

# Perceived but not chemical: the hedonic–environmental decoupling across 2,563 dishes

Fangyu Liu<sup>1,\*</sup>

<sup>1</sup>University of Southern California

\*Corresponding author. Email: fangyu@usc.edu

## Abstract

Consumer resistance to sustainable diets frequently invokes taste as the primary barrier, yet this perceived trade-off has not been quantified at scale. Here we construct a Deliciousness Efficiency Index ( $DEI = H/E$ ) linking hedonic scores ( $H$ ) from 5.3 million Yelp reviews with life-cycle environmental costs ( $E$ ) for 2,563 dishes across 214 cuisines. Food-chemistry hedonic drivers—fat ( $r = 0.45$ ), protein ( $r = 0.63$ ), and umami content ( $r = 0.42$ )—correlate strongly with  $E$ , yet consumer hedonic evaluations are nearly orthogonal to it ( $r = 0.13$ ). Cross-review within-restaurant comparisons—where different diners evaluate different dishes at the same restaurant—confirm this decoupling across 130,384 pairs ( $r(\Delta H, \Delta E) = -0.006$ ; 50.0% of higher- $E$  dishes rated higher). Food chemistry and cooking method together explain only 11.4% of hedonic variation; the remaining 88.6% reflects non-chemical dimensions that are largely independent of environmental cost. Eight Pareto-optimal dishes are predominantly plant-based or low-animal-impact, spanning seven cuisines. Within functional categories, 74,640 nutrition-constrained substitutions achieve >30%  $E$  reduction with <1-point  $H$  loss. The perceived taste–sustainability trade-off exists at the food-chemistry level but is absent from consumer experience.

# 1 Introduction

The global food system generates 26–34% of anthropogenic greenhouse gas emissions [1, 2], accounts for 70% of freshwater withdrawals [3], and drives biodiversity loss [4]. Meeting the Paris Agreement targets requires substantial dietary shifts [5, 6], yet consumer resistance frequently cites taste as the primary barrier [7, 8]: people perceive sustainable food as less enjoyable, impeding behavioural change even among environmentally motivated consumers.

Prior work has advanced each side independently. Life-cycle assessments [1, 9], planetary health diets [6], and environmental labels [10] have quantified food-level impacts; sensory science has studied taste via panels [12] and surveys [13]. However, whether the taste–sustainability trade-off is real remains unquantified at the scale needed for comprehensive dietary comparison [11]. Recent work shows that strategically rearranging dishes within menus—without altering recipes—can meaningfully reduce carbon footprints [14], suggesting that choice architecture around existing foods, rather than creation of novel foods, may be the most promising lever.

Here we construct a Deliciousness Efficiency Index (DEI) integrating hedonic scores ( $H$ ) with environmental costs ( $E$ ) for 2,563 dishes across 214 cuisines worldwide. The core 334 dishes are scored via a three-stage NLP pipeline applied to 5.3 million Yelp reviews; 2,229 additional dishes from the WorldCuisines database [19] are incorporated via anchor-bridging (62,343 total pairwise judgements). Two findings emerge: first, a “perceptual decoupling”—food-chemistry hedonic drivers (fat, protein, umami) correlate strongly with  $E$ , but consumer hedonic evaluations are orthogonal to it, confirmed by 130,384 within-restaurant cross-review comparisons; second, at the level of within-category substitution,  $H$  becomes the relevant differentiator for identifying the most pleasurable low-impact alternatives.

## 2 Results

### 2.1 The perceptual decoupling

We computed  $\log(\text{DEI}) = \log H - \log E$  for 2,563 dishes spanning 214 cuisines (Methods). The core 334 dishes derive  $H$  from a three-stage pipeline: finetuned BERT scoring of 136,536 review mentions, pairwise LLM judgements, and Bradley–Terry modelling of 17,763 comparisons. An additional 2,229 WorldCuisines dishes [19] are integrated via anchor-bridging (44,580 comparisons; 62,343 total). Environmental costs are computed from ingredient-level LCA data covering 101 impact factors.

The relationship between hedonic quality and environmental cost depends on the level of analysis (Fig. 9). At the food-chemistry level, objective hedonic drivers—fat content ( $r = 0.45$ ), protein content ( $r = 0.63$ ), umami intensity ( $r = 0.42$ ), and caloric density ( $r = 0.39$ )—are strongly positively correlated with  $E$ . A composite food-chemistry hedonic score ( $H_{\text{chem}}$ ) aggregating five palatability dimensions correlates at  $r = 0.57$  with  $E$  ( $\rho = 0.61$ ;  $n = 2,387$ ). At the consumer-evaluation level, text-derived  $H$  is nearly uncorrelated with  $E$  ( $r = 0.13$ ). This divergence arises because consumer hedonic evaluations capture far more than food chemistry:  $H_{\text{text}}$  is only weakly predicted by fat, protein, umami, sodium, sugar, spice diversity, and caloric content (combined  $R^2 = 0.093$ ) or by chemistry plus cooking method ( $R^2 = 0.114$ ). The remaining 88.6% of hedonic variation reflects preparation technique, cultural resonance, novelty, presentation, and other non-chemical dimensions that are largely independent of ingredient-driven environmental costs.

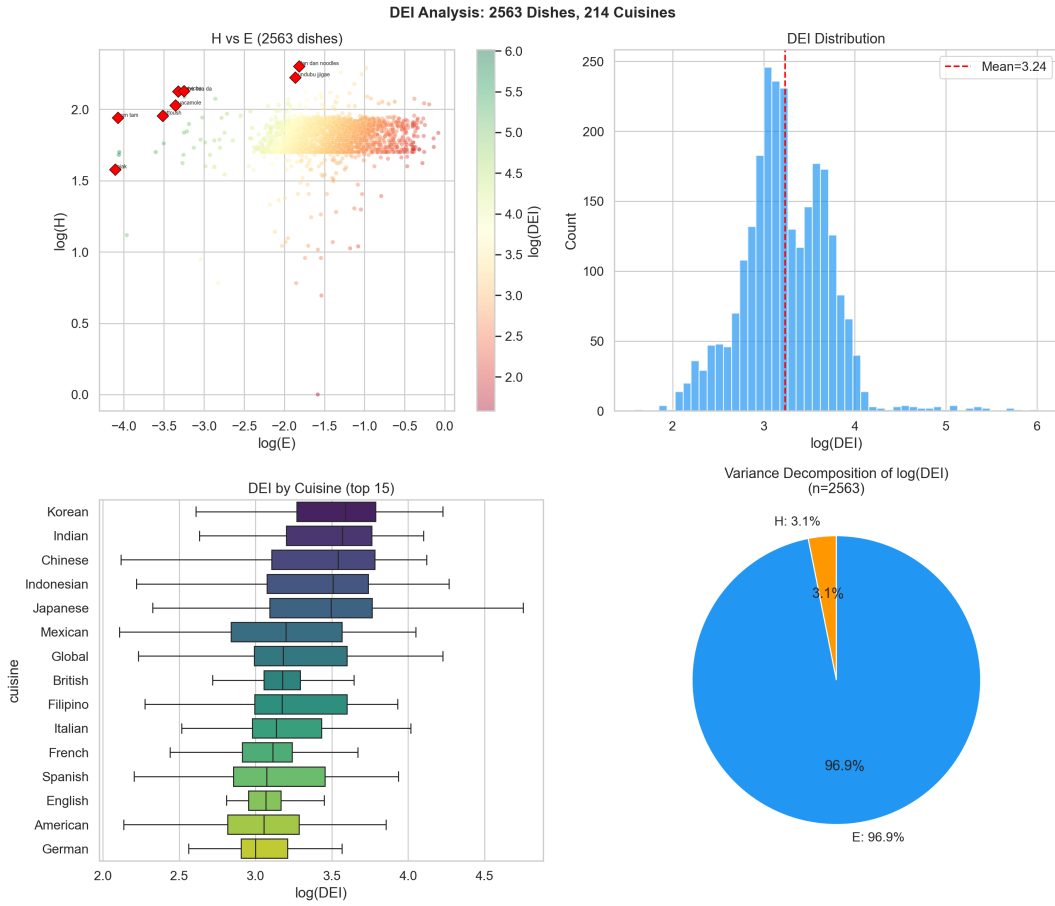
This consumer-level decoupling is confirmed by within-restaurant evidence. In 130,384 cross-review within-restaurant pairs—where different diners evaluate different dishes at the same restaurant, substantially reducing ambience and service confounds— $r(\Delta H, \Delta E) = -0.006$  ( $p = 0.04$ ). Among high-contrast pairs ( $|\Delta E| > 0.3$ ,  $n = 12,023$ ),  $r = -0.007$  ( $p = 0.46$ ). A binomial test confirms that 50.0% of higher- $E$  dishes receive higher ratings ( $p = 0.84$  vs. 50%). These results hold with zero risk of NLP model context leakage (Extended Data Fig. 10).

