

Mapping the language of fragrance with FragranceDive

Tyler M. Moore, Ph.D.^{1,2}

¹ Department of Psychiatry, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104 USA

² The Lifespan Brain Institute (LiBI) of the Children's Hospital of Philadelphia (CHOP) and Penn Medicine, Philadelphia, PA 19104

Corresponding author

Tyler M. Moore, Ph.D.

3700 Hamilton Walk - Office B502, Philadelphia, PA 19104

tymoore@pennmedicine.upenn.edu

Article type: Original Article

[Supplementary materials](#) · Use the PDF bookmarks to navigate sections, tables, and figures.

Abstract

Commercial fragrance descriptions combine scent, performance, and market information. FragranceDive organizes 35,449 fragrances using language model annotations on 90 question columns representing 60 distinct wordings. An exploratory analysis retained 14 correlated factors for interpretability. Web-assisted correction was evaluated in 200 randomly sampled and 50 recognized fragrances. In the random subset, median matched polychoric correlation between repeated questions increased from .790 to .944, and absolute-agreement intraclass correlation increased from .580 to .869. Directional agreement with a frozen source-based audit reference increased from 90.7% to 96.9% on 432 assessable matched cells; weighted evidence coverage was 36.8%. Improvement persisted after excluding rows exposed during an initial audit preview. Newly filled cells remained unaudited. Fixed reference scoring and archived audit records support comparison across annotation versions. The resource provides interpretable maps of documented product characteristics for stimulus selection and descriptor development, with perceptual correspondence and consumer relevance to be established through human validation studies.

Practical Applications

FragranceDive helps sensory researchers assemble candidate stimulus sets, identify products with contrasting documented profiles, and develop descriptor lists. Separate scent, performance, and commercial dimensions support sampling tailored to the research question, such as comparing similarly described fragrances at different price levels. Fixed reference scoring makes annotation revisions visible on stable axes, while the audit records help direct further verification. Researchers can use these profiles to plan studies, verify the exact release and concentration before purchasing samples, and establish attribute meanings with assessors. Sensory panels and consumer studies can then test whether the selected contrasts correspond to perceived differences, liking, or choice.

Keywords: fragrance; sensory semantics; language models; factor analysis; annotation quality; sensory validation

Introduction

Commercial fragrances are described through overlapping vocabularies of odor quality, performance, and product identity. Floral, woody, and fruity convey scent character, whereas longevity, projection, price, and intended audience describe additional properties relevant to comparison and selection. Sensory and consumer science increasingly considers extrinsic product information, context, emotions, and behavior alongside liking (Meiselman, Jaeger, Carr, & Churchill, 2022). A useful fragrance representation should accommodate these different properties while specifying what its coordinates measure. Establishing that representation across a large catalog requires scalable annotation and evidence about the resulting data.

Statistical analyses have long identified structure in perfumery language. Chastrette, Elmouaffek, and Sauvegrain (1988) examined 74 notes in descriptions of 2,467 odoriferous substances, finding associations that reflected both chemical similarities and semantics. Perception-based odor arrangements also depend on stimuli, assessors, data-collection procedures, and analytic methods (Kaeppeler & Mueller, 2013). Dimensions therefore require interpretation within the framework that produced them.

Analyses of perfumery materials recovered broad contrasts related to light versus heavy odors and feminine versus masculine cosmetic scents (Zarzo, 2015; Zarzo & Stanton, 2009). In finished products, structured descriptions of 140 feminine fragrances and a separate analysis of 176 perfumes supported sensory wheels and dimensions related to gender associations and daytime versus nighttime wear (Zarzo, 2020a, 2020b). The first two components in the latter analysis explained less than 30% of descriptor variability, illustrating the limits of two-dimensional maps. Different measurement frameworks recover different organizations: the leading component of the Dravnieks odor-profile database was interpreted as pleasantness (Khan et al., 2007), whereas non-negative matrix factorization of the same database yielded 10 sparse dimensions (Castro, Ramanathan, & Chennubhotla, 2013). These findings support multidimensional descriptions without establishing universal fragrance categories.

Direct sensory work provides a benchmark for description-based maps. Trained assessors and consumers sorting 15 commercial feminine perfumes produced broadly similar configurations but differed in some relationships and descriptions (Veramendi, Herencia, & Ares, 2013). Lindqvist (2012) found graded, overlapping gender associations across 12 perfumes; femininity predicted from online olfactory descriptions correlated with panel ratings ($r = .87$). Descriptions can thus track perceptual judgments, but their correspondence requires testing for the products and attributes being represented.

Language itself contains olfactory structure. Hörberg, Larsson, and Olofsson (2022) organized 243 English odor descriptors into Offensive, Malodorous, Fragrant, and Edible clusters, with partial correspondence to perceptual data. Kurfalı, Herman, Pierzchajło, Olofsson, and Hörberg (2025) compared language models against actual odor similarity, imagined odor similarity from labels, and odor-to-label ratings. Performance depended on the benchmark: GPT-4o performed best on the language-mediated tasks, while static embeddings achieved the highest numerical correlation with actual odor similarity, without significantly outperforming GPT-4o. Recovering olfactory semantics and reproducing sensory relationships are consequently related but distinguishable achievements.

Product reviews offer a further source of descriptive data. Hamilton and Lahne (2020) derived a whisky flavor wheel from online reviews, recovering several established categories but no off-flavor terms, which were absent from the corpus. Espinosa-Ramos et al. (2025) found comparable product discrimination from online reviews and a rate-all-that-apply test of five coffees with 105 consumers, although individual attributes differed. Fragrance discourse is also culturally situated. Zhou, Zhou, Zhu, and Zou (2026) analyzed 29,372 Chinese reviews of 36 perfumes and identified entity-based, abstract, and evaluative descriptions, with source-based vocabulary predominating and more than 70% of reviews classified as positive. These studies support using product text while showing that coverage depends on what people discuss and how they describe it.

This descriptive approach complements molecular odor prediction. Models have predicted perceptual attributes from chemical features (Keller et al., 2017), organized odorants in a principal odor map (Lee et al., 2023), and extended molecular representations to mixture similarity (Tom et al., 2025). Named commercial products pose a different problem: their formulations are generally unavailable, and relevant characteristics include performance and commercial positioning as well as odor. Continuous product dimensions can express overlapping characteristics that would be compressed by assigning each fragrance to one family.

For finished fragrances, Teixeira, Barrault, Rodríguez, Carvalho, and Rodrigues (2014) developed Perfumery Radar 2.0, combining chemical composition, vapor-liquid equilibrium, odor thresholds, and ingredient classifications to predict profiles across established olfactory families and finer descriptors. Its eight basic families were supplemented by outer and inner descriptor layers. The authors compared predictions with new expert classifications of five formulated mixtures and with published classifications of 36 chemically analyzed commercial perfumes. This provides a composition-based route to describing finished products, with different evidential requirements from a language-derived catalog. Where formulations or chemical analyses are unavailable, structured annotation can summarize documented characteristics across many releases, but correspondence with composition-based predictions or sensory judgments must be evaluated separately. The shared goal is an interpretable representation; the inputs, coordinate meanings, and validation criteria differ.

Large language models (LLMs) can apply common questions across products and consult web sources during annotation. Successful text-annotation applications coexist with substantial differences across tasks, making explicit coding rules and task-specific validation essential (Abdurahman, Salkhordeh Ziabari, Moore, Bartels, & Dehghani, 2025; Gilardi, Alizadeh, & Kubli, 2023; Törnberg, 2024; Ziems et al., 2024). Structured entity annotation extends this approach by synthesizing answers across sources; previous applications produced interpretable quantitative matrices for organisms, people, and compounds (Moore, 2025, 2026a, 2026b). Each response remains an estimate whose reliability and accuracy require evaluation in its own domain.

Uneven documentation is a particular obstacle. Factual question answering is less successful for knowledge that appears rarely in training data, although retrieval can help (Kandpal, Deng, Roberts, Wallace, & Raffel, 2023). Widely discussed fragrances may therefore be easier to annotate than obscure releases. Search does not resolve every omission or distinguish every similarly named product, motivating explicit Unknown responses, release-specific identification, and evaluation beyond familiar examples.

Data quality also has several levels. Repeated questions assess consistency; external evidence assesses source agreement; sensory measurement tests perceptual correspondence. A model can be internally consistent while inaccurate, and imperfect annotations can bias downstream analyses even when aggregate accuracy appears high (Egami, Hinck, Stewart, & Wei, 2023; Ziems et al., 2024). Validation must therefore report what was assessable and distinguish source support from sensory evidence.

For sensory methodology, these distinctions determine the question a representation can answer. A family label summarizes a classification, a descriptive profile records specified attributes, and a similarity judgment concerns relationships among products. Agreement at one level does not establish agreement at the others. We therefore treat the annotation matrix as an explicit set of product-level judgments whose structure and source support can be examined separately. This makes both the question wording and the limits of the available evidence part of the method, rather than treating a plausible-looking map as sufficient validation.

The present study introduces FragranceDive, a database of 35,449 commercial fragrances annotated on 90 question columns representing 60 distinct wordings. Exploratory factor modeling summarizes scent characteristics, performance, price, and market positioning on a common reference metric. We evaluate targeted, web-assisted correction of missing or statistically unexpected responses in a random sample of 200 fragrances and 50 widely recognized fragrances, alongside a separate source-based audit of sampled responses from the random subset. The study asks whether annotation yields interpretable dimensions, whether targeted review improves consistency and source agreement, and how evidence coverage constrains these conclusions.

Materials and Methods

The workflow combined web-assisted LLM annotation, exploratory factor modeling, targeted correction, and evaluation of repeated-item consistency and source agreement. Supplementary Methods documents the complete collection, parsing, correction, scoring, and audit procedures; Tables S1–S14 and Figures S1–S6 provide extended results. The unit of analysis was the named fragrance, and the measured responses were model-generated annotations of documented product characteristics. No fragrances were presented to human assessors in this study. The sensory implications of the resulting representation are therefore proposed applications to be tested, rather than observed panel outcomes.

Fragrance Corpus and Item Set

The corpus comprised 35,449 commercial fragrances, stored as Name (House) to distinguish releases with shared names. The 90 columns represented 60 distinct questions spanning notes, olfactive character, gender coding, price, projection, longevity, and market familiarity (Table S1). Twenty-eight questions occurred twice, one three times, and 31 once; repetitions were separated and not identified to the model.

All items used 0 = Definitely No, 1 = Probably No, 2 = Unknown or not applicable, 3 = Probably Yes, and 4 = Definitely Yes. Code 2 was a collection-stage marker for Unknown or not applicable. These entries were converted to blanks (missing values) before statistical analysis; no separate indicators or analyses of code-2 responses were used. The probably/definitely distinction expressed certainty that an attribute criterion was met, not its measured sensory intensity.

Reuse license. The FragranceDive data and questionnaire are dedicated to the public domain under CC0 1.0 Universal. No separate permission is required to reuse, modify, redistribute, or use them commercially. The accompanying scripts are released under the MIT License, which permits reuse, modification, redistribution, and commercial use subject to retention of the copyright and license notice.

Annotation

Primary annotation used Grok 4.3 through the xAI Batch API at medium reasoning effort, with web search enabled (xAI, 2026a). Requests contained five fragrances, the full item list, response definitions, and CSV formatting requirements. The model was instructed to search thoroughly and use Unknown when evidence remained insufficient. An earlier Claude-based collection was discontinued; the analysis pipeline reads the Grok batch output without a step merging the earlier ratings.

Outputs were aligned to question positions and requested fragrances, with ambiguous matches rejected and parsing problems logged for retry. The collection matrix was initialized to code 2, so Unknown responses and unreturned values could share a code. Provenance logs distinguished clean parsing, column realignment, partial returns, unmatched rows, and other failures, supporting retries and review. These are collection and data-processing records; all remaining original code-2 entries were converted to blanks before statistical analysis. Supplementary Methods documents the parsing rules and the general-purpose correction code.

Dimension Reduction and Factor Scoring

An exploratory item-factor model used `irt.fa` in `psych` to derive graded-response-type parameters from polychoric correlations (Revelle, 2025; Samejima, 1969). Both price-below-\$30-per-ounce columns were reverse-keyed. Original code-2 responses had been converted to missing values in the reference and Famous50 data. The substantive responses 0, 1, 3, and 4 were recoded consecutively as 0, 1, 2, and 3; the new category 2 therefore represents original Probably Yes responses, not Unknown. Solutions with 8 through 15 factors were examined, and 14 factors were retained on interpretability grounds, as the 15-factor solution was less clearly interpretable. A subsequent check using Velicer's minimum average partial (MAP) criterion also suggested 14 factors (Velicer, 1976). Rotation was oblimin. The standardized root mean-square residual (SRMR) summarized residual fit and was not used to select the factor count. Repeated columns remained separate. The reported loadings are the oblimin-rotated pattern matrix, and $|\text{loading}| \geq .30$ identified salient indicators. The oblique factor-correlation matrix (Φ) is reported separately in Table S14.

The model describes covariance among annotated questions, including repeated administrations of selected wordings. The four substantive categories are ordered by strength and direction of endorsement, but the analysis does not turn those categories into physical concentrations or calibrated sensory intensities. Retaining duplicates preserves the questionnaire as administered while giving some contents additional representation; the implications of that weighting are considered in the Discussion.

