penny, M.; Sánchez, A.; Vimalaswaran, K.S. Development of an online food frequency questionnaire and estimation of misreporting of energy intake during the COVID-19 pandemic among young adults in Peru. *Front. Nutr.* **2022**, *9*, 949330. [[CrossRef](#)] [[PubMed](#)]
47. Nunes, C.L.; Jesus, F.; Oliveira, M.V.; Thomas, D.M.; Sardinha, L.B.; Martins, P.; Minderico, C.S.; Silva, A.M. The impact of body composition on the degree of misreporting of food diaries. *Eur. J. Clin. Nutr.* **2023**, *78*, 209–216. [[CrossRef](#)] [[PubMed](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.