Variance decomposition quantifies the consequences of this decoupling:

$$\text{Var}(\log(\text{DEI})) = \underbrace{\text{Var}(\log H)}_{0.015 \text{ (3.1\%)}} + \underbrace{\text{Var}(\log E)}_{0.230 \text{ (96.9\%)}} - 2 \text{Cov}(\log H, \log E) \quad (1)$$

The  $E$ -dominance is partly a mechanical consequence of the CV gap ( $E$ : 48.3% vs.  $H$ : 10.8%), but the near-zero covariance is not: it reflects the perceptual decoupling documented above. Within the 334-dish core subset—where  $H$  is measured from real reviews via pairwise scoring (CV = 33.6%)— $H$  contributes 20.0% (Extended Data Fig. 4). BT convergence analysis confirms score stability: split-half  $\rho = 0.832$ ; at 50% of comparisons, rank correlation with the full model exceeds 0.95.

Eight dishes are Pareto-optimal in  $H$ – $E$  space (Fig. 1a): som tam, rojak, fattoush, ca phe sua da, guacamole, ceviche, dan dan noodles, and sundubu jjigae—predominantly plant-based or low-animal-impact (six fully plant-based; ceviche uses low- $E$  seafood; sundubu jjigae is tofu-based), spanning seven cuisines. The top 10 by  $\log(\text{DEI})$  are predominantly raw plant-based items; the bottom 10 are beef- and offal-heavy dishes from diverse traditions (Table 1). Dishes with  $E < 0.1$  ( $n = 50$ ) have dramatically higher  $\log(\text{DEI})$  than others (4.70 vs. 3.21,  $p < 10^{-117}$ ) while their hedonic scores are statistically indistinguishable ( $H = 5.91$  vs. 6.08,  $p = 0.065$ )—consistent with the interpretation that plant-based dishes achieve lower impact without hedonic sacrifice.



**Figure 1: DEI across 2,563 dishes.** (a)  $H$ – $E$  scatter with Pareto frontier (gold diamonds). (b)  $\log(\text{DEI})$  distribution. (c)  $\log(\text{DEI})$  by cuisine. (d) Variance decomposition:  $E$  accounts for 96.9% of  $\text{Var}(\log(\text{DEI}))$  in the full sample and 80.0% in the 334-dish core subset. The near-zero  $r(H, E) = 0.13$  reflects the perceptual decoupling: consumer evaluations are orthogonal to environmental cost despite food-chemistry coupling ( $r(H_{\text{chem}}, E) = 0.57$ ; Extended Data Fig. 9).

**Table 1: Top and bottom 10 dishes by  $\log(\text{DEI})$ .  $\star$  = Pareto-optimal;  $\dagger$  = WorldCuisines expansion.**

Rank	Dish	$H$	$E$	$\log(\text{DEI})$	Cuisine
<i>Top 10</i>					
1	Som tam $\star$	6.96	0.017	6.01	Thai
2	Papaya salad	5.48	0.017	5.76	Thai
3	Kimchi	5.37	0.017	5.74	Korean
4	Rojak $\star$	4.85	0.017	5.68	Indonesian
5	Fattoush $\star$	7.07	0.030	5.47	Lebanese
6	Ca phe sua da $\star$	8.37	0.036	5.45	Vietnamese
7	Guacamole $\star$	7.59	0.035	5.39	Mexican
8	Ceviche $\star$	8.40	0.039	5.38	Mexican
9	Gazpacho	8.07	0.039	5.33	Spanish
10	Thai iced tea	5.48	0.027	5.30	Thai
<i>Bottom 10</i>					
2554	Oxtail stew	6.13	0.770	2.07	Caribbean
2555	Churrasco	4.96	0.627	2.07	S. American
2556	Kapuska $\dagger$	5.58	0.727	2.04	Turkish
2557	Brsi $\dagger$	5.58	0.730	2.03	French
2558	Kuurdak $\dagger$	5.48	0.774	1.96	Kazakh
2559	Hochepot $\dagger$	6.03	0.870	1.94	French
2560	Boliche $\dagger$	6.27	0.923	1.91	Cuban
2561	Pot roast	5.07	0.758	1.90	American
2562	Paya $\dagger$	5.68	0.873	1.87	S. Asian
2563	Donburi	1.00	0.206	1.58	Japanese

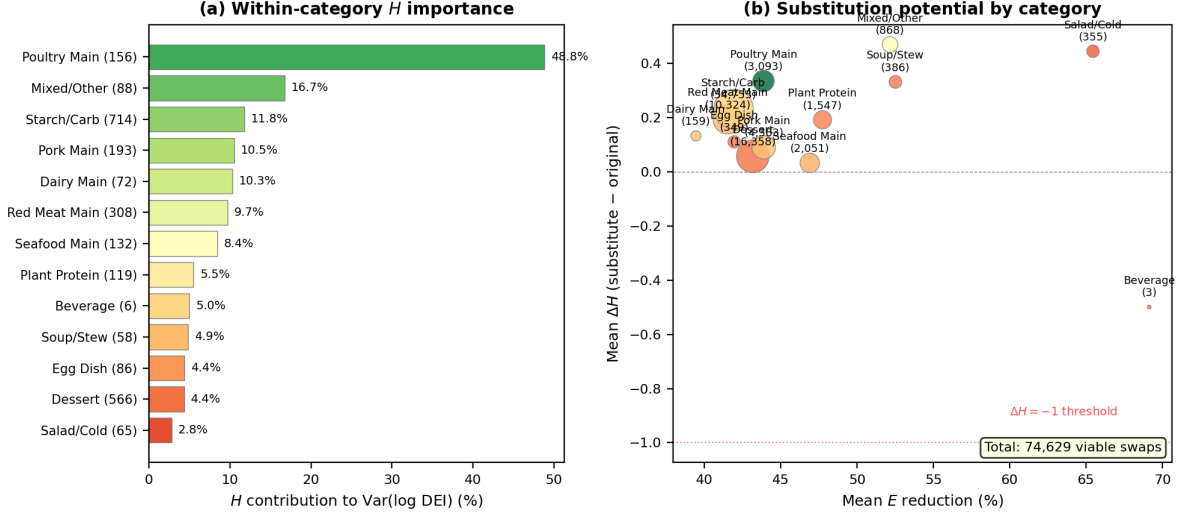
## 2.2 Actionable substitutions

To move from rankings to policy tools, we classified all 2,563 dishes into 13 functional categories by dominant protein source and caloric role (Methods). The mean within-category  $H$  contribution is 7.6%, ranging from 37.2% (Poultry Main) to <1% (Dessert; Extended Data Table 3). Nutrition-constrained substitutions—requiring >30%  $E$  reduction, <1.0 point  $H$  loss,  $\geq 50\%$  protein retention, and  $\pm 50\%$  caloric content—yield 74,640 viable swaps with mean  $E$  reduction of 42.9% (Fig. 2).