The full corpus served as the reference set for all scoring. Factor scores (θ) were computed with `scoreIrt` using an absolute IRT discrimination cutoff of .30 ($\text{cut} = .30$), distinct from the .30 pattern-loading criterion used for interpretation. To place the 14 factors on a common and interpretable display metric, a single pair of scaling constants was estimated per factor from the reference set alone: the 0.1st and 99.9th percentiles of that factor's θ distribution. Scores

were clamped to that interval, winsorizing the most extreme 0.1% in each tail so that a small number of outlying estimates cannot compress the rest of the scale (Wilcox, 2021), and then mapped linearly onto a -4 to $+4$ range and rounded to two decimals. These display units are not standard deviations; zero is the midpoint of the percentile anchors rather than necessarily the reference mean. The scaling constants were written to a parameter file and held fixed thereafter, so that the display metric is a property of the reference set rather than of any particular analytic file. The model and reference constants were then held fixed when scoring the original and corrected Famous50 ratings; they were not re-estimated within that subset.

Each bivariate plot displays all 50 fragrances with 25 selected labels; points indicate locations when labels are displaced. Original and corrected panels use identical axes and scaling. The diagonal is a visual guide, not a test of equivalence between factors.

Data Quality Correction Procedure

Correction predicted each rating column from the others to prioritize statistically unexpected cells for web-assisted review. Evaluation used Random200, 200 fragrances sampled without replacement from the 35,449-row `fragrance_data_raw_forPub.csv` file using `set.seed(2026)` in R, and Famous50, 50 widely recognized releases selected with an LLM using sales, media recognition, and bottle/brand familiarity (OpenAI, 2026). Famous50 supplied recognizable illustrations rather than a representative sample.

Blanks, including recoded Unknown responses, and invalid entries were treated as missing. Within each column, missing cells received priority; remaining review slots went to substantive responses with high out-of-bag surprise and ordinal discrepancy under a 500-tree probability random forest (Breiman, 2001; Wright & Ziegler, 2017). The correction code uses a fixed numeric placeholder for missing predictors. This computational convention does not preserve the original Unknown responses or distinguish why a value is missing.

Random forests used out-of-bag class probabilities, so a row's prediction came from trees that excluded it during fitting. Review priority combined the negative log probability of the observed response with distances from the probability-weighted expected category and modal category. Low probability thus triggered investigation rather than automatic replacement. Up to five cells per column were selected in a sweep; when missing cells filled this budget, no forest-based selections were made. The Supplement specifies the complete score, forest settings, and handling of missing predictors, allowing the review rule to be implemented independently of the adjudicating language model.

Selected cells were reviewed by Grok 4.6 with high reasoning effort and required web search (xAI, 2026b). Each request addressed one question for up to five fragrances. Prompts required release-specific identification, distinguished concentrations and flankers, preferred authoritative sources, and instructed the model to treat triage statistics as reasons for review rather than evidence for an answer. Unsupported cases could return Unknown, which was converted to a blank before analysis. The source hierarchy favored house materials, specialist coverage, reference works, and authorized retailers. User-contributed databases were acceptable for note pyramids and launch information but treated as weaker support for performance and reformulation claims. The current widely sold formulation was assumed unless a vintage was specified, and poorly documented releases were not to be inferred from related flankers. Conflicting sources were to produce qualified rather than definite responses.

Recommendations were applied only when valid, different from the current value, and accompanied by model-reported confidence $\geq .70$; this was an operational threshold, not a validated probability of correctness. Random200 used a configuration allowing 20 sweeps with up to five selections per column per sweep. Famous50 underwent 10 sweeps, during which every one of its 4,500 cells was reviewed. This was comprehensive review of Famous50, not selective review of only its outliers. The matrix was refreshed between columns. Each response supplied a recommended category, confidence, change flag, rationale, and sources; review did not necessarily produce a change. Previously reviewed cells were excluded from repeat selection unless their value had changed. The Random200 limit of 100 selections per column was a budget, not a count of distinct corrections. Supplementary Methods details the review rules, and the accompanying data archive includes the supplied Random200 API response records.

Reliability of Duplicated Items

All 31 within-question pairs across 29 repeated questions were evaluated in each subset and version; the triplicated question contributed three overlapping pairs. Only observed substantive responses were included; original Unknown responses had already been converted to missing values. We calculated exact and No/Yes polarity agreement, polychoric correlation (Olsson, 1979), and the single-measurement, two-way, absolute-agreement intraclass correlation, ICC(A,1) (McGraw & Wong, 1996; Shrout & Fleiss, 1979). ICC used numeric distances on the original 0, 1, 3, 4 scale; polychoric correlation used category order.

Mean signed and absolute differences supplied additional information about response shifts. ICC estimates were checked against the corresponding analysis-of-variance calculation, and negative estimates were retained. Constant responses, sparse marginal categories, boundary estimates, and warnings were recorded rather than silently treated as perfect agreement. Reporting both association and absolute agreement is necessary because repeated ratings can retain their ordering while changing category or level.

Before-and-after summaries emphasize matched complete cases observed on both occurrences in both versions, and the same pairs with estimable coefficients. Available-case matrices describe each file's complete pairs and may differ in sample size. Sparse-category and estimation diagnostics were retained. Because duplicate columns were predictors during correction, post-correction agreement measures consistency of the corrected matrix, not independent annotation reliability.

External Accuracy Audit

The audit used GPT 6 Sol through ChatGPT (OpenAI, 2026). This separate LLM auditor sampled six originally nonblank cells per Random200 fragrance, yielding 1,200 targets. Each cell received inverse-probability weight $M/6$, where M was that fragrance's number of originally entered ratings. Target ratings were masked during reference assessment, although an initial preview had exposed 11 fragrance rows. Yes/No reference judgments distinguished direct source support from provisional inference; unresolved cases remained in coverage calculations. References were frozen before joining ratings.

The corrected file was compared with the same targets, references, evidence tiers, and weights. Paired analyses excluded cells that became missing. Conditional 95% intervals used 10,000 fragrance-cluster bootstrap replicates with shared resamples across versions. They capture sampling variation conditional on the frozen evidence, excluding reference error and

nonrandom assessability. Newly filled cells were outside the original sample; exact four-category accuracy and confidence calibration were not tested.

A masking sensitivity analysis excluded all targets from the 11 preview-exposed fragrance rows, retaining the original references and cell weights and resampling all 189 remaining fragrance rows, including those without assessable cells.

Correction preceded both audits. Corrector and auditor nevertheless used web-assisted judgments with potentially overlapping sources, so the benchmark assesses source agreement rather than independent sensory accuracy. Complete audit methods, chronology, evidence limitations, and sensitivity analyses appear in the Supplement.

Analyses used R 4.5.1 (R Core Team, 2025) and psych, polycor, irr, and ranger; audit sampling and bootstrap calculations used Python 3.13.5 and NumPy 2.3.5. Package versions, seeds, scripts, and machine-readable outputs are documented in Supplementary Methods and the accompanying archive. Numerical reproduction from saved ratings and regeneration of those ratings are distinct reproducibility targets: new annotation runs may differ as models or web evidence change.

Results

Factor Structure

The 14-factor solution retained for interpretability had a standardized root mean-square residual (SRMR) of .04 and organized the 90 rating columns into interpretable dimensions of scent character, performance, price, and market positioning (Table 1). Eighty-one columns (90.0%) had at least one absolute loading of .30 or greater, including 16 columns with salient loadings on more than one factor. Nine columns had no loading reaching .30. Because repeated questions were modeled separately, these counts describe rating columns rather than distinct question contents. The complete item key and all 1,260 pattern loadings are presented in Tables S1 and S2.

Several factors were anchored by strong, relatively specific indicators. Resinous/Smoky was marked by resinous and incense notes (loadings = .894 and .797), Aquatic/Ozonic by ozonic and aquatic notes (.895 and .859), and Minty by mint prominence (.817). Gourmand/Sweet was marked by gourmand notes (.820 and .767 across repeated occurrences), caramel (.688), and vanilla (.659); Fruity was marked by fruity notes (.919 and .847), peach (.581), and citrus (.428). The Masculine factor combined positive loadings for leather prominence (.554 and .515) and marketing to men (.552, .532, and .503) with negative loadings for floral notes (−.497) and indolic notes (−.421). Factor labels summarize the loading patterns and do not imply that the dimensions are statistically independent.

Interfactor correlations ranged from −.29 to .35. The largest linked Resinous/Smoky with Animalic (.35), Intense/Powerful (.32), and Expensive (.31; Table S14).

Table 1*Representative Indicators of the 14 Fragrance Factors*

Factor	n†	Representative indicators (loadings)
Resinous/Smoky	8	Resinous notes (.894); Incense notes (.797)
Animalic	6	Animalic notes (.706); Fecal notes (.647)
Expensive	4	Price < \$30/oz (reversed) (.937); Price > \$100/oz (.846)
Intense/Powerful	9	Sillage/projection (.645); Longevity ≥ 8 hours (.617)
Earthy	6	Mossy notes (.768); Chypre notes (.630)
Gourmand/Sweet	7	Gourmand notes (.820); Caramel (.688)
Masculine	11	Leather (.554); Marketed mostly to men (.552)
Aquatic/Ozonic	4	Ozonic notes (.895); Aquatic notes (.859)
Spicy/Woody	6	Cardamom (.806); Cumin (.568)
Carnation/Aldehydic	4	Carnation (.671); Aldehydic notes (.600)
Minty	4	Mint (.817); Camphoraceous notes (.474)
Fruity	4	Fruity notes (.919); Peach (.581)
Vintage/Waxy	4	Formulated before 1930 (.881); Waxy notes (.455)
US Mainstream	4	U.S. department-store availability (.851); Described as strange (–.393)

Note. Indicators are distinct question contents selected to characterize each factor; when repeated, the largest absolute loading is shown. The full signed pattern matrix appears in Table S2; factor correlations appear in Table S14. The price-below-\$30-per-ounce indicators were reverse-keyed before modeling. † Number of columns whose largest absolute loading is on that factor and is at least .30; repeated occurrences count separately. Counts sum to 81.

Factor Score Illustration

Figure 1 plots Gourmand/Sweet against Fruity for Famous50. Angel (Mugler) occupies the corrected upper-right region, whereas Good Girl (Carolina Herrera) is gourmand with a lower Fruity score. Sauvage Eau de Toilette (Dior) moves from high to low Fruity after correction, and Chanel No 5 Eau de Parfum (Chanel) moves upward. These changes illustrate annotation effects on coordinates; their sensory accuracy was not independently tested. Figures S1–S2 provide additional factor pairs.

Repeated Item Agreement

Random200 contained 31 repeated-question pairs representing 29 distinct questions. Matched sample sizes ranged from 74 to 163 fragrances per pair (median = 140). Across pairs with coefficients estimable in both versions, the median polychoric correlation increased from .790 to .944 and the median ICC(A,1) from .580 to .869 (30 pairs for each coefficient). Each coefficient increased for 28 of these 30 pairs. Across all 31 pairs, median exact agreement increased from 80.8% to 85.4%, and median polarity agreement increased from 88.7% to 97.9%. Exact agreement improved for 23 pairs, declined for seven, and was unchanged for one (Tables S4–S6).

The selected available-case matrices illustrate why correlation, absolute agreement, and exact agreement should be distinguished (Figure 2). For vetiver, ICC(A,1) increased from .560 to .895 while exact agreement decreased from 80.4% to 71.5%. For musk prominence, the polychoric correlation increased from .765 to .806 while ICC(A,1) decreased from .637 to .611. The price-above-\$100-per-ounce item improved on all three indices, with exact agreement increasing from 72.1% to 91.8% and ICC(A,1) from .392 to .930. Available-case sample sizes increased from 153 to 186, 107 to 180, and 122 to 183, respectively. Across all Random200 pairs, available-case median exact agreement decreased from 80.4% to 75.5%, emphasizing that this comparison also reflects changes in response availability.

Famous50 also showed higher matched-case coefficients and exact agreement after correction (Tables S4–S9). Full matrices appear in Figures S3–S6. Because repeated columns participated in correction, these increases describe post-correction consistency and do not independently validate annotation reliability.

Source Based Accuracy Audit

Of 1,200 sampled cells, 433 received reference judgments: 70 had direct source support and 363 provisional support; 767 remained unresolved. One assessable cell became missing, leaving 432 paired cells from 112 of the 200 fragrances. Weighted evidence coverage among still-answered audit targets was 36.8%.

Design-weighted directional agreement increased from 90.7% to 96.9%, a 6.2-percentage-point gain (conditional 95% fragrance-cluster bootstrap interval [4.2, 8.6]; Table 2). Thirty disagreements became agreements, 13 remained, and one new provisional disagreement appeared. Direct-tier agreement increased from 87.4% to 97.2% across 70 cells; sensitivity to affirmative references increased from 70.5% to 90.2%.

Excluding the 11 preview-exposed fragrance rows left 403 paired cells from 105 fragrances. Weighted agreement increased from 90.3% to 97.1%, a gain of 6.8 percentage points (conditional 95% fragrance-cluster bootstrap interval [4.6, 9.4]). The gain therefore persisted outside the preview-exposed rows; shared-source and reference-quality limitations remain.

Figure Legends

Figure 1. Gourmand/Sweet and Fruity factor scores before and after correction. A: Original ratings. B: Corrected ratings. All 50 Famous50 fragrances are plotted on the fixed reference metric; 25 labels are shown in each panel. Points identify score locations when labels are displaced. Red labels highlight selected examples and do not indicate statistical significance. The diagonal is an identity guide; the axes represent distinct factors.