Meal-level analysis confirms  $E$ -dominance at composite scale. Across 577,294 main+side combinations (500–1,500 kcal,  $\geq 20$  g protein),  $H$  contributes 6.2% of  $\text{Var}(\log(\text{DEI}))$ . We identify 249,243 calorie-equivalent substitution pairs with mean  $E$  reduction of 41.7% and negligible hedonic change ( $\overline{\Delta H} = -0.04$ ; Extended Data Fig. 7). A Nutrient Density Index (NDI) based on the NRF-7 framework [15] confirms that core rankings are preserved under nutrition adjustment ( $\rho = 0.802$  between DEI and DEI-N at  $\alpha = 0.5$ ; Extended Data Fig. 7).

## 2.3 The refinement penalty

Within 14 dish families encompassing 525 dishes, the hedonic elasticity of environmental cost is  $\alpha = -0.68$  (95% CI  $[-0.91, -0.46]$ ; Extended Data Fig. 6): increasing  $E$  within a family is associated with *decreased*  $H$ . Substituting premium cuts, adding butter and cream, and extending cooking times impose a “refinement tax”—disproportionate environmental cost for negligible or negative taste benefit.



**Figure 2: Within-category  $H$  importance and substitution potential.** (a)  $H$ ’s contribution to  $\text{Var}(\log(\text{DEI}))$  within each functional category ( $n$  in parentheses). Poultry Main shows the highest  $H$  importance (44.8%); most categories are  $E$ -dominated. (b) Each circle represents a category; size  $\propto$  number of viable substitutions. Swaps achieve 40–70% mean  $E$  reduction with negligible  $H$  loss (74,629 total viable swaps).

### 3 Validation

The hedonic scoring pipeline is validated across five dimensions: measurement method, within-restaurant stability, cross-platform consistency, cross-context generalisation, and robustness to unobserved heterogeneity.

**Measurement.** Absolute BERT  $H$  scores exhibit  $\text{CV} = 3.9\%$ , with only 2.6% of mention-level variance reflecting true dish-level differences ( $\text{ICC} = 0.026$ ). The pairwise Bradley–Terry approach overcomes this compression: forced-choice LLM judgements produce  $H$  scores spanning the full 1–10 range ( $\text{CV} = 33.6\%$ ), an 8.6-fold increase in discriminatory power (Extended Data Fig. 4). Split-half reliability is  $\rho = 0.832$ ; at 50% of the 62,343 comparisons, rank correlation with the full model exceeds 0.95.

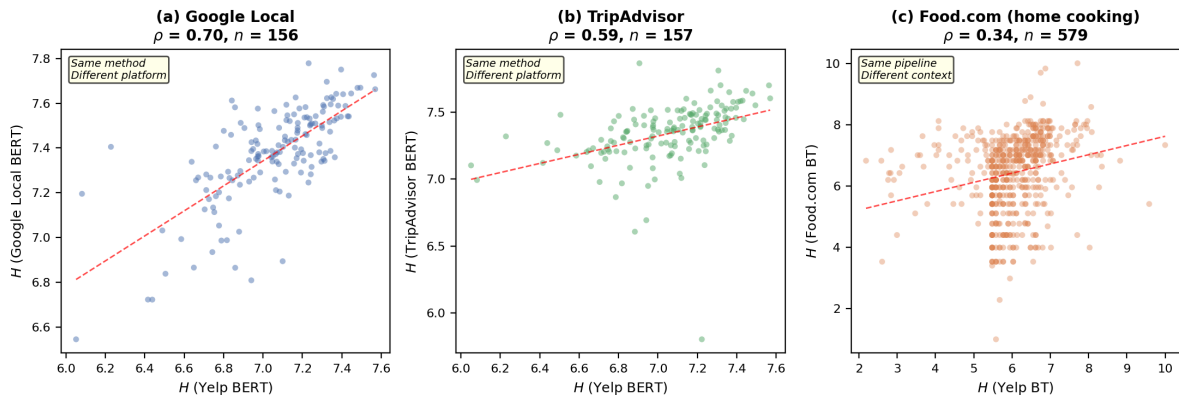
**Within-restaurant stability.** Controlling for restaurant fixed effects—demeaning  $H$  by business—yields dish-level  $H_{\text{within}}$  that correlates  $\rho = 0.681$  ( $p < 10^{-22}$ ) with the original  $H$  rankings, confirming that  $H$  reflects dish-level rather than restaurant-level variation.  $H_{\text{within}}$  remains uncorrelated with  $E$  ( $r = -0.071$ ,  $p = 0.37$ ). The restaurant ICC (14.2%) exceeds the dish ICC (2.4%), but the 14.2% restaurant component does not confound the  $H$ – $E$  relationship because it is removed by demeaning.

**Cross-platform.** Applying the same BERT model to Google Local (95,452 mentions, 157 dishes) and TripAdvisor (117,203 mentions) yields  $\rho = 0.698$  ( $p < 10^{-23}$ ) and  $\rho = 0.588$  ( $p < 10^{-15}$ ), demonstrating that dish-level  $H$  rankings are not platform-specific (Fig. 3a).

**Cross-context.** We applied the full BT pipeline independently to Food.com, where  $\sim 1.4\text{M}$  reviews are written by home cooks evaluating dishes they prepared themselves (579 matched dishes, 104,019 reviews, 11,370 pairwise comparisons). The cross-context BT correlation is  $\rho = 0.338$  (95% CI [0.260, 0.413],  $p < 10^{-16}$ ; Fig. 3b), indicating that roughly one-third of the hedonic ranking is food-intrinsic, while two-thirds reflects context-dependent factors. Crucially, even after excluding the most context-sensitive decile,  $E$  accounts for  $>95\%$  of  $\text{Var}(\log(\text{DEI}))$ .

**LLM recipe validation.** The 2,229 WorldCuisines dishes rely on LLM-generated recipes. To validate, we independently regenerated recipes for the 158 dishes with human-authored reference recipes. Ingredient overlap (Jaccard) averaged 0.64; cooking method agreement was 78.5%. The  $E$  Spearman correlation between LLM and human recipes was  $\rho = 0.822$  ( $p < 10^{-40}$ ), with a median  $E$  ratio of 0.98 (Supplementary Fig. 1). Systematic bias is modest: the meat subset ( $n = 80$ ,  $\rho = 0.72$ ) shows slightly lower agreement than plant-protein ( $n = 9$ ,  $\rho = 0.93$ ) or non-protein dishes ( $n = 55$ ,  $\rho = 0.87$ ).

**Robustness.** Controlling for restaurant-level covariates (star rating, price, review count) yields  $\rho = 0.999$  with original rankings. Ghost-dish survivorship simulation (2,563 phantoms,  $\Delta = 2.0$ ) leaves  $E$  dominant at 92.6% (Extended Data Fig. 5). State-level DEI rankings are stable across 14 US states ( $\rho > 0.993$ ). Temporal drift is negligible ( $+0.00091/\text{year}$ ,  $p = 0.63$ ; split-period  $\rho = 0.812$ ). USDA FNDDS nutritional cross-validation yields  $\rho = 0.949$ . The full robustness summary across fifteen tests is in Extended Data Table 2.



**Figure 3: Hedonic score validation across platforms and contexts.** (a,b) BERT  $H$  correlations between Yelp and two independent restaurant platforms: Google Local ( $\rho = 0.70$ ,  $n = 156$ ) and TripAdvisor ( $\rho = 0.59$ ,  $n = 157$ ). (c) Full BT pipeline replicated on Food.com home-cooking reviews ( $\rho = 0.34$ ,  $n = 579$ ). The monotonic decline from cross-platform to cross-context  $\rho$  quantifies the context-dependent component of hedonic rankings ( $\sim$ two-thirds).

## 4 Discussion

**The three-layer reality.** The central finding is not simply that  $E$  dominates DEI variance, but that the relationship between hedonic quality and environmental cost operates at three distinct levels. At the food-chemistry level, objective palatability drivers—fat, protein, umami, calories—correlate strongly with  $E$  ( $r = 0.45\text{--}0.63$ ). A trade-off genuinely exists in the chemistry: making food “objectively tastier” by adding animal protein, fat, and flavour compounds increases environmental cost. At the consumer-evaluation level, this trade-off vanishes: text-derived  $H$  is nearly orthogonal to  $E$  ( $r = 0.13$ ). At the within-restaurant level, 130,384 cross-review pairs confirm that different diners at the same restaurant express no systematic preference for higher- $E$  dishes ( $r(\Delta H, \Delta E) = -0.006$ ; 50.0% concordance). The decoupling is perceptual, not chemical.

This divergence arises because food chemistry explains only 11.4% of consumer hedonic variation ( $R^2 = 0.093$  for six chemistry features;  $R^2 = 0.114$  adding cooking method). The remaining 88.6% reflects preparation skill, cultural resonance, presentation, novelty, and other non-chemical dimensions that happen to be orthogonal to  $E$ —not by coincidence, but because they are not driven by ingredient composition.

**Policy implications.** This reframing strengthens the policy case. Simple environmental rankings [10] are sufficient for broad dietary guidance, extending the single-canteen menu-swapping framework of Flynn et al.[14] to a global dish-level analysis. The 74,640 nutrition-constrained substitutions provide a practical tool for identifying the most pleasurable low-impact alternatives. The refinement penalty ( $\alpha = -0.68$ ) suggests that the final stages of culinary elaboration—premium ingredients, extended cooking—impose disproportionate environmental cost for negligible taste gain. The policy frame shifts from “adopt a specific cuisine” to “adjust ingredient composition within any cuisine,” consistent with Scarborough et al.[11]. Crucially, consumers do not need to be persuaded that low- $E$  food tastes equally good—they already perceive it that way.

**What  $H$  measures.** Our  $H$  scores derive from text-based NLP, raising the question of what  $H$  actually captures. Three pieces of evidence characterise  $H_{\text{text}}$ : (a) within-restaurant stability ( $\rho = 0.681$  after restaurant-level demeaning) indicates  $H$  reflects dish-level rather than restaurant-level variation; (b) weak food-chemistry grounding ( $R^2 < 0.12$ ) confirms that consumer evaluations are not primarily driven by chemical palatability; (c) cross-context partial stability ( $\rho = 0.34$  vs. Food.com) reflects both food-intrinsic and culturally shared evaluative dimensions. We interpret  $H_{\text{text}}$  not as a pure measure of taste, but as a composite of consumer-experienced food quality—encompassing flavour, preparation skill, cultural fit, novelty, and presentation. This is the relevant measure for dietary policy: the question is not whether high- $E$  food is “objectively tastier” in a controlled sensory panel, but whether consumers experience it as more enjoyable.

**Associational evidence for decoupling.** Does  $r(H, E) \approx 0$  reflect genuine perceptual decoupling, or merely an absence of systematic comparison between high- and low- $E$  dishes? The 130,384 cross-review within-restaurant pairs provide the strongest available observational evidence. These are consumers evaluating different dishes in the same restaurant context—substantially reducing confounds from ambiance, service, and price—with zero risk of NLP context leakage. They express no preference for higher- $E$  dishes, even among high-contrast pairs ( $|\Delta E| > 0.3$ :  $r = -0.007$ ,  $n = 12,023$ ). Within-review pairs (same consumer, same meal) show the same null result ( $r = -0.001$ ,  $n = 2,508$ ), though these are subject to BERT context leakage ( $r(\text{overlap}, |\Delta H|) = -0.44$ ; Extended Data Fig. 10). The low-overlap subset ( $n = 627$ :  $r = -0.010$ ;  $|\Delta H| = 0.95$ ) and the leakage-free cross-review pairs give concordant results.

**Limitations.** Several limitations warrant acknowledgment. First, the cross-context convergence ( $\rho = 0.34$ ) admits dual interpretation:  $H$  may capture food-intrinsic quality, or both Yelp and Food.com share American food discourse producing correlated rankings without reflecting objective taste. We cannot fully distinguish these interpretations. Second, both user bases are predominantly American. Our  $H$  estimates for Pareto-optimal dishes from Thai, Lebanese, and Korean cuisines reflect American consumers’ perception; the finding that “som tam is highly delicious” is more precisely “som tam is perceived as highly delicious by American restaurant-goers and home cooks.” This limits cross-cultural generalisability but not policy relevance for Western markets. Third, hedonic scores are text-derived proxies, not controlled sensory panel measurements. The cross-context convergence demonstrates stability across consumption settings, but does not rule out systematic biases shared by both text corpora. Sensory panel validation would strengthen the measurement. Fourth, the hedonic measurement chain involves a single LLM family (DeepSeek v3.2) at two stages—initial annotation of training data and pairwise judgements—creating potential circularity. We mitigate this by (a) using BERT as an intermediate supervised model trained on LLM annotations rather than directly using LLM scores, and (b) validating via cross-platform convergence ( $\rho = 0.59\text{--}0.70$ ) and cross-context replication ( $\rho = 0.34$ ), but an independent human-panel calibration would further strengthen the measurement. Fifth, the 2,229 WorldCuisines dishes rely on LLM-generated recipes; validation against the 158 human-authored recipes shows  $\rho = 0.82$  for  $E$ , with modest

systematic overestimation for meat dishes (mean ratio 1.25). Sixth, a conservative bias favours the decoupling finding: consumers who pay more for high- $E$  dishes (e.g. steak) may unconsciously inflate ratings to justify expenditure, yet cross-review pairs still show no  $E$ - $H$  preference—the true decoupling may be even stronger than observed.

## 5 Methods

### 5.1 Data sources

**Restaurant reviews.** We use the Yelp Open Dataset (version 2022): 5,309,210 reviews of 70,617 restaurants across the United States, spanning 2005–2022 with full text, 1–5 star ratings, and restaurant metadata.