Figure 2. Selected Random200 confusion matrices before and after correction. Original and corrected matrices use all complete cases in their respective files; sample sizes differ. Cell percentages use the total number of complete pairs in each matrix. Responses are 0 = Definitely No, 1 = Probably No, 3 = Probably Yes, and 4 = Definitely Yes. Polychoric correlations and single-measurement, two-way, absolute-agreement intraclass correlations [ICC(A,1)] summarize each comparison. The items are illustrative selections. Matched results for these and all other pairs appear in Table S6 and Figure S3.

Figure 1

Gourmand/Sweet and Fruity Factor Scores Before and After Correction

A. Original ratings

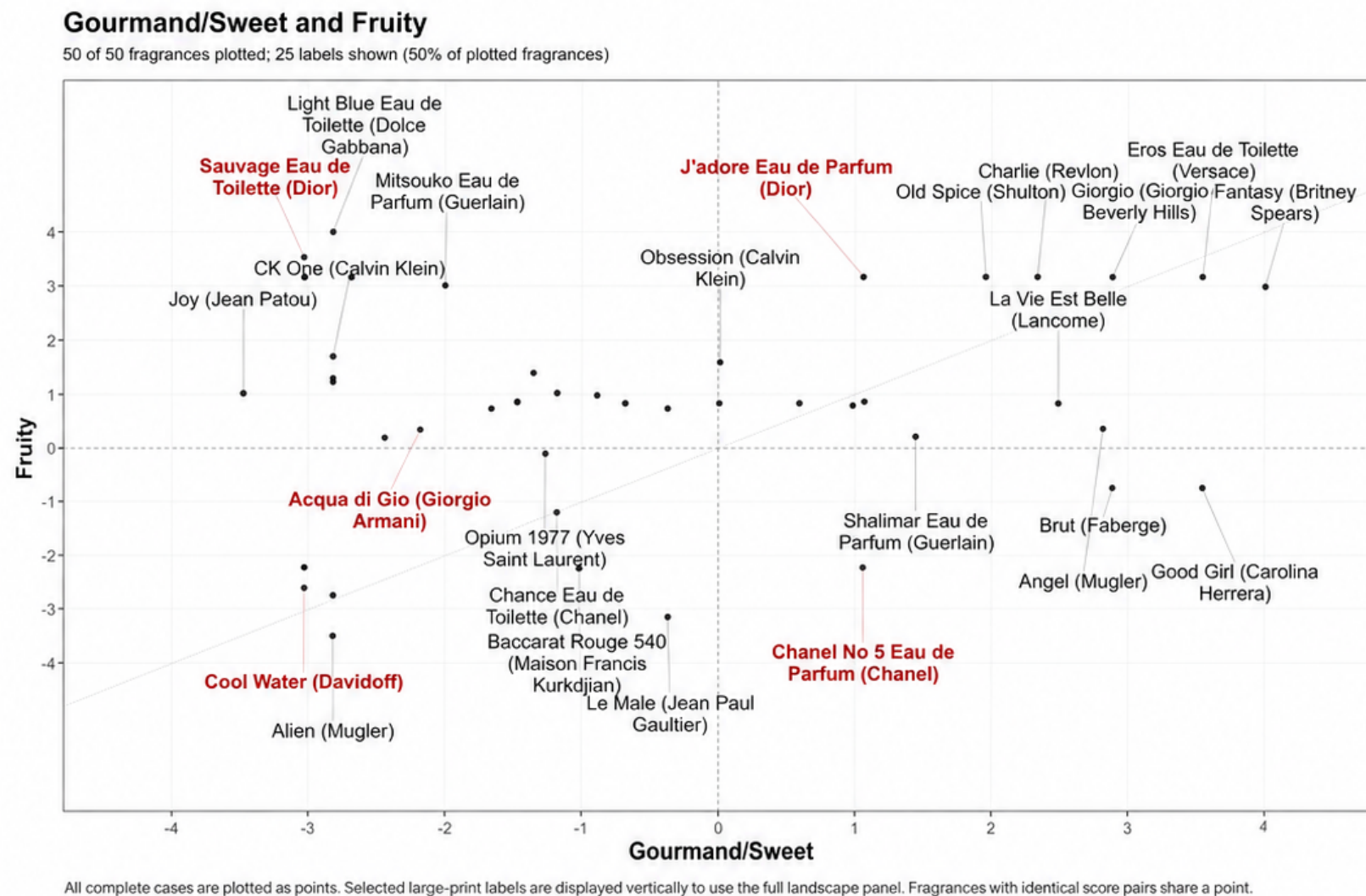


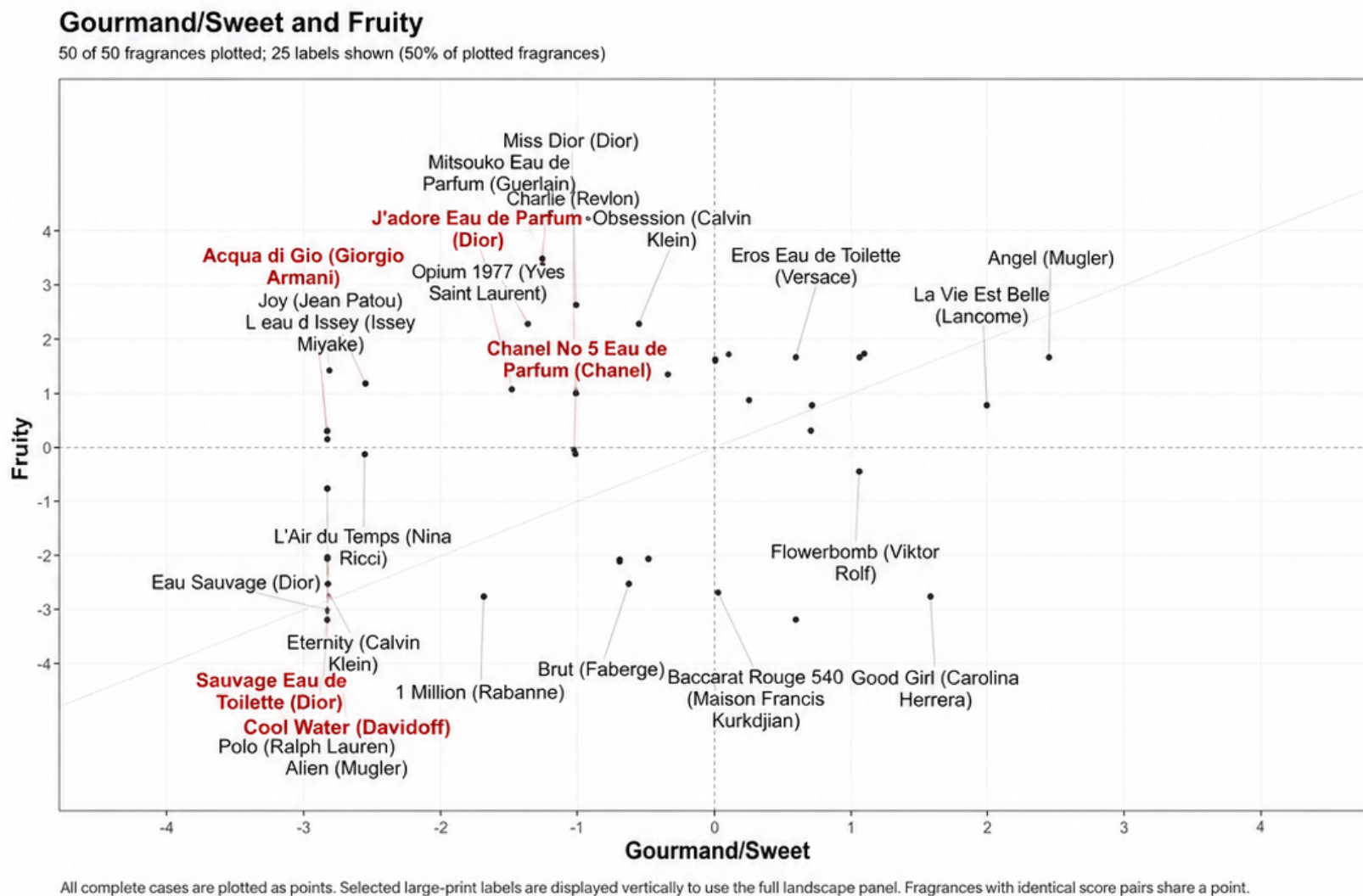
Figure 1 (continued)*Gourmand/Sweet and Fruity Factor Scores Before and After Correction***B. Corrected ratings**

Figure 2*Selected Random200 Confusion Matrices Before and After Correction***Agreement between repeated fragrance ratings**

Illustrative items; all available complete pairs within each dataset

A. Does this fragrance have vetiver notes?**Original | n = 153 | Exact = 80.4%**

Polychoric r = 0.772 ICC(A,1) = 0.560

Occurrence 1	0	114 74.5%	0 0.0%	1 0.7%	4 2.6%
	1	4 2.6%	0 0.0%	1 0.7%	0 0.0%
	3	4 2.6%	0 0.0%	0 0.0%	6 3.9%
	4	9 5.9%	0 0.0%	1 0.7%	9 5.9%
		0	1	3	4
		Occurrence 2			

Corrected | n = 186 | Exact = 71.5%

Polychoric r = 0.850 ICC(A,1) = 0.895

Occurrence 1	0	87 46.8%	23 12.4%	0 0.0%	0 0.0%
	1	27 14.5%	20 10.8%	1 0.5%	0 0.0%
	3	0 0.0%	1 0.5%	0 0.0%	0 0.0%
	4	1 0.5%	0 0.0%	0 0.0%	26 14.0%
		0	1	3	4
		Occurrence 2			

B. Is musk a leading note in this fragrance?**Original | n = 107 | Exact = 57.0%**

Polychoric r = 0.765 ICC(A,1) = 0.637

Occurrence 1	0	45 42.1%	5 4.7%	9 8.4%	1 0.9%
	1	7 6.5%	1 0.9%	9 8.4%	1 0.9%
	3	2 1.9%	0 0.0%	10 9.3%	6 5.6%
	4	0 0.0%	1 0.9%	5 4.7%	5 4.7%
		0	1	3	4
		Occurrence 2			

Corrected | n = 180 | Exact = 56.1%

Polychoric r = 0.806 ICC(A,1) = 0.611

Occurrence 1	0	59 32.8%	24 13.3%	6 3.3%	1 0.6%
	1	11 6.1%	25 13.9%	28 15.6%	0 0.0%
	3	0 0.0%	3 1.7%	12 6.7%	6 3.3%
	4	0 0.0%	0 0.0%	0 0.0%	5 2.8%
		0	1	3	4
		Occurrence 2			

C. Does this fragrance cost more than \$100 per ounce?**Original | n = 122 | Exact = 72.1%**

Polychoric r = 0.690 ICC(A,1) = 0.392

Occurrence 1	0	80 65.6%	7 5.7%	8 6.6%	12 9.8%
	1	2 1.6%	0 0.0%	0 0.0%	1 0.8%
	3	0 0.0%	0 0.0%	1 0.8%	2 1.6%
	4	2 1.6%	0 0.0%	0 0.0%	7 5.7%
		0	1	3	4
		Occurrence 2			

Corrected | n = 183 | Exact = 91.8%

Polychoric r = 0.983 ICC(A,1) = 0.930

Occurrence 1	0	158 86.3%	7 3.8%	0 0.0%	1 0.5%
	1	2 1.1%	0 0.0%	0 0.0%	0 0.0%
	3	0 0.0%	0 0.0%	0 0.0%	5 2.7%
	4	0 0.0%	0 0.0%	0 0.0%	10 5.5%
		0	1	3	4
		Occurrence 2			

Cells show count and percentage of complete pairs; outlined diagonal cells indicate exact agreement.

0 = definitely no; 1 = probably no; 3 = probably yes; 4 = definitely yes. ICC = absolute agreement, single measurement.

Note presence or prominence accounted for 379 of 432 assessments (87.7%) and 28 of 30 resolved disagreements. Performance/evolution had eight assessments, subtle descriptors four, and consumer strangeness none. Correction filled 2,194 previously blank Random200 cells that were outside the audit sampling frame. The estimate therefore does not establish accuracy for filled cells, unresolved cases, unassessed domains, or the full corpus (Tables S10–S13).

Discussion

FragranceDive organizes documented characteristics of 35,449 commercial fragrances into 14 interpretable dimensions. Targeted review improved repeated-response consistency and directional agreement with a fixed source-based reference. The fixed scoring framework and archived audit records make these improvements traceable across annotation versions. Together, these features support large-scale product comparison and selection for sensory research, with direct sensory and consumer validation providing the next step.

Interpreting the Fragrance Dimensions

The structure extends earlier analyses of perfumery descriptions to a larger catalog and broader product attributes (Zarzo, 2015, 2020a, 2020b; Zarzo & Stanton, 2009). Separating Gourmand/Sweet from Fruity permits a fragrance to express either or both; distinguishing Resinous/Smoky, Animalic, and Earthy preserves combinations compressed by broad family labels. These coordinates offer a flexible vocabulary for products whose descriptions overlap without being interchangeable.