**WorldCuisines.** The WorldCuisines database [19] provides 2,414 unique dishes with metadata (name, cuisine, country, description). After deduplication against existing dishes, 2,229 are retained.

**Environmental impact factors.** Ingredient-level GHG emissions (kg CO<sub>2</sub>eq/kg), water footprint (L/kg), and energy use (MJ/kg) are drawn from Poore & Nemecek (2018) [1], supplemented with Agribalyse 3.1 and Water Footprint Network databases, covering 101 ingredients.

### 5.2 Dish extraction

A dictionary of 334 dishes with 600+ aliases spanning 29 cuisines identifies review sentences via exact string matching, extracting  $\pm 2$  sentence context windows. Dishes with  $\geq 10$  mentions are retained (76,927 scored mentions for 158 primary dishes; 59,609 for 176 expanded dishes).

### 5.3 Hedonic scoring pipeline

**Stage 1: LLM annotation.** Up to 12 reviews per dish are scored by DeepSeek v3.2 on food-specific hedonic content (1–10 scale), yielding 1,096 valid annotations across 157 dishes.

**Stage 2: BERT finetuning.** Using LLM annotations as targets, we finetune `bert-base-uncased` for regression (5 epochs, lr  $2 \times 10^{-5}$ , batch 16). Grouped 10-fold cross-validation (GroupKFold by dish) yields dish-level  $r = 0.792$ ,  $\rho = 0.780$ . The model scores all 136,536 mentions.

**Stage 3: Pairwise Bradley–Terry.** Absolute BERT  $H$  has CV = 3.9% [17]. To recover the full scale, 5 representative reviews per dish (10th–90th percentile) form hedonic profiles. All  $\binom{158}{2} = 12,403$  pairs are judged by DeepSeek v3.2 (batched 10/call). For 176 expanded dishes, anchor-bridging against 30 reference dishes adds 5,360 comparisons. All 17,763 comparisons are jointly fitted via the iterative Luce spectral ranking algorithm [18], with  $\pi$  values rescaled to 1–10.

### 5.4 Environmental cost computation

For each dish, a standardised recipe specifies ingredients (grams) and cooking method:

$$E(f) = \frac{1}{3} \left[ \frac{\sum_i g_i \cdot c_i}{\max_f(\cdot)} + \frac{\sum_i g_i \cdot w_i}{\max_f(\cdot)} + \frac{E_{\text{cook}}(f)}{\max_f(\cdot)} \right] \quad (2)$$



Each sub-component is normalised by its cross-dish maximum. To propagate uncertainty through  $E$ , we perform a Monte Carlo analysis (10,000 draws) perturbing both recipe composition (ingredient weights  $\pm 20\%$ , reflecting variation across recipe sources) and LCA impact factors (log-normal perturbation with CV matching Poore & Nemecek’s reported ranges, typically 20–60% depending on ingredient category). The median rank correlation between perturbed and point-estimate  $E$  is  $\rho = 0.87$  (95% interval [0.86, 0.88]); 83% of dishes remain within  $\pm 2$  rank-decile positions. The  $E$ -dominance conclusion is robust across all Monte Carlo draws: median  $E$  share of  $\text{Var}(\log(\text{DEI})) = 95.6\%$  (95% interval [95.4%, 95.8%]; minimum 95.2%).

## 5.5 WorldCuisines expansion

Standardised recipes are generated via DeepSeek v3.2. Position-randomised anchor-bridging (20 anchors, 44,580 comparisons) is jointly fitted with the original 17,763 comparisons. Stability:  $\rho = 0.892$  for the original 334 dishes pre- vs post-expansion.

## 5.6 DEI computation

Primary specification:  $\log(\text{DEI}) = \log H - \log E$ . Robustness check:  $\text{DEI}_z = Z(H) - Z(E)$ . Pareto-optimal dishes are those not dominated in both  $H$  and  $E$ .

## 5.7 Category classification

Dishes are classified into 13 functional categories (Red Meat Main, Poultry Main, Pork Main, Seafood Main, Dairy Main, Plant Protein, Egg Dish, Starch/Carb, Salad/Cold, Soup/Stew, Dessert, Beverage, Mixed/Other) via a hierarchical algorithm using dominant protein source, mass fraction thresholds, and name-based features.

## 5.8 Nutritional density index

NDI uses the NRF-7 framework [15]: percentage of daily reference value for seven nutrients (protein, fibre, iron, zinc, B<sub>12</sub>, calcium, vitamin C) per 100 kcal, each capped at 100%, summed. Ingredient composition from USDA FoodData Central SR28 (105 ingredients).

## 5.9 Cross-context validation (Food.com)

We matched 579 Food.com dishes via two-pass matching (exact + token-set) and extracted 104,019 reviews ( $\geq 20$  words, max 1,000/dish). The same BERT model and full BT pipeline are applied independently: stratified sampling (5 reviews at 10/30/50/70/90th BERT percentiles), anchor-bridging (20 anchors, 11,370 comparisons via DeepSeek v3.2), and BT fitting. Domain-shift calibration via Food.com star ratings shows negligible impact ( $\Delta\rho < 0.001$ ). Correlations are Spearman  $\rho$  with bootstrap 95% CIs (2,000 iterations).

## 5.10 Food-chemistry hedonic score

For each dish with a recipe (2,387 of 2,563), we compute five palatability dimensions from ingredient composition: (1) fat content (g/serving from USDA FoodData Central), (2) umami intensity (glutamate/nucleotide-based scores for 106 ingredients, weighted by mass fraction), (3) sodium content (mg/serving), (4) sugar content (g/serving), and (5) spice diversity (count of distinct spice/herb ingredients). Each dimension is standardised;  $H_{\text{chem}}$  is the equal-weight composite, rescaled to 1–10.

### 5.11 Within-restaurant analysis

To test the robustness of the decoupling, we construct two types of within-restaurant pairs from the 76,927 scored mentions across 19,905 restaurants:

**Cross-review pairs.** Different reviewers evaluating different dishes at the same restaurant. For each restaurant with  $\geq 2$  reviews mentioning different dishes, we enumerate cross-review pairs (capped at 50 per restaurant). This yields 130,384 pairs across 9,886 restaurants with zero BERT context leakage.

**Within-review pairs.** Same reviewer, same meal, different dishes ( $n = 2,508$ ). These are subject to BERT context leakage (shared text context), diagnosed by the correlation  $r(\text{overlap}, |\Delta H|) = -0.44$ , where overlap is the Jaccard similarity of context tokens.

**Restaurant fixed effects.**  $H_{\text{within}} = H_{ij} - \bar{H}_{.j}$ , where  $\bar{H}_{.j}$  is the restaurant mean. Dish-level  $H_{\text{within}}$  is aggregated over mentions.

### 5.12 LLM recipe validation

For the 158 dishes with human-authored recipes, we independently regenerate recipes via DeepSeek v3.2 using the same ingredient vocabulary (101 items). Validation metrics: ingredient Jaccard overlap, gram-weight MAE, cooking method agreement, and  $E$  Spearman  $\rho$ . Systematic bias is assessed by protein-source category (meat, seafood, plant, other).

### 5.13 Robustness analyses

**Proxy-bias control.** Mention-level  $H$  is regressed on restaurant star rating, price range, log review count, and text length; controlled  $H$  uses residuals.

**Survivorship bounds.**  $K \in \{100, \dots, 2,563\}$  ghost dishes with  $H \sim \text{Unif}[\bar{H} - \Delta, \bar{H}]$  for  $\Delta \in \{0, \dots, 3.0\}$ ;  $E$  sampled from observed distribution; 200 repetitions per scenario.

**Multi-dimensional hedonic.** Satiety extracted via keyword matching; composite  $H_{\text{comp}} = (1 - w)H_{\text{taste}} + wH_{\text{satiety}}$  for  $w \in [0, 1]$ .

**Meal-level.** Dishes classified into four roles by calorie content. Combinations: main + side, 500–1,500 kcal,  $\geq 20$  g protein. Calorie-equivalent substitution:  $\pm 25\%$  kcal,  $\pm 50\%$  protein,  $> 20\%$   $E$  reduction.

### 5.14 Statistical methods

Bootstrap CIs (5,000 iterations) unless noted. OLS with HC3 robust SEs. Monte Carlo: 10,000 simulations. All  $P$  values two-sided.

## References

- [1] Poore, J. & Nemecek, T. Reducing food’s environmental impacts through producers and consumers. *Science* **360**, 987–992 (2018).
- [2] Crippa, M. et al. Food systems are responsible for a third of global anthropogenic GHG emissions. *Nat. Food* **2**, 198–209 (2021).
- [3] Mekonnen, M. M. & Hoekstra, A. Y. Four billion people facing severe water scarcity. *Sci. Adv.* **2**, e1500323 (2016).
- [4] Benton, T. G. et al. Food system impacts on biodiversity loss. *Chatham House Report* (2021).
- [5] Springmann, M. et al. Options for keeping the food system within environmental limits. *Nature* **562**, 519–525 (2018).
- [6] Willett, W. et al. Food in the Anthropocene: the EAT–Lancet Commission on healthy diets from sustainable food systems. *Lancet* **393**, 447–492 (2019).
- [7] Hartmann, C. & Siegrist, M. Consumer perception and behaviour regarding sustainable protein consumption: a systematic review. *Trends Food Sci. Technol.* **61**, 11–25 (2017).
- [8] Onwezen, M. C. et al. A systematic review on consumer acceptance of alternative proteins. *Appetite* **159**, 105058 (2021).
- [9] Clark, M. A. et al. Global food system emissions could preclude achieving the 1.5° and 2°C climate change targets. *Science* **370**, 705–708 (2020).
- [10] Julia, C. et al. Front-of-pack environmental labelling: overview and perspectives. *Lancet Planet. Health* **6**, e696–e703 (2022).
- [11] Scarborough, P. et al. Vegans, vegetarians, fish-eaters and meat-eaters in the UK show discrepant environmental impacts. *Nat. Food* **4**, 565–574 (2023).
- [12] Lawless, H. T. & Heymann, H. *Sensory Evaluation of Food* (Springer, 2010).
- [13] Stubbs, R. J. et al. Measuring the difference between actual and reported food intakes. *Br. J. Nutr.* **85**, 415–430 (2001).
- [14] Flynn, A. N. et al. Dish swap across a weekly menu can deliver health and sustainability gains. *Nat. Food* **6**, 843–847 (2025).
- [15] Drewnowski, A. Defining nutrient density: development and validation of the Nutrient Rich Foods Index. *J. Am. Coll. Nutr.* **28**, 421S–426S (2009).
- [16] Stylianou, K. S. et al. Small targeted dietary changes can yield substantial gains for human health and the environment. *Nat. Food* **2**, 616–627 (2021).
- [17] Hu, N. et al. Overcoming the J-shaped distribution of product reviews. *Commun. ACM* **52**, 144–147 (2009).
- [18] Maystre, L. & Grossglauser, M. Fast and accurate inference of Plackett–Luce models. In *NeurIPS* **28** (2015).

- [19] Winata, G. I. et al. WorldCuisines: A massive-scale benchmark for multilingual and multicultural visual question answering on food images. In *NAACL* (2025).
- [20] Lehner, M. et al. Nudging—a promising tool for sustainable consumption behaviour? *J. Cleaner Prod.* **134**, 166–177 (2016).
- [21] Thaler, R. H. & Sunstein, C. R. *Nudge* (Penguin, 2009).

## Extended Data

**Table 2: Extended Data Table 1 | Robustness summary across fifteen validation tests.**

Test	Result	Notes
BT split-half reliability	$\rho = 0.832$	$\pm 0.003$ (20 splits)
Pairwise vs BERT $H$ rank $\rho$	0.272	$p < 10^{-6}$ (334 dishes)
$H$ contribution (pairwise, core 334)	20.0%	vs 0.8% (BERT)
$H$ contribution (pairwise, full 2,563)	3.1%	conservative bound
Yelp vs Google Local BERT $\rho$	0.698	$p < 10^{-23}$
Yelp vs TripAdvisor BERT $\rho$	0.588	$p < 10^{-15}$
Food.com cross-context BT $\rho$	0.338	$p < 10^{-16}$ , 579 dishes
Proxy-bias control rank $\rho$	0.999	$H$ contrib. drops to 0.1%
Survivorship ( $K = 2563$ , $\Delta = 2$ )	$E$ : 92.6%	$H$ : 7.4%
State-level DEI stability	$\rho > 0.993$	14 US states
Temporal stability	$\rho = 0.812$	Split-period (2005–17 vs 18–22)
FNDDS nutritional validation	$\rho = 0.949$	DEI-N rank stability
Within-restaurant FE: $\rho(H, H_{\text{within}})$	0.681	$p < 10^{-22}$ (158 dishes)
Cross-review pairs: $r(\Delta H, \Delta E)$	−0.006	130,384 pairs, $p = 0.04$
LLM recipe validation: $\rho(E_{\text{human}}, E_{\text{LLM}})$	0.822	$p < 10^{-40}$ (158 dishes)