Perfumery Radar 2.0 makes the distinction between a predefined sensory classification and an empirically estimated factor space concrete (Teixeira et al., 2014). Its profiles start from composition and calculate odor contributions within established families and descriptor layers. FragranceDive instead starts from question responses and estimates their covariance across products. The methods may converge on broad fruity, woody, or resinous contrasts while differing because of ingredient proportions, mixture behavior, descriptor conventions, or market information. Each representation is interpreted in relation to its own inputs and descriptive purpose.

Commercial dimensions add information relevant to product experience. Expensive reflects price-threshold responses, US Mainstream concerns the specified market, and Intense/Powerful combines reported performance characteristics. Their inclusion accords with Meiselman et al.'s (2022) account of sensory and consumer science expanding to extrinsic cues, expectations, context, and outcomes beyond liking. It can support comparisons among similarly described scents at different price or performance levels.

No broad hedonic factor emerged. This contrasts with pleasantness-related organization in perceptual profiles and odor language (Hörberg et al., 2022; Khan et al., 2007), but fits an instrument focused on notes, performance, price, and positioning rather than liking. Product text shows a comparable shift: across about 25 million words of English perfume reviews, Hörberg, Kurfalı, and Olofsson (2025) found concreteness rather than valence to be the leading dimension of perfume vocabulary, separating source-based terms from abstract and evaluative ones. Likewise, the absence of a separate Floral factor may reflect questionnaire coverage and rotation, with floral content distributed across dimensions.

Comparison at the level of individual axes also requires attention to what generates a score. In a composition-based radar, a value depends on modeled vapor concentrations, odor thresholds,

Table 2*Directional Agreement in the Fixed Reference LLM Accuracy Audit*

Measure	Original	Corrected
Paired assessable cells, n	432	432
Directional agreements, n	389	418
Design-weighted agreement, % [95% CI]	90.7 [87.7, 93.3]	96.9 [94.9, 98.6]
Direct-evidence agreement, % (n = 70)	87.4	97.2
Sensitivity to reference Yes, %	70.5	90.2
Specificity to reference No, %	94.5	98.2
Balanced accuracy, %	82.5	94.2
Always-No comparator, %	84.0	84.0

Note. All percentages use the original design weights and the same 432 assessable cells, except the direct tier, which uses its 70-cell subset. No = ratings 0/1; Yes = ratings 3/4. Balanced accuracy is the mean of sensitivity and specificity against the reference. The paired gain is 6.2 percentage points, conditional 95% interval [4.2, 8.6]. Direct source support does not establish sensory ground truth. The initial 433-cell audit reported 90.5%; exclusion of one now-missing prior disagreement yields the paired baseline of 90.7%. Bracketed values are conditional 95% fragrance-cluster bootstrap intervals; the original agreement interval uses the paired 432-cell sample.

and the family assignments of ingredients. In FragranceDive, a value depends on the loading pattern and the pattern of uncertain or definite endorsements of product attributes. A higher score can therefore reflect more strongly supported descriptions without an equivalent increase in perceived intensity. For applied sensory research, the first use of these coordinates is to select and characterize candidate stimuli; treating distances as sensory quantities would require additional measurement.

Masculine links leather and male marketing with lower floral and indolic ratings. Lindqvist's (2012) graded perceptual gender associations support interpreting such characteristics continuously, but the present loadings remain properties of these products and annotations. They need not reproduce blind judgments or persist across cultures. Analyses separating scent items from market information would test how much of the structure depends on commercial associations.

Language and cultural context deserve explicit treatment. Jahai speakers can name odors consistently using abstract odor vocabulary, unlike the predominantly source-based descriptions of matched English speakers (Majid & Burenhult, 2014). In Chinese perfume reviews, Zhou et al. (2026) found source-based terms alongside scene-related, abstract, and evaluative language, with variation even within the same perfume. Their selected 36-product corpus and predominantly positive sentiment cannot establish equivalent dimensions across markets. It instead suggests that multilingual extensions should test attribute meanings and include consumer context and emotion rather than merely translate the English questionnaire. Nikitina's (2024) bilingual fragrance lexicon offers candidate vocabulary for that work.

The 14-factor solution provides one useful organization of this instrument and corpus. Interpretability guided retention, with MAP providing subsequent support for the selected count. The aim was a detailed, usable exploratory representation, rather than a claim about the intrinsic dimensionality of olfaction. Repeated content weights some attributes more heavily: male marketing appears three times, and four Expensive indicators represent two distinct questions. Future analyses can evaluate structural stability by consolidating duplicates and replicating across fragrance houses, release periods, and independent annotations. The relevant outcomes are the stability of salient loadings, cross-loadings, and product scores under matched scaling, together with preservation of the substantive factor interpretations.

What Targeted Correction Improves

Correction increased matched median polychoric correlation from .790 to .944 and ICC(A,1) from .580 to .869 in Random200. The audit gain supplies an additional criterion because making duplicates agree does not establish which response is supported. Multivariate prediction identifies unusual answers; web-assisted adjudication assesses them. External-tool correction has precedent in other LLM tasks (Gou et al., 2024), but rare combinations may be defining fragrance features and must survive review when supported.

The comparison evaluates the workflow as a whole; Famous50 received comprehensive review, whereas Random200 used a limited review budget. It does not isolate statistical triage from the newer model, higher reasoning effort, focused prompts, or additional searching. Randomly assigning equally reviewed cells to outlier-based versus random selection would test efficiency; withholding current answers and predictions could test anchoring. Full-corpus correction performance and stability of a refitted factor solution also remain untested. Supplementary Discussion specifies how such a comparison should be designed and scored.

Duplicate predictors make post-correction agreement partly an intended product of the procedure. Independent reassessment requires new judgments without access to those responses. Matched samples offer the clearer before-and-after comparison, whereas available cases describe the more complete corrected matrix. The corrected matched Random200 petrichor pair illustrates why category frequencies matter: identical constant responses produced 100% exact agreement but undefined polychoric and ICC estimates (Table S5). Perfect agreement in that case provides no information about distinguishing fragrances with and without the characteristic. Conversely, vetiver showed a higher available-case ICC alongside lower exact agreement. Neither coefficient nor percentage agreement should stand alone. Cicchetti and Feinstein (1990) showed that a single omnibus concordance index can be driven by marginal imbalance and by disparity between agreement on positive and on negative decisions, and recommended reporting those directional components separately, as Table 2 does for the audit. The full confusion matrices, sample sizes, and estimability diagnostics are part of interpreting the method.

Accuracy and the Coverage of Evidence

The fixed benchmark provides clear evidence of improvement within its assessable domain: 30 disagreements were resolved against one introduced, and balanced accuracy increased from 82.5% to 94.2%. An all-No rule would achieve 84.0% weighted agreement on these cells, underscoring the value of reporting affirmative-reference sensitivity alongside overall agreement. Freezing the audit ledger, targets, weights, and reference judgments preserved a consistent basis for comparison, and correction preceded the audits.

The 96.9% agreement estimate applies to assessable, previously answered cells. Weighted evidence coverage was 36.8%, most assessments concerned notes, and all 2,194 newly filled Random200 cells lay outside the audit sample. This coverage defines the scope of the result: improvement on the assessable benchmark, with overall database accuracy remaining an open empirical question. The conditional bootstrap intervals quantify sampling variation given the frozen evidence; they do not capture reference error or nonrandom assessability.

Coverage reflects the source material. The absence of off-flavor vocabulary in Hamilton and Lahne's (2020) review-derived whisky wheel illustrates that unrecorded attributes need not be random. Similarly, fragrance notes receive more documentary attention than standardized performance over time. Zhou et al.'s (2026) predominantly positive review corpus further shows why evaluative distributions in product text need not mirror an unrestricted sensory sample. Agreement where documentation is dense should not be generalized to sparse domains.

The improvement also persisted after excluding the 11 preview-exposed fragrance rows: weighted agreement increased from 90.3% to 97.1%, a gain of 6.8 percentage points. This sensitivity analysis strengthens the comparison while leaving reference quality and shared sources as limitations. Only 70 paired cells had direct support; most judgments were provisional, and documentary support is distinct from sensory verification. A stronger benchmark would combine independently adjudicated documentary evidence with sensory measurements, preserve unresolved cases, and report results by evidence type and attribute domain. Miller, Hamilton, and Lahne (2021) illustrate one workable construction for the documentary component: sensory-science annotators labeled whisky-review terms independently, and a two-of-three majority across raters resolved disagreements into a single standard. Their descriptor classifier also lost recall when moved from professional to hobbyist

reviews, demonstrating the value of testing both annotations and reference standards across source types.

Evaluator independence deserves direct assessment. Panickssery, Bowman, and Feng (2024) found that language-model evaluators favored their own generated summaries beyond differences supported by human judgments and linked this tendency to self-recognition through fine-tuning experiments and controls. The present audit used a different provider's model to evaluate numeric Grok annotations, so their experiments identify a methodological concern rather than an estimate of bias in this study. Future audits can strengthen independence by deriving judgments from underlying evidence before revealing candidate responses, masking the generating model and correction status, and including independent human adjudication. These measures would also help assess dependence arising from shared sources or similar evidence conventions.

Uncertainty and the Meaning of the Response Scale

Definitely Yes to a vanilla question expresses stronger certainty that the criterion is met than Probably Yes; it does not measure stronger vanilla intensity. Ordered-category modeling may therefore capture variation in evidence strength as well as fragrance characteristics. This differs from rate-all-that-apply sensory measurement, where respondents identify applicable attributes and rate perceived intensity. Across seven consumer studies, Vidal, Ares, Hedderley, Meyners, and Jaeger (2018) found broadly similar results from check-all-that-apply and rate-all-that-apply questions, without a general discrimination advantage from the added rating step. Ares et al. (2014), who introduced the format, separated two versions of that step: rating perceived intensity and rating how well a term applies. The applicability version is the nearer analogue to the present gradations, which grade certainty that a criterion is met rather than perceived strength. Rating modestly improved sample discrimination in three of their four studies, yet consumers used the full range of scale points for terms they had already endorsed. Graded endorsement can therefore add information while remaining unevenly used. Together, these findings motivate direct evaluation of what additional response gradations contribute. The present audit evaluated No/Yes direction; future comparisons can test the probably/definitely distinction, threshold comparability across questions, and the relative usefulness of four substantive categories versus binary endorsements against appropriate sensory or documentary criteria.

The adjudicator's continuous confidence also warrants calibration. Studies find that verbalized confidence can carry useful information but varies with model, task, and prompting (Tian et al., 2023; Xiong et al., 2024). The .70 acceptance threshold was an operational rule whose relationship to correctness can be evaluated empirically. Original code-2 responses were converted to blanks before analysis and treated as missing; they were not retained as substantive categories or separate informative indicators. Future versions could distinguish unsupported answers, inapplicable questions, and technical nonresponse to investigate whether these missingness categories provide useful information. Supplementary Discussion develops these prospective analyses.

Perceptual Validation and Product Identity

The study by Kurfalı et al. (2025) provides a direct rationale for sensory validation: a language model's fit to odor-related judgments depends on whether those judgments come from smelling, imagining odors from words, or matching labels to odors. Their results support recoverable olfactory information in language, but do not validate any model's commercial-fragrance ratings. FragranceDive should similarly be tested against distinct criteria rather than assuming that coherent semantics imply equivalent perceptual distances.

The study by Teixeira et al. (2014) further illustrates why the validation endpoint must be specified. Its five formulated mixtures received new expert sensory classifications, with generally favorable correspondence for the dominant family, although the first mixture differed and perfumers did not recover most finer model-predicted nuances. The separate 36-product analysis compared calculated profiles with existing fragrance-house and expert classifications rather than newly collected sensory ratings of all 36 products. Its results therefore supply both a sensory proof of concept and a documentary comparison, with different strengths. For FragranceDive, a matched study could compare fixed factor scores, composition-based profiles where analytical data are available, and blind sensory judgments of the same releases. Agreement should be evaluated separately for broad family membership, individual descriptors, and interproduct distances; success at one resolution cannot be assumed to extend to another.

A feasible study could freeze database profiles for a small product set before collecting blind attribute ratings and sorting or similarity judgments. Consumer profiling and mapping methods offer different balances of burden and detail (Varela & Ares, 2012), while the coffee comparison of Espinosa-Ramos et al. (2025) illustrates evaluation of text-derived and sensory profiles on matched products. Samples should span the factor space, include poorly documented releases and large correction-induced shifts, and compare original and corrected predictions against the same assessors. Distance metrics and factor weights should be specified in advance because equal display ranges do not establish perceptually equivalent units. Supplementary Discussion sets out the corresponding similarity design.

Analytical sensory tests and affective consumer tests address different outcomes (Drake, Watson, & Liu, 2023). A trained descriptive panel could evaluate scent attributes using explicit definitions, reference materials, and instruction in intensity scaling, with replicate presentations to quantify repeatability. Separately recruited target consumers could judge liking, situational appropriateness, and recommendation usefulness. Panel expertise should be treated as domain- and task-specific. Croijmans and Majid (2016) found greater descriptive consistency for wine experts describing wine, but no corresponding advantage for coffee experts describing coffee and no general superiority in identifying everyday odors or basic tastes. Both expert groups nevertheless used more source-based vocabulary within their domains. Familiarity with specialist terminology is therefore insufficient evidence of agreement. A fragrance validation report should document training, attribute interpretation, and repeatability rather than treating professional status as a substitute for those measurements.

Blind and identity-visible conditions would distinguish sensory correspondence from effects of names, houses, and marketing. Semantic labels can alter odor pleasantness even when the stimulus is unchanged (de Araujo, Rolls, Velazco, Margot, & Cayeux, 2005). Randomized or counterbalanced presentation would help control order and exposure effects. Price and

availability instead require verified commercial records, and longevity requires standardized observation over time. Each dimension needs a criterion appropriate to its content.

Validation must match the annotated release, concentration, and formulation. Teixeira et al. (2014) chemically compared commercial variants, including eau de toilette, eau fraîche, and aftershave products, and found compositional differences accompanied by differences in predicted intensity or character. A shared product name is therefore insufficient to establish a match. Ingredient restrictions provide a further reason to retain formulation dates. Following the Scientific Committee on Consumer Safety's (2012) assessment, Regulation (EU) 2017/1410 prohibited placing cosmetic products containing HICC, atranol, or chloroatranol on the Union market from 23 August 2019 and making them available from 23 August 2021 (European Commission, 2017). Separately, Regulation (EU) 2023/1545 expanded individual fragrance-allergen labeling, with transition deadlines of 31 July 2026 for placing noncompliant products on the market and 31 July 2028 for making them available (European Commission, 2023). These labeling deadlines do not govern the earlier ingredient prohibitions, and relabeling alone does not demonstrate a perceptual change.

Declared allergens provide verifiable label information beyond the attributes assessed here. Prospective studies applying fragrance to skin should use appropriate ethics review, exclusion criteria, and safety procedures for their protocol. The present correction prompt prioritized specific releases and current formulations, but sources may not identify them clearly. Validation records should link the sampled product to its concentration and batch or formulation information where available, and measurements at specified times after application should distinguish immediate character from subsequent evolution. Vintage/Waxy does not establish a chemical history of perfume, and product coordinates should retain the source or retrieval dates needed to interpret later updates.

Applications and Further Development

The immediate application is structured exploration and selection of products for closer study. Profiles can guide experiments comparing fragrances similar on some attributes and different on others. Recommendation systems could weight scent, price, and performance explicitly, but should be evaluated against consumer judgments or behavior and simpler benchmarks such as note overlap and family labels. Kalashi, Saed, and Teimourpour (2026) supply one behavioral benchmark of this kind. Using reviews from a Persian retail platform, they linked perfumes endorsed by the same users and recovered communities corresponding to fragrance families, gender orientation, and warm versus fresh character. Because that structure comes from which products people liked together rather than from documented attributes, it offers an external comparison for the scent-related factors here. The authors did not separate shared affect from co-review frequency, so the comparison would test convergence between descriptive profiles and co-preference structure rather than sensory equivalence. Similarity, liking, and choice are separate outcomes; none was directly measured here.

Fixed reference parameters allow annotation revisions to be compared on stable axes. Because both the factor model and scaling constants were held fixed, movements in the plots can be attributed to changes in the input ratings. The -4 to $+4$ display metric uses percentile-defined units, with zero at the midpoint of the reference anchors. This design makes annotation changes easy to inspect and provides a consistent basis for subsequent tests of sensory correspondence. Future evaluation can additionally assess uncertainty arising from incomplete

or unstable responses. Supplementary Methods records the precision limits of this display metric.

The archive supports verification and reuse through scripts, machine-readable outputs, file inventories, and checksums. Its documented provenance and version information address core requirements for reusable chemosensory data discussed by Parma et al. (2026). FragranceDive also complements molecular and mixture modeling (Tom et al., 2025): linking product profiles to formulations, concentrations, and ingredient identities would connect these levels in future work.

Further accuracy assessment should probability-sample original, newly filled, and unresolved cells across attribute domains and documentation levels. Methods combining abundant predictions with smaller validated samples offer a framework for specified inferential targets (Angelopoulos, Bates, Fannjiang, Jordan, & Zrnic, 2023; Egami et al., 2023); confidence-guided sampling can improve efficiency when appropriately designed (Gligorić, Zrnic, Lee, Candès, & Jurafsky, 2025). These methods require suitable reference measurements and do not automatically validate factor scores. Extended methodological implications appear in Supplementary Discussion.

FragranceDive demonstrates scalable organization of documented commercial-fragrance characteristics and measurable improvement of an annotation matrix through review. Interpretable dimensions, stronger repeated-response consistency, and improved agreement with a frozen source-based reference provide complementary evidence for the approach. Fixed reference scoring and an inspectable archive support reuse and further evaluation. These results provide a practical foundation for targeted stimulus selection and subsequent sensory and consumer validation.

Acknowledgments

ChatGPT (OpenAI), Claude (Anthropic), and Grok (xAI) assisted with drafting and revising manuscript text, integrating literature supplied by the author, and preparing submission materials. The author retains responsibility for the interpretation, references, and final content. Research uses of language models for annotation, correction, and audit are described in Materials and Methods.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Competing Interests

The author declares no competing interests.

Author Contributions

Tyler M. Moore: Conceptualization, methodology, software, formal analysis, investigation, data curation, visualization, writing – original draft, and writing – review and editing.

Data and Code Availability

All data underlying this study, including the full reference corpus and the original and corrected validation-subset ratings, will be publicly released at publication through Zenodo (<https://doi.org/10.5281/zenodo.22755429>). The FragranceDive data and questionnaire are dedicated to the public domain under CC0 1.0 Universal; the scripts are released under the MIT License. These terms permit reuse, modification, redistribution, and commercial use, with the MIT copyright and license notice retained in copies of the software.

FragranceDive_Supplementary_Data.zip is supplied directly with the article and includes the license notices, working R source, Random200 sampling code, pattern loadings and factor correlations, repeated-item outputs, both complete audit packages, correction API response records, and the masking sensitivity analysis, with file inventories and checksums. The study steps, including manual preprocessing, are documented, but the workflow is not a single-execution pipeline. The article archive reproduces the audit calculations and permits checks of the distributed factor and reliability outputs; reconstructing the full factor model and scatterplots additionally requires the repository data and the documented stage-specific operations.

Release update — October 2026

The public FragranceDive research archive now includes the full raw reference matrix (35,449 fragrances × 90 questionnaire items) and the secondary factor-score dataset. Both are downloadable directly from the FragranceDive website.

References

- Abdurahman, S., Salkhordeh Ziabari, A., Moore, A. K., Bartels, D. M., & Dehghani, M. (2025). A primer for evaluating large language models in social-science research. *Advances in Methods and Practices in Psychological Science*, 8, Article 25152459251325174. <https://doi.org/10.1177/25152459251325174>
- Angelopoulos, A. N., Bates, S., Fannjiang, C., Jordan, M. I., & Zrnic, T. (2023). Prediction-powered inference. *Science*, 382, 669–674. <https://doi.org/10.1126/science.adi6000>
- Ares, G., Bruzzone, F., Vidal, L., Silva Cadena, R., Giménez, A., Pineau, B., . . . Jaeger, S. R. (2014). Evaluation of a rating-based variant of check-all-that-apply questions: Rate-all-that-apply (RATA). *Food Quality and Preference*, 36, 87–95. <https://doi.org/10.1016/j.foodqual.2014.03.006>
- Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5–32. <https://doi.org/10.1023/A:1010933404324>
- Castro, J. B., Ramanathan, A., & Chennubhotla, C. S. (2013). Categorical dimensions of human odor descriptor space revealed by non-negative matrix factorization. *PLoS ONE*, 8, Article e73289. <https://doi.org/10.1371/journal.pone.0073289>
- Chastrette, M., Elmouaffek, A., & Sauvegrain, P. (1988). A multidimensional statistical study of similarities between 74 notes used in perfumery. *Chemical Senses*, 13, 295–305. <https://doi.org/10.1093/chemse/13.2.295>
- Cicchetti, D. V., & Feinstein, A. R. (1990). High agreement but low kappa: II. Resolving the paradoxes. *Journal of Clinical Epidemiology*, 43, 551–558. [https://doi.org/10.1016/0895-4356\(90\)90159-M](https://doi.org/10.1016/0895-4356(90)90159-M)
- Croijmans, I., & Majid, A. (2016). Not all flavor expertise is equal: The language of wine and coffee experts. *PLOS ONE*, 11, Article e0155845. <https://doi.org/10.1371/journal.pone.0155845>
- de Araujo, I. E., Rolls, E. T., Velazco, M. I., Margot, C., & Cayeux, I. (2005). Cognitive modulation of olfactory processing. *Neuron*, 46, 671–679. <https://doi.org/10.1016/j.neuron.2005.04.021>
- Drake, M. A., Watson, M. E., & Liu, Y. (2023). Sensory analysis and consumer preference: Best practices. *Annual Review of Food Science and Technology*, 14, 427–448. <https://doi.org/10.1146/annurev-food-060721-023619>
- Egami, N., Hinck, M., Stewart, B. M., & Wei, H. (2023). Using imperfect surrogates for downstream inference: Design-based supervised learning for social science applications of large language models. *Advances in Neural Information Processing Systems*, 36, 68589–68601. <https://doi.org/10.52202/075280-3000>
- Espinosa-Ramos, J. I., Hutchings, S. C., Dixit, Y., Jin, D., Ryan, G. S., Marsden, M., . . . Reis, M. M. (2025). Comparison of free-comment online product reviews and central location product testing for sensory product characterisation: A case study with coffee consumers. *Food Quality and Preference*, 125, Article 105377. <https://doi.org/10.1016/j.foodqual.2024.105377>
- European Commission. (2017). Commission Regulation (EU) 2017/1410 of 2 August 2017 amending Annexes II and III to Regulation (EC) No 1223/2009 of the European Parliament and of the Council on cosmetic products. *Official Journal of the European Union*, L 202, 1–3. <http://data.europa.eu/eli/reg/2017/1410/oj>
- European Commission. (2023). Commission Regulation (EU) 2023/1545 of 26 July 2023 amending Regulation (EC) No 1223/2009 of the European Parliament and of the Council as regards labelling of fragrance allergens in cosmetic products. *Official Journal of the European Union*, L 188, 1–23. <http://data.europa.eu/eli/reg/2023/1545/oj>
- Gilardi, F., Alizadeh, M., & Kubli, M. (2023). ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120, Article e2305016120. <https://doi.org/10.1073/pnas.2305016120>

- Gligorić, K., Zrnic, T., Lee, C., Candès, E. J., & Jurafsky, D. (2025). Can unconfident LLM annotations be used for confident conclusions? In L. Chiruzzo, A. Ritter, & L. Wang (Eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)* (pp. 3514–3533). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2025.naacl-long.179>
- Gou, Z., Shao, Z., Gong, Y., Shen, Y., Yang, Y., Duan, N., & Chen, W. (2024). CRITIC: Large language models can self-correct with tool-interactive critiquing. In *The Twelfth International Conference on Learning Representations*. <https://openreview.net/forum?id=Sx038qxjek>
- Hamilton, L. M., & Lahne, J. (2020). Fast and automated sensory analysis: Using natural language processing for descriptive lexicon development. *Food Quality and Preference*, 83, Article 103926. <https://doi.org/10.1016/j.foodqual.2020.103926>
- Hörberg, T., Kurfali, M., & Olofsson, J. K. (2025). Chemosensory vocabulary in wine, perfume and food product reviews: Insights from language modeling. *Food Quality and Preference*, 124, Article 105357. <https://doi.org/10.1016/j.foodqual.2024.105357>
- Hörberg, T., Larsson, M., & Olofsson, J. K. (2022). The semantic organization of the English odor vocabulary. *Cognitive Science*, 46, Article e13205. <https://doi.org/10.1111/cogs.13205>
- Kaeppler, K., & Mueller, F. (2013). Odor classification: A review of factors influencing perception-based odor arrangements. *Chemical Senses*, 38, 189–209. <https://doi.org/10.1093/chemse/bjs141>
- Kalashi, K., Saed, S., & Teimourpour, B. (2026). Sentiment-driven community detection in a network of perfume preferences. *Applied Network Science*, 11, Article 13. <https://doi.org/10.1007/s41109-025-00757-0>
- Kandpal, N., Deng, H., Roberts, A., Wallace, E., & Raffel, C. (2023). Large language models struggle to learn long-tail knowledge. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, & J. Scarlett (Eds.), *Proceedings of the 40th International Conference on Machine Learning* (Vol. 202, pp. 15696–15707). PMLR. <https://proceedings.mlr.press/v202/kandpal23a.html>
- Keller, A., Gerkin, R. C., Guan, Y., Dhurandhar, A., Turu, G., Szalai, B., . . . Meyer, P. (2017). Predicting human olfactory perception from chemical features of odor molecules. *Science*, 355, 820–826. <https://doi.org/10.1126/science.aal2014>
- Khan, R. M., Luk, C.-H., Flinker, A., Aggarwal, A., Lapid, H., Haddad, R., & Sobel, N. (2007). Predicting odor pleasantness from odorant structure: Pleasantness as a reflection of the physical world. *The Journal of Neuroscience*, 27, 10015–10023. <https://doi.org/10.1523/JNEUROSCI.1158-07.2007>
- Kurfali, M., Herman, P., Pierzchajlo, S., Olofsson, J., & Hörberg, T. (2025). Representations of smells: The next frontier for language models? *Cognition*, 264, Article 106243. <https://doi.org/10.1016/j.cognition.2025.106243>
- Lee, B. K., Mayhew, E. J., Sanchez-Lengeling, B., Wei, J. N., Qian, W. W., Little, K. A., . . . Wiltschko, A. B. (2023). A principal odor map unifies diverse tasks in olfactory perception. *Science*, 381, 999–1006. <https://doi.org/10.1126/science.ade4401>
- Lindqvist, A. (2012). Perfume preferences and how they are related to commercial gender classifications of fragrances. *Chemosensory Perception*, 5, 197–204. <https://doi.org/10.1007/s12078-012-9119-7>
- Majid, A., & Burenhult, N. (2014). Odors are expressible in language, as long as you speak the right language. *Cognition*, 130, 266–270. <https://doi.org/10.1016/j.cognition.2013.11.004>
- McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, 1, 30–46. <https://doi.org/10.1037/1082-989X.1.1.30>
- Meiselman, H. L., Jaeger, S. R., Carr, B. T., & Churchill, A. (2022). Approaching 100 years of sensory and consumer science: Developments and ongoing issues. *Food Quality and Preference*, 100, Article 104614. <https://doi.org/10.1016/j.foodqual.2022.104614>