**Table 3: Extended Data Table 2 | Within-category variance decomposition (2,563 dishes).**

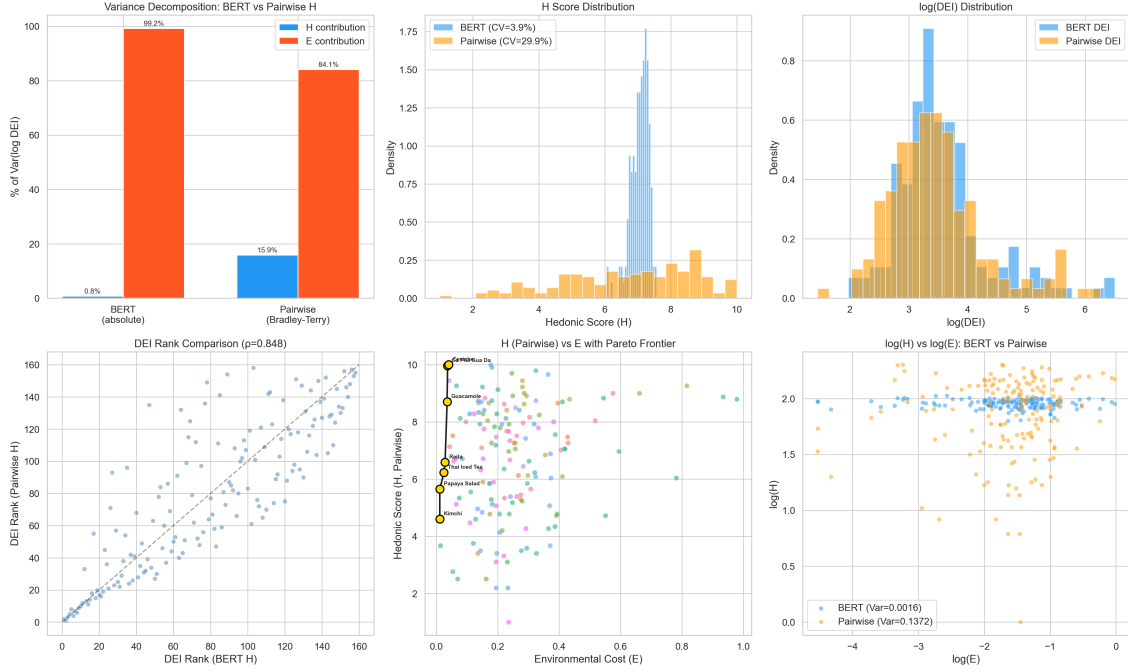
Category	$n$	$H$ CV (%)	$E$ CV (%)	$H$ % of Var	Mean log(DEI)
Poultry Main	156	18.3	31.1	37.2	3.13
Starch/Carb	714	10.4	30.1	9.6	3.37
Red Meat Main	308	9.1	26.8	9.5	2.54
Mixed/Other	88	18.8	45.0	8.4	3.54
Beverage	6	20.2	78.6	7.6	4.58
Soup/Stew	58	11.3	45.8	7.5	3.35
Plant Protein	119	9.2	37.0	6.6	3.49
Pork Main	193	10.7	32.0	6.3	3.09
Dairy Main	72	9.4	26.3	3.5	3.19
Seafood Main	132	10.1	37.4	3.2	3.39
Dessert	566	7.8	33.7	1.2	3.26
Salad/Cold	65	12.2	59.1	−0.4	3.91
Egg Dish	86	7.7	43.9	−1.7	3.42

## Data availability

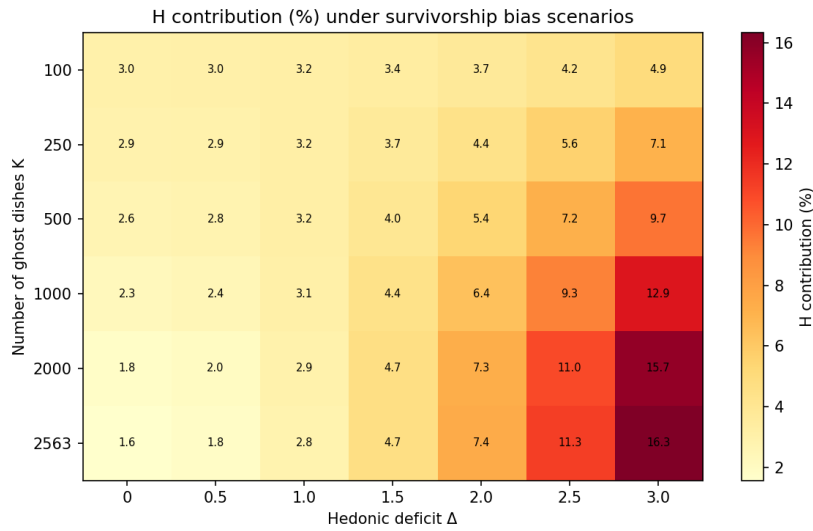
The Yelp Open Dataset is available at <https://www.yelp.com/dataset>. The WorldCuisines database is available at <https://github.com/worldcuisines/worldcuisines>. Environmental impact factors are from Poore & Nemecek (2018), supplemented with Agribalyse 3.1 (<https://agribalyse.ademe.fr/>) and Water Footprint Network (<https://www.waterfootprint.org/>). Nutritional data are from USDA FoodData Central SR28 (<https://fdc.nal.usda.gov/>). All derived datasets—including dish-level DEI scores, pairwise comparison records, LLM-generated recipes, and environmental cost calculations—are available at <https://github.com/fangyuliu/dei-replication>.

**Table 4: Extended Data Table 3 | OLS regression of  $\log(\text{DEI})$  on environmental sub-components (HC3 robust SEs).**

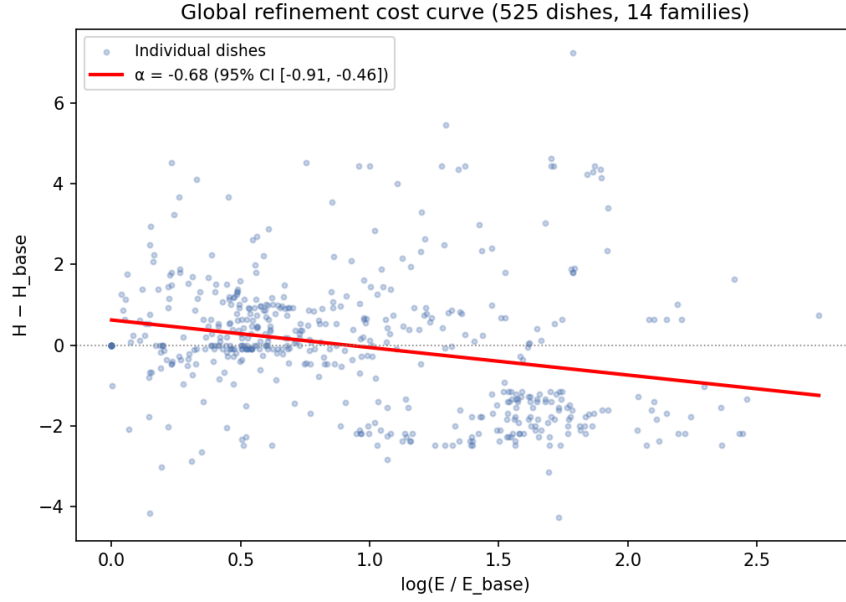
Predictor	Coeff.	SE	$t$	$p$
Intercept	+3.771	0.102	36.87	< 0.001
$\log(E_{\text{carbon}})$	-0.245	0.013	-19.48	< 0.001
$\log(E_{\text{water}})$	-0.100	0.016	-6.25	< 0.001
$\log(E_{\text{energy}})$	-0.565	0.008	-68.88	< 0.001
$R^2 = 0.882, n = 2,520$				



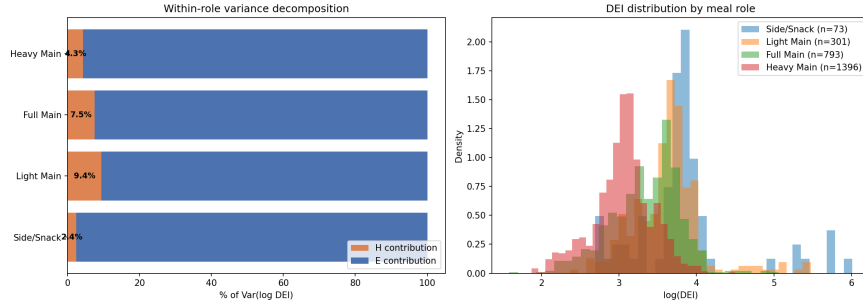
**Figure 4: Extended Data Fig. 1 | BERT versus pairwise hedonic scoring.** The transition from absolute ( $\text{CV} = 3.9\%$ ) to pairwise scoring ( $\text{CV} = 33.6\%$ ) increases  $H$ 's variance contribution from 0.8% to 20.0% while preserving  $E$ -dominance (80.0%).