- Miller, C., Hamilton, L., & Lahne, J. (2021). Sensory descriptor analysis of whisky lexicons through the use of deep learning. *Foods*, 10, Article 1633. <https://doi.org/10.3390/foods10071633>
- Moore, T. M. (2025). *Using large language models to obtain quantitative data* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/aqu9x_v1
- Moore, T. M. (2026a). *A language model generates seven million phenotype estimates across the tree of life* [Manuscript submitted for publication]. Department of Psychiatry, Perelman School of Medicine, University of Pennsylvania.
- Moore, T. M. (2026b). *PharmaDive: A language model generates an uncertainty-graded property matrix for more than 16,000 compounds* [Manuscript submitted for publication]. Department of Psychiatry, Perelman School of Medicine, University of Pennsylvania.
- Nikitina, L. (2024). Fragrance lexicon for analysis of consumer-generated perfume reviews in Russian and English. *MethodsX*, 12, Article 102627. <https://doi.org/10.1016/j.mex.2024.102627>
- Olsson, U. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika*, 44, 443–460. <https://doi.org/10.1007/BF02296207>
- OpenAI. (2026). *ChatGPT* [Large language model]. <https://chatgpt.com>
- Panickssery, A., Bowman, S. R., & Feng, S. (2024). LLM evaluators recognize and favor their own generations. *Advances in Neural Information Processing Systems*, 37, 68772–68802. <https://doi.org/10.52202/079017-2197>
- Parma, V., Mainland, J. D., Hwang, L.-D., Kedziora, R. J., Pineau, N., Gerkin, R. C., & Niv, M. Y. (2026). FAIR chemosensory data: Unlocking AI for flavour, food and health. *Trends in Food Science & Technology*, 169, Article 105545. <https://doi.org/10.1016/j.tifs.2026.105545>
- R Core Team. (2025). *R: A language and environment for statistical computing* [Computer software]. R Foundation for Statistical Computing. <https://www.R-project.org/>
- Revelle, W. (2025). *psych: Procedures for psychological, psychometric, and personality research* [R package]. Northwestern University. <https://CRAN.R-project.org/package=psych>
- Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, 34(4, Pt. 2), 1–97.
- Scientific Committee on Consumer Safety. (2012). *Opinion on fragrance allergens in cosmetic products* (SCCS/1459/11). European Commission. https://ec.europa.eu/health/scientific_committees/consumer_safety/docs/sccs_o_102.pdf
- Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86, 420–428. <https://doi.org/10.1037/0033-2909.86.2.420>
- Teixeira, M. A., Barrault, L., Rodríguez, O., Carvalho, C. C., & Rodrigues, A. E. (2014). Perfumery Radar 2.0: A step toward fragrance design and classification. *Industrial & Engineering Chemistry Research*, 53, 8890–8912. <https://doi.org/10.1021/ie403968w>
- Tian, K., Mitchell, E., Zhou, A., Sharma, A., Rafailov, R., Yao, H., . . . Manning, C. D. (2023). Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 5433–5442). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.emnlp-main.330>
- Tom, G., Ser, C. T., Rajaonson, E. M., Lo, S., Park, H. S., Lee, B. K., & Sanchez-Lengeling, B. (2025). Does this smell the same? Learning representations of olfactory mixtures using inductive biases. *Machine Learning: Science and Technology*, 6, Article 035063. <https://doi.org/10.1088/2632-2153/adfffc>
- Törnberg, P. (2024). Best practices for text annotation with large language models. *Sociologica*, 18, 67–85. <https://doi.org/10.6092/issn.1971-8853/19461>

- Varela, P., & Ares, G. (2012). Sensory profiling, the blurred line between sensory and consumer science. A review of novel methods for product characterization. *Food Research International*, 48, 893–908. <https://doi.org/10.1016/j.foodres.2012.06.037>
- Velicer, W. F. (1976). Determining the number of components from the matrix of partial correlations. *Psychometrika*, 41, 321–327. <https://doi.org/10.1007/BF02293557>
- Veramendi, M., Herencia, P., & Ares, G. (2013). Perfume odor categorization: To what extent trained assessors and consumers agree? *Journal of Sensory Studies*, 28, 76–89. <https://doi.org/10.1111/joss.12025>
- Vidal, L., Ares, G., Hedderley, D. I., Meyners, M., & Jaeger, S. R. (2018). Comparison of rate-all-that-apply (RATA) and check-all-that-apply (CATA) questions across seven consumer studies. *Food Quality and Preference*, 67, 49–58. <https://doi.org/10.1016/j.foodqual.2016.12.013>
- Wilcox, R. R. (2021). *Introduction to robust estimation and hypothesis testing* (5th ed.). Academic Press.
- Wright, M. N., & Ziegler, A. (2017). ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, 77(1), 1–17. <https://doi.org/10.18637/jss.v077.i01>
- xAI. (2026a). Grok 4.3 [Large language model]. <https://x.ai>
- xAI. (2026b). Grok 4.6 [Large language model]. <https://x.ai>
- Xiong, M., Hu, Z., Lu, X., Li, Y., Fu, J., He, J., & Hooi, B. (2024). Can LLMs express their uncertainty? An empirical evaluation of confidence elicitation in LLMs. In *The Twelfth International Conference on Learning Representations*. <https://openreview.net/forum?id=gjeQKFxFpZ>
- Zarzo, M. (2015). Underlying dimensions in the descriptive space of perfumery odors: Part II. *Food Quality and Preference*, 43, 79–87. <https://doi.org/10.1016/j.foodqual.2015.02.018>
- Zarzo, M. (2020a). Multivariate analysis of olfactory profiles for 140 perfumes as a basis to derive a sensory wheel for the classification of feminine fragrances. *Cosmetics*, 7, Article 11. <https://doi.org/10.3390/cosmetics7010011>
- Zarzo, M. (2020b). Understanding the perceptual spectrum of commercial perfumes as a basis for a standard sensory wheel of fragrances. *Cosmetics*, 7, Article 3. <https://doi.org/10.3390/cosmetics7010003>
- Zarzo, M., & Stanton, D. T. (2009). Understanding the underlying dimensions in perfumers' odor perception space as a basis for developing meaningful odor maps. *Attention, Perception, & Psychophysics*, 71, 225–247. <https://doi.org/10.3758/APP.71.2.225>
- Zhou, H., Zhou, K., Zhu, Z., & Zou, L. (2026). The language of scents: A corpus-based study of Chinese perfume descriptions. *Humanities and Social Sciences Communications*, 13, Article 1149. <https://doi.org/10.1057/s41599-026-07494-4>
- Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z., & Yang, D. (2024). Can large language models transform computational social science? *Computational Linguistics*, 50, 237–291. https://doi.org/10.1162/coli_a_00502

FragranceDive Supplementary Materials

Mapping the language of fragrance with FragranceDive

Tyler M. Moore, Ph.D.

[Return to manuscript](#)

This supplement provides detailed annotation, correction, scoring, and audit methods; additional methodological interpretation; the complete factor pattern matrix, interfactor correlations, and item wording; repeated-item statistics and confusion matrices; additional factor-score illustrations; and detailed accuracy-audit results. The accompanying archive provides machine-readable results, original figures, the R workflow, complete audit records, and the supplied correction API responses.

Contents

Supplementary Methods: Collection provenance, parsing, factor scoring, correction, agreement statistics, audit procedures, and software.

Supplementary Discussion: Annotation precedents, correction mechanisms, confidence calibration, language, and additional validation details. Interpretive examples and panel-design recommendations are developed in the main manuscript.

Tables S1–S2: Item wording and all factor loadings.

Tables S3–S9: Response availability, all-pair agreement summaries, estimation diagnostics, and complete pairwise statistics.

Tables S10–S13: Audit coverage, changes, subgroup results, and sensitivity analyses.

Table S14: The oblique factor-correlation matrix (Φ).

Figures S1–S2: Additional original and corrected factor-score illustrations.

Figures S3–S6: Every repeated-item matrix, under matched and available sampling, for both cohorts.

Supplementary Methods

These methods provide implementation and provenance details for the concise procedures in the main manuscript. The analyzed corpus derives from the Grok batch output; the earlier collection is documented separately below.

Fragrance Corpus and Item Set

The corpus comprised 35,449 fragrances (perfumes, colognes, eaux de toilette, extraits, and related products). Each entry was stored as a single string of the form Name (House), with the parenthetical house serving as the authoritative disambiguator, because fragrance names collide frequently across houses.

Each fragrance was rated on 90 characteristic-question columns representing 60 distinct question wordings, including note presence and prominence, olfactive family, perceived gender coding, price, projection, longevity, and market familiarity. Twenty-eight questions occurred

twice, one occurred three times, and 31 occurred once. Full wording is provided in Table S1. Every item used the same collection scale: 0 = Definitely No, 1 = Probably No, 2 = Unknown or not applicable, 3 = Probably Yes, and 4 = Definitely Yes. Code 2 was a collection-stage marker for Unknown or not applicable. All such entries were converted to blanks (missing values) before statistical analysis. No separate code-2 indicators, uncertainty weights, or analyses of the reasons for missingness were used.

Reuse license. FragranceDive data and questionnaire materials are dedicated to the public domain under CC0 1.0 Universal, allowing reuse for any purpose, including commercial use, without separate permission. The accompanying scripts are released under the MIT License, which also permits commercial reuse, modification, and redistribution provided the copyright and license notice is retained.

A subset of questions was repeated at separated positions within the same instrument without identifying the repetitions to the annotating model. These repeated columns provide a check on within-model response consistency and form the basis of the repeated-item analyses reported below.

Initial Annotation

Initial annotation used the Anthropic Messages API with Claude Haiku 4.5 (claude-haiku-4-5-20251001; Anthropic, 2025). Fragrances were processed in batches of four using a CSV-only system prompt, the full item list, the server-side web search tool (web_search_20250305; up to five searches per request), and a sampling temperature of 0. The request wrapper allowed up to six attempts, with backoff for transient failures, and continued paused responses. A separate batch-level loop allowed up to four attempts in total for failed or entirely Unknown output. Unparseable and out-of-range cells were assigned code 2; short response blocks were padded with 2, and excess rows or columns were truncated. Progress was checkpointed periodically and at completion.

This initial approach was discontinued on cost grounds after annotating approximately 15,000 fragrances on a 60-item version of the instrument. Subsequent annotation used the batch procedure described below. The supplied analysis pipeline reads the batch-annotation output and contains no step that merges the initial Claude ratings into that file.

Batch Annotation

Primary annotation used the xAI Batch API with Grok 4.3 (xAI, 2026a) at medium reasoning effort.

Fragrance names were sanitized for JSON transport by replacing double quotes and backticks with apostrophes and stripping backslashes, while preserving valid UTF-8 (including Cyrillic and accented characters). Requests were constructed in batches of five fragrances, each carrying the complete item list, the response scale, the web search tool, and explicit formatting constraints: exactly one row per fragrance in the order listed, exactly one integer per item, no empty cells, and no text outside the CSV. Prompts instructed the model that many entries are obscure and that thorough web search would be required, and that 2 should be used when an answer could not be established after effortful search.

Requests were serialized to JSONL, uploaded, submitted as a single batch job, and polled to completion; results were retrieved with pagination. The full result set was written to disk before parsing so that parsing could be repeated without re-incurring inference cost.

Response Parsing and Provenance

The results matrix was initialized to 2 and overwritten by successfully parsed values. Research-based Unknown responses and initialization artifacts therefore shared the same numeric code; batch provenance was retained to identify likely unreturned responses. Parsing proceeded in three steps. First, answer columns were aligned to item positions using returned header labels; a batch was flagged and discarded when its headers could not be mapped and its column count did not match the expected count. Second, returned rows were matched to requested fragrances using normalized names, attempting exact matching followed by unambiguous substring matching and rejecting ambiguous matches. The normalization lowercased names and removed parenthetical houses and non-alphanumeric characters. Third, positional assignment was used as a fallback only when no name matched and the returned row count was exactly correct.

Parsed responses received statuses distinguishing clean parsing, column realignment, partially returned rows, unmatched rows, column-alignment failure, and parsing error. The problem log also recorded request and assignment failures, keyed to fragrance index ranges. These records identified batches for retry. The script generated retry requests for problematic batches and supported resumption from a saved batch identifier or checkpoint. Any remaining original code-2 entries were converted to blanks before statistical analysis.