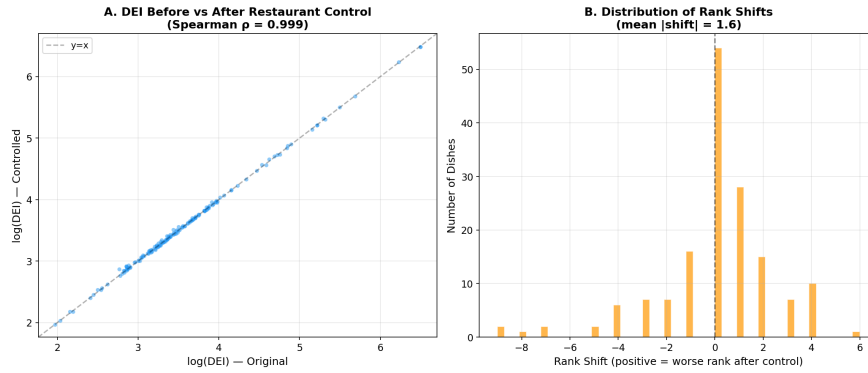
**Figure 5: Extended Data Fig. 2 | Survivorship bias bounds.**  $H$ 's contribution to  $\text{Var}(\log(\text{DEI}))$  as a function of ghost dishes ( $K$ ) and hedonic deficit ( $\Delta$ ). Even doubling the sample ( $K = 2,563, \Delta = 2$ ) leaves  $E$  dominant at 92.6%.



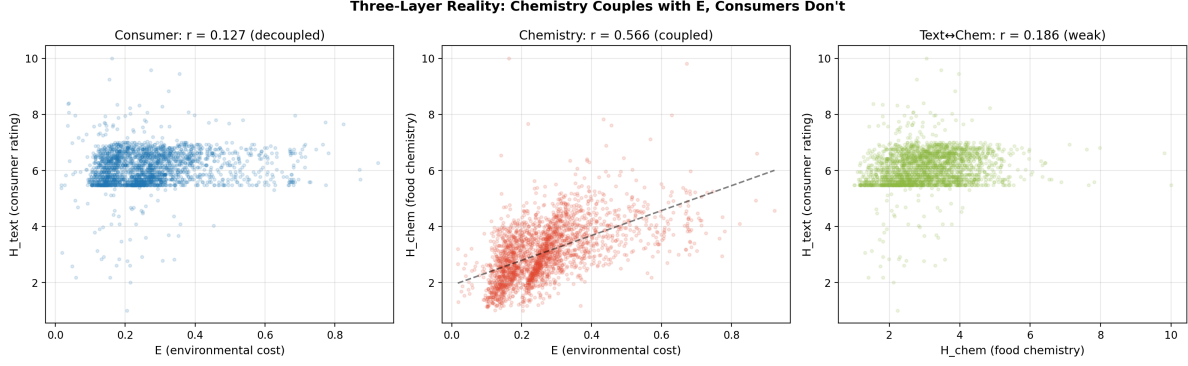
**Figure 6: Extended Data Fig. 3 | Global refinement cost curve.** Pooled relationship across 14 dish families (525 dishes). Slope  $\alpha = -0.68$  (95% CI  $[-0.91, -0.46]$ ): increasing  $E$  within families *decreases*  $H$ .



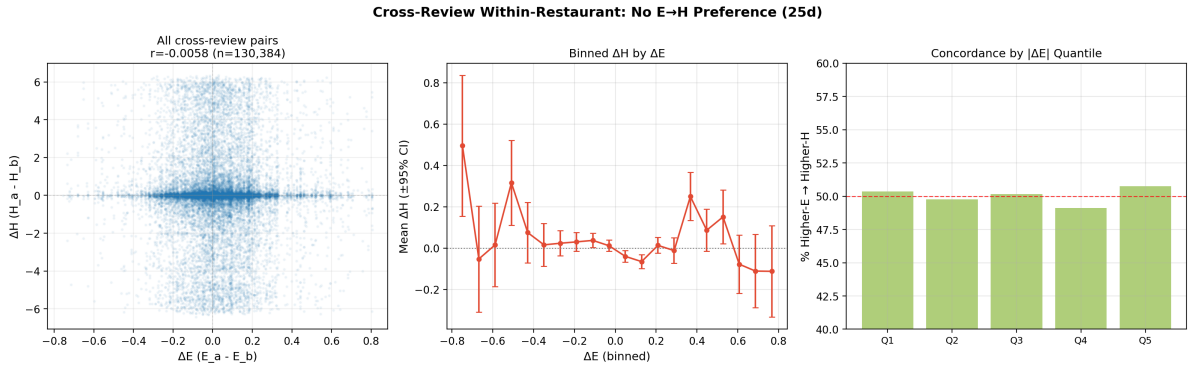
**Figure 7: Extended Data Fig. 4 | Meal-level DEI and nutritional adjustment.** Within-role variance decomposition and DEI distributions by meal role, confirming  $E$ -dominance within each functional category. Rankings under nutrition-adjusted DEI-N ( $\alpha = 0.5$ ) preserve core structure ( $\rho = 0.802$ ); nutritionally dense dishes gain ranks.



**Figure 8: Extended Data Fig. 5 | Proxy-bias control and nutritional validation.** Original vs controlled DEI ( $\rho = 0.999$ ). Restaurant-level confounding explains only 7.7% of mention-level  $H$  variance. USDA FNDDS cross-validation: recipe-based protein vs. USDA ( $r = 0.814$ ,  $n = 113$ ); DEI-N rank stability  $\rho = 0.949$ .



**Figure 9: Extended Data Fig. 6 | Three-layer reality.** (a) Consumer  $H_{\text{text}}$  vs.  $E$ :  $r = 0.13$  (decoupled). (b) Food-chemistry  $H_{\text{chem}}$  vs.  $E$ :  $r = 0.57$  (coupled). (c)  $H_{\text{text}}$  vs.  $H_{\text{chem}}$ :  $r = 0.19$  (weak). Consumer evaluations capture primarily non-chemical hedonic dimensions that are orthogonal to environmental cost.



**Figure 10: Extended Data Fig. 7 | Cross-review within-restaurant evidence.** (a)  $\Delta H$  vs.  $\Delta E$  across 130,384 cross-review pairs ( $r = -0.006$ ). (b) Binned mean  $\Delta H$  by  $\Delta E$  with 95% CIs: no monotonic trend. (c) Concordance rate by  $|\Delta E|$  quantile: 50% throughout, consistent with the null. Different diners at the same restaurant show no preference for higher- $E$  dishes. Within-review pairs ( $n = 2,508$ ) show the same null ( $r = -0.001$ ), though subject to BERT context leakage ( $r(\text{overlap}, |\Delta H|) = -0.44$ ). Low-overlap within-review pairs ( $n = 627$ ,  $r = -0.010$ ) corroborate the cross-review result.



## **Code availability**

All analysis code for data processing, hedonic scoring, environmental cost computation, pairwise Bradley–Terry modelling, validation analyses, and figure generation is available at <https://github.com/fangyuliu/dei-replication>.

## **Acknowledgements**

Manuscript drafting, data analysis, and code development were assisted by Claude Opus 4.6 (Anthropic). The author takes full responsibility for the content of this work.

## **Author contributions**

F.L. conceived and designed the study, collected and analysed the data, and wrote the paper.

## **Competing interests**

The authors declare no competing interests.