Dimension Reduction and Factor Scoring

We summarized the rating matrix with an exploratory item-factor model using `irt.fa` in `psych` (Revelle, 2025), which derives graded-response-type item parameters from a factor analysis of polychoric correlations (Samejima, 1969). The two occurrences of the price-below-\$30-per-ounce question were reverse-keyed before modeling. Original Unknown responses (code 2) were treated as missing in both the reference-model and Famous50 scoring data. For the reference-model input, this conversion had been performed manually before the factor-analysis script was run; the Famous50 scoring code performs the same conversion explicitly. The remaining substantive responses were then recoded identically from 0, 1, 3, and 4 to the consecutive ordered categories 0, 1, 2, and 3. The recoded category 2 represents original Probably Yes responses and is unrelated to the original Unknown code. Exploratory solutions containing 8 through 15 factors were examined with oblimin rotation. Fourteen factors were retained because the solution provided interpretable distinctions, whereas the 15-factor solution was less clearly interpretable. A subsequent check using Velicer's minimum average partial (MAP) criterion (Velicer, 1976) also suggested 14 factors. The standardized root mean-square residual (SRMR = .04) describes residual fit of the retained solution; fit indices did not determine retention. The distributed factor-fitting block represents the selected 14-factor solution. The factor labels were Resinous/Smoky, Animalic, Expensive, Intense/Powerful, Earthy, Gourmand/Sweet, Masculine, Aquatic/Ozonic, Spicy/Woody, Carnation/Aldehydic, Minty, Fruity, Vintage/Waxy, and US Mainstream. The reported loadings are the oblimin-rotated pattern matrix. They were interpreted descriptively, using an absolute pattern loading of .30 to identify salient indicators; repeated question columns were retained separately.

The full corpus served as the reference set for all scoring. Factor scores (θ) were computed with `scoreIrt` using an absolute IRT discrimination cutoff of .30 ($\text{cut} = .30$), not a factor-loading cutoff. The separate $|\text{loading}| \geq .30$ criterion concerns interpretation of the pattern matrix. To place the 14 factors on a common and interpretable display metric, a single pair of scaling

constants was estimated per factor from the reference set alone: the 0.1st and 99.9th percentiles of that factor's θ distribution. Scores were clamped to that interval, winsorizing the most extreme 0.1% in each tail so that a small number of outlying estimates cannot compress the rest of the scale (Wilcox, 2021), and then mapped linearly onto a -4 to $+4$ range and rounded to two decimals. These display units are not standard deviations; zero is the midpoint of the percentile anchors rather than necessarily the reference mean. The scaling constants were written to a parameter file and held fixed thereafter, so that the display metric is a property of the reference set rather than of any particular analytic file.

The percentile transformation improves display comparability within a fixed model but does not equalize measurement precision across factors. A sparsely supported factor and a well-supported factor can occupy the same numerical range. Reporting which item responses contributed to a product's score would help users judge whether an apparent match rests on detailed information or on a limited response profile. Any future uncertainty intervals should account for the estimation or annotation process actually being varied, rather than treating the bounded display scale itself as a measure of certainty.

The fitted item parameters and reference-set scaling constants were applied to the original and corrected ratings of 50 widely recognized fragrances (Famous50). This set was selected with a general-purpose LLM (OpenAI, 2026), using sales, media recognition, and familiarity of the bottle and brand as indicators of well-knownness. Parameters and scaling constants were not re-estimated within this subset, so its scores were expressed on the reference metric. Values beyond the reference bounds were reported at the corresponding bound.

Bivariate factor-score plots were produced with ggplot2 (Wickham, 2016) and ggrepel (Slowikowski, 2024). Each displayed panel contains all 50 fragrances and 25 selected labels. Labels were prioritized using isolation, distance from the origin, and local density and could be displaced vertically with leader lines; score locations are indicated by points rather than label positions. Original and corrected panels use the same factor axes and reference metric. The faint identity line is a visual guide, not a fitted relationship or a test of equivalence between constructs.

Data Quality Correction Procedure

The correction procedure combined multivariate outlier detection with targeted LLM re-adjudication. Each rating column was predicted from the remaining columns, allowing cells that were inconsistent with their row profiles to be prioritized for review.

Validation Subsets

Evaluation used a random sample of 200 fragrances (Random200) and a set of 50 widely recognized fragrances (Famous50). Both subsets were analyzed before and after correction. Random200 was sampled without replacement from the final, ordered 35,449-row `fragrance_data_raw_forPub.csv` file using the following author-supplied R commands (seed 2026); these commands apply no fragrance-level exclusions. The same source file and row order are needed to reproduce the draw. The sampling script is also archived as `code/source/select_Random200.R`.

```
x <- read.csv("fragrance_data_raw_forPub.csv")
set.seed(2026)
write.csv(x[sample(1:nrow(x),200,replace=FALSE)],,
"fragrance_data_raw_200.csv",na="",row.names=FALSE)
```

Random200 also formed the sampling frame for the source-based accuracy audit. Familiarity-based selection of Famous50 was intended to provide a comparison among well-documented releases; it does not make that subset representative of the corpus.

Definition of Missing Data

For this study, every original code-2 response was converted to a blank by recoding or manual search-and-replace before the data entered the statistical analyses. Blanks, NA-like strings, and invalid entries were treated as missing. The analytical files therefore contained only substantive responses (0, 1, 3, and 4) and missing values; there was no analyzed Unknown category and the causes of missingness were not used as predictors or indicators. The general-purpose correction script retains code paths that can accept or return 2 for use with raw or future data, but those provisions do not give original code-2 responses an analytical role in the results reported here. Any newly returned Unknown response was likewise represented as a blank before analysis.

Multivariate Outlier Detection

When review slots remained after missing cells were prioritized, the target column was predicted from all remaining item columns with a probability random forest in ranger (Breiman, 2001; Wright & Ziegler, 2017). The supplied implementation fits each forest within the currently loaded validation subset, rather than loading a predictor trained on the full corpus. Forests used 500 trees, the integer part of the square root of the predictor count for `mtry`, and a minimum terminal node size of 10. Rows with a missing target were excluded from fitting. For software compatibility, the generic correction code temporarily substitutes the numeric value 2 for blank predictor entries while a forest is fit. This temporary computational placeholder is not an observed Unknown response, is not retained in the analysis data, and was not interpreted as evidence about the fragrance or the reason a value was missing.

Out-of-bag predicted class probabilities were used for scoring, so that each observation is evaluated by trees that did not see it. For every row, we computed the predicted probability of the observed category, the probability-weighted expected value, and the modal predicted category. The outlier score combined a surprisal term with two ordinal-distance terms:

$$\text{score} = -\log(p_{\text{observed}}) + 1.25 \times |\text{observed} - \text{expected}| + 0.75 \times |\text{observed} - \text{modal}|$$

The score increases when an observed substantive response has low model probability or differs from the expected value or modal category. Missing targets receive review priority separately. Original Unknown responses had already been replaced by blanks, so they did not enter this score as an intermediate No/Yes category. The outlier score prioritizes review and is not itself a test of factual correctness.

Adjudication and Correction

Within each target column and each sweep, up to five cells were selected for review. Eligible missing cells were selected first in their existing row order; when these filled the budget, the random-forest step was skipped. Otherwise, the remaining slots were filled with the highest-scoring eligible nonmissing cells. Selected cells were submitted in requests containing up to five fragrances and exactly one question.

Adjudication used Grok 4.6 via the xAI Responses API (xAI, 2026b), configured for high reasoning effort and required web search, with no temperature parameter sent and a 300-s timeout. The request wrapper allowed up to five attempts per request configuration and could

start another retry sequence after omitting parameters explicitly rejected by the endpoint. The adjudication prompt supplied the full item text, the response scale, the current value, and the model-derived triage fields, with an explicit instruction that the triage fields explain why the cell was selected but must not determine the answer. Two further prompt components were included to control known failure modes in this domain: a source-quality hierarchy preferring house materials, specialist trade and editorial coverage, reference works, and authorized retailer listings over user-contributed databases, which were designated acceptable for note pyramids and launch data but weak evidence for performance and reformulation claims; and a fragrance interpretation policy requiring that the specific release be identified before answering, that concentrations and flankers be treated as distinct products, that the current widely sold formulation be assumed absent explicit reference to a vintage, and that thinly documented releases be coded Unknown rather than inferred from siblings in the same line. The prompt further specified that an existing 2 must not be treated as evidence, that conflicting sources should yield 1 or 3 rather than 0 or 4, and that at least one source be supplied whenever confidence exceeded 0.5.

Responses were required in a fixed JSON schema containing the recommended value, a continuous confidence in $[0, 1]$, a change flag, a one-sentence rationale, and a source list. A recommendation was applied only when it parsed successfully, differed from the current value, was a valid integer in 0–4, and carried confidence of at least 0.70; the same threshold governed the filling of missing cells. Any accepted recommendation of Unknown was converted to a blank before statistical analysis. All adjudications were written to a review log regardless of whether they resulted in a change, and previously reviewed cells were excluded from later selection unless their value had changed in the interim, so that review effort was not spent twice on the same cell.

Random200 correction was configured for 20 sweeps with up to five review selections per column per sweep, yielding at most 100 selection slots per column. This is a review budget rather than a count of distinct cells reviewed or changed. Famous50 correction involved 10 sweeps, and every one of its 4,500 cells was reviewed. Thus, missingness and outlier scores determined review order in Famous50, but did not restrict final review to a subset of cells. Review could retain a value or leave uncertainty unresolved; it did not imply a change to every cell. The working matrix was refreshed before each target column was scored, allowing earlier corrections to inform later predictions. The random seed was 20260612, with each forest using that seed plus the target-column index. Sweep-level output, checkpoints, and review logs supported resumption after interruptions.

Reliability of Duplicated Items

Agreement between repeated occurrences of the same question was assessed separately in the original and corrected versions of Random200 and Famous50. Duplicate columns were identified by position and wording, including identical headers and numerically suffixed variants. All 31 within-question pairs were enumerated across 29 repeated questions; the question appearing three times contributed three overlapping pairs. Original Unknown responses had already been converted to missing values; complete-case analyses used only the substantive values 0, 1, 3, and 4.

For each pair we computed exact agreement, a two-step polychoric correlation with polycor (Fox, 2022; Olsson, 1979), and the single-measurement, two-way, absolute-agreement intraclass correlation, ICC(A,1), with model-based 95% confidence limits from irr (Gamer,

Lemon, & Fellows, 2019; McGraw & Wong, 1996; Shrout & Fleiss, 1979). Additional measures were the mean signed difference, mean absolute difference, and polarity agreement after coding 0/1 as No and 3/4 as Yes. ICC estimates were checked against the corresponding analysis-of-variance formula, and negative coefficients were retained. The ICC and difference statistics use the numeric distances of the original response codes; polychoric correlations use their ordering. Sparse marginal categories, boundary estimates, and estimation warnings were recorded.

Each pair was analyzed using matched complete cases, restricted to fragrances with both occurrences observed in both versions, and all available complete cases within each version. Numerical original-versus-corrected summaries emphasize matched samples. Median coefficients were compared across the same item pairs with estimable coefficients in both versions. All 31 pairs contributed to summaries of exact and polarity agreement. The selected main-text matrices use all available cases and can consequently differ in sample size. They illustrate contrasting patterns and are not a random selection of items. Supplementary figures display every pair under both sampling rules. Because other occurrences of a repeated question were available to the correction model as predictors, post-correction agreement describes the consistency of the corrected matrix and does not provide an independent estimate of annotation reliability.

External Accuracy Audit

GPT 6 Sol, accessed through ChatGPT (OpenAI, 2026), conducted a source-based accuracy audit of Random200. The auditor sampled six originally nonblank cells without replacement from each of the 200 fragrances, using Python `random.Random` with seed 20260910, yielding 1,200 targets. Duplicate columns were eligible on the same basis as other columns, and obscure fragrances were retained. For a fragrance with M originally entered ratings, each selected cell received the inverse-probability weight $M/6$. Reference assessments used a sample file with target ratings masked, and the reference ledger was frozen before joining it to the ratings. An initial structural preview had displayed the first 11 fragrance rows, so masking was incomplete. The auditor assigned a Yes or No reference judgment when source evidence permitted, distinguished direct source support from provisional inference, and retained unresolved references in coverage accounting. Directional agreement compared 0/1 with No and 3/4 with Yes; exact four-category accuracy and confidence calibration were not assessed.

The corrected file was compared against the identical sampled cells, frozen reference judgments, evidence tiers, question classifications, and original sampling weights. The paired comparison excluded cells that became missing. Conditional 95% intervals used 10,000 bootstrap replicates resampling whole fragrance rows with shared draws across versions, using NumPy seed 20260911. These intervals quantify sampling variation conditional on the available evidence and frozen references; they do not capture reference error or nonrandom assessability. Newly filled cells had no chance of entering the original sample and were not externally audited.

The masking sensitivity analysis excluded every sampled target from the first 11 fragrance rows exposed during the initial preview. It retained the frozen reference labels, paired-scoreability rule, and original $M/6$ weights for all other cells. A new set of 10,000 shared cluster draws used `numpy.random.default_rng(20260911)`, resampling all 189 remaining sampled fragrance rows, including rows without assessable references. The archived code also reproduces the main 200-row analysis as a numerical check.

A further LLM review examined the audit arithmetic, reported checks of sampled-cell identities and reference-file hashes, and spot-checked selected reference judgments. Its revised review established the reported sequence as data collection, correction, original-file audit, and corrected-file audit, and withdrew an earlier suggestion that the correction had been fitted to the audit flags. Correction and audit nevertheless both relied on web-search-assisted judgments and potentially overlapping sources. Audit results are therefore described as agreement with a source-based reference, rather than independent sensory validation. The audit reports and revised methodological review are included in the supplementary archive.

Software

Repeated-item analyses used R 4.5.1, with psych 2.5.6, polycor 0.8-2, and irr 0.84.1, as recorded in the analysis session information. The broader workflow used httr (Wickham, 2023), jsonlite (Ooms, 2014), data.table (Dowle & Srinivasan, 2025), dotenv (Csárdi, 2023), and ranger (Wright & Ziegler, 2017); graphics used ggplot2, ggrepel, and base R grid graphics. Confusion-matrix exports include vector PDFs and high-resolution raster versions. The accuracy audit used Python 3.13.5 and NumPy 2.3.5 for sampling, weighting, and conditional bootstrap calculations.

The article's supplementary archive contains the working R source, the Random200 sampling script, complete repeated-item summary tables, confusion-matrix counts, diagnostics, sample-membership records, session information, the pattern-loading workbook and machine-readable exports, and the oblique factor correlations. It also includes both Random200 rating files, original and corrected audit scripts, frozen references, sampling metadata, evidence ledgers, complete audit outputs, and the masking sensitivity script and results. The supplied correction API response records are included directly in the archive, with a per-file inventory. The R source is a sequence of working blocks, not a single-execution pipeline; its appended free-text conversation is commented. The staged workflow and manual conversion of original Unknown responses to missing values are described above. The full corpus and remaining study data will be publicly released at publication in Zenodo (<https://doi.org/10.5281/zenodo.22755429>); the article archive and that deposit have complementary scopes.

Correction Response Archive

The supplied archive contains 1,713 JSON records: 1,705 completed Grok 4.6 responses and eight HTTP 520 error records representing four error events recorded twice. The completed responses contain 8,525 recommendations, all matching Random200 fragrance names and row indices. Of these recommendation records, 2,899 have a missing current value and 5,626 have a nonmissing current value. These are counts of returned recommendations, not certified counts of distinct cells or accepted corrections. The response envelopes retain search activity, source URLs, model identifiers, and returned judgments, but not the full input prompts or explicit question-column and sweep identifiers. The prompt templates remain in the R source. These records do not independently document Famous50's complete review; its 10-sweep, all-cell coverage is recorded above from the study's run history. Selected response fields are illustrated below, and the data archive provides a full file inventory.

Factor Loadings and Item Identification

The 90 loading rows represent 60 distinct question wordings. Table S1 maps each wording to its row identifiers in the loading table. These identifiers follow the loading workbook, which is grouped by factor, and do not denote the original administration order. Repeated occurrences are preserved individually; their order within the loading workbook should not be equated with the occurrence numbering in the repeated-item pair map.

At an absolute loading threshold of .30, 81 of 90 rows had at least one salient loading, 16 had multiple salient loadings, and nine had none. At .40, the counts were 64, six, and 26. Table S2 retains all pattern coefficients. Table S14 gives the oblique factor-correlation matrix (Φ); its 91 distinct off-diagonal entries range from $-.29$ to $.35$ at the supplied two-decimal precision.

These are fitted-factor correlations, not correlations among the bounded display scores.

Repeated questions share content and do not independently confirm a factor. Fourteen factors were retained for interpretability; a later MAP check also suggested 14. The retained solution had SRMR = .04. Original Unknown responses had been converted to blanks before reference-model estimation and Famous50 scoring.

Table S1*Full Question Wording and Loading Row Identifiers*

Loading rows	Question wording
Q01, Q04	Does this fragrance have resinous notes?
Q02	Does this fragrance have incense notes?
Q03	Does this fragrance have balsamic notes?
Q05, Q06	Does this fragrance have smoky notes?
Q07	Is amber a dominant note in this fragrance?
Q08, Q88	Would many U.S. consumers describe this fragrance as strange?
Q09, Q10	Does this fragrance have animalic notes?
Q11, Q12	Does this fragrance have fecal notes?
Q13, Q14	Is musk a leading note in this fragrance?
Q15, Q16	Does this fragrance cost less than \$30 per ounce?
Q17, Q18	Does this fragrance cost more than \$100 per ounce?
Q19, Q20	Does this fragrance have significant sillage/projection?
Q21, Q23	Does this fragrance have notable longevity (8+ hours)?
Q22, Q24	Is this fragrance considered a beast mode projector?
Q25	Is this fragrance sold as a perfume extract?
Q26, Q28	Does this fragrance have hay notes?
Q27	Is oud a leading note in this fragrance?
Q29, Q30	Does this fragrance have mossy notes?
Q31	Does this fragrance have chypre notes?
Q32, Q34	Is patchouli a leading note in this fragrance?
Q33, Q58	Is petrichor a prominent note in this fragrance?
Q35, Q85	Does this fragrance have vetiver notes?
Q36, Q37	Does this fragrance have gourmand notes?
Q38	Is caramel a prominent note in this fragrance?
Q39	Does this fragrance have vanilla notes?
Q40	Is almond a leading note in this fragrance?
Q41	Is coconut a leading note in this fragrance?
Q42	Is honey a dominant note in this fragrance?
Q43, Q46	Is leather a leading note in this fragrance?
Q44, Q45, Q47	Is this fragrance marketed mostly to men?
Q48	Does this fragrance have floral notes?

Loading rows	Question wording
Q49	Does this fragrance have indolic notes?
Q50, Q51	Does this fragrance have rose notes?
Q52	Does this fragrance have tobacco notes?
Q53	Is jasmine a leading note in this fragrance?
Q54	Is tuberose a prominent note in this fragrance?
Q55, Q56	Does this fragrance have ozonic notes?
Q57	Does this fragrance have aquatic notes?
Q59	Does this fragrance have metallic notes?
Q60, Q61	Does this fragrance have cardamom notes?
Q62, Q63	Is cumin a leading note in this fragrance?
Q64, Q65	Does this fragrance have cedar notes?
Q66	Does this fragrance have sandalwood notes?
Q67, Q68	Does this fragrance have carnation notes?
Q69, Q70	Does this fragrance have aldehydic notes?
Q71	Was this fragrance formulated after 2015?
Q72	Does this fragrance have violet notes?
Q73, Q74	Is mint a prominent note in this fragrance?
Q75	Does this fragrance have camphoraceous notes?
Q76	Is lavender a dominant note in this fragrance?
Q77	Is saffron a prominent note in this fragrance?
Q78, Q79	Does this fragrance have fruity notes?
Q80	Is peach a prominent note in this fragrance?
Q81	Does this fragrance have citrus notes?
Q82	Was this fragrance formulated before 1930?
Q83	Does this fragrance have waxy notes?
Q84	Is iris a leading note in this fragrance?
Q86, Q87	Can this fragrance be found at most U.S. department stores?
Q89	Does this fragrance have soapy notes?
Q90	Is this primarily a linear fragrance?

Note. Q identifiers denote the 90 data rows of the factor-loading workbook in order. Repeated text is consolidated in this key only; all separate coefficients are retained in Table S2. Straight quotation marks and numeric duplicate suffixes were normalized for grouping. Q15 and Q16 are the reverse-keyed price-below-\$30-per-ounce administrations; wording is shown before reversal. The questionnaire used 0 = Definitely No, 1 = Probably No, 2 = Unknown or not applicable, 3 = Probably Yes, and 4 = Definitely Yes. Original code-2 entries were converted to blanks before statistical analysis.

Table S2*Complete Oblimin-Rotated Factor Pattern Matrix*

Row	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12	F13	F14
Q01	.894	-.033	.018	.050	.033	.013	-.091	-.017	.009	-.020	.031	-.070	-.051	.016
Q02	.797	-.020	.054	-.034	-.065	-.076	.007	-.053	-.003	.066	-.017	-.074	-.049	-.065
Q03	.730	-.096	-.001	.044	.013	.144	-.089	-.027	.085	.095	-.027	-.019	.097	-.016
Q04	.643	-.062	.134	.112	.027	.054	.087	-.066	.126	.153	-.083	-.063	-.110	-.098
Q05	.619	.161	.115	-.043	.018	-.068	.320	-.108	-.033	-.047	-.118	-.074	-.069	-.076
Q06	.617	.113	.062	-.033	.036	-.010	.233	-.069	-.041	-.121	-.031	-.067	.181	.048
Q07	.471	.011	-.032	.161	-.060	.078	.081	-.025	.048	.169	-.075	-.080	-.230	-.036
Q08	.411	.356	.036	-.042	.068	.057	-.185	.106	-.039	-.157	.148	.015	.232	.031
Q09	.064	.706	.053	.164	-.046	-.115	.170	-.027	-.007	.154	-.126	-.063	-.061	-.080
Q10	.068	.693	.074	.136	.010	-.064	.112	-.053	.008	.061	-.086	-.071	-.077	-.015
Q11	.016	.647	.151	.025	.060	-.031	.031	-.071	-.008	.128	.017	-.038	.030	-.116
Q12	-.017	.641	.058	.041	.083	.009	-.118	-.036	.074	.028	.018	-.081	.125	-.053
Q13	-.123	.394	-.020	.183	-.131	-.065	-.135	.121	-.086	.183	.028	-.067	-.260	.142
Q14	-.204	.342	-.069	.179	-.149	-.111	-.134	.086	-.051	.234	-.030	-.032	-.087	-.016
Q15	.027	-.013	.937	.030	-.006	-.023	-.015	-.011	.005	.002	.041	-.011	-.012	.141
Q16	.022	.002	.850	.042	-.011	.001	-.013	-.023	-.024	-.053	.030	-.044	-.041	.034
Q17	.018	.022	.846	.077	-.022	-.040	-.001	-.044	.019	.046	.013	-.015	-.003	.014
Q18	.133	.078	.519	.205	-.059	-.047	-.128	-.013	.023	-.079	.049	-.005	.291	.003
Q19	-.009	.040	.200	.645	.100	.150	.101	-.033	.069	.067	-.002	-.022	-.006	-.113
Q20	.231	.141	.097	.618	.102	.133	-.052	.037	.022	-.004	.045	.046	-.096	.050
Q21	.011	.027	.287	.617	.092	.142	.080	-.014	.058	.071	-.027	-.077	.002	-.145
Q22	.101	.181	.130	.604	.091	.124	.135	-.019	.063	-.014	-.037	-.031	-.018	-.202
Q23	.120	.068	.187	.602	.082	.122	-.002	.022	.036	.031	-.019	-.059	.013	-.035
Q24	.289	.219	-.125	.519	.022	.073	-.129	.030	-.029	-.169	.169	.041	.463	.156
Q25	.030	-.033	.334	.393	-.050	.083	-.030	.014	-.044	-.025	.037	-.003	.136	-.226
Q26	.058	.231	.308	-.338	.215	.177	.101	.008	-.042	.016	.181	-.007	-.002	-.161
Q27	.240	.250	.016	.322	-.139	-.266	.054	-.174	.021	-.225	-.144	-.108	-.060	-.218
Q28	.135	.249	.227	-.293	.208	.142	-.047	.060	-.046	-.068	.240	-.037	.122	-.052
Q29	-.029	.034	-.034	.035	.768	-.022	.085	.063	-.011	.184	.070	.030	-.052	-.047
Q30	-.026	.032	-.048	.034	.754	-.024	.043	.058	-.020	.159	.087	.011	-.045	-.033

Note. Bold indicates $|\text{loading}| \geq .30$, evaluated before rounding. F1 = Resinous/Smoky; F2 = Animalic; F3 = Expensive; F4 = Intense/Powerful; F5 = Earthy; F6 = Gourmand/Sweet; F7 = Masculine; F8 = Aquatic/Ozonic; F9 = Spicy/Woody; F10 = Carnation/Aldehydic; F11 = Minty; F12 = Fruity; F13 = Vintage/Waxy; F14 = US Mainstream. Pattern coefficients are reproduced in the keyed orientation of the fitted model; original question wording is in Table S1.

Table S2 (continued)*Complete Oblimin-Rotated Factor Pattern Matrix*

Row	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12	F13	F14
Q31	-.014	.035	-.007	.034	.630	-.092	-.129	-.123	-.029	.178	.007	.069	.238	.130
Q32	.045	-.126	-.006	.278	.578	-.067	.011	-.194	.083	-.119	-.127	.004	-.089	.118
Q33	.015	-.036	.162	-.157	.552	-.073	.029	.228	.005	-.096	-.054	-.150	.180	-.131
Q34	.086	-.031	-.073	.281	.533	-.060	-.090	-.202	.057	-.210	-.143	-.032	-.086	.128
Q35	.037	-.181	.170	.050	.295	-.180	.213	.012	.225	.110	.053	.080	.114	.013
Q36	-.021	-.036	.021	.084	-.036	.820	.067	-.140	.023	.051	-.108	.126	.048	-.046
Q37	.069	-.024	-.086	.096	-.076	.767	-.121	-.098	-.074	-.167	.031	.085	.011	.091
Q38	.034	-.030	-.037	.053	.004	.688	.044	-.055	-.038	-.027	-.073	-.029	-.142	-.026
Q39	.033	-.072	-.012	.138	-.049	.659	-.034	-.137	.047	.158	-.074	-.024	-.167	.061
Q40	-.052	-.073	.111	.008	-.056	.391	.005	-.093	.032	.088	-.055	-.036	.109	.017
Q41	-.122	-.020	-.011	.011	-.094	.360	-.009	.173	.027	-.002	-.145	.035	.299	.013
Q42	.066	.154	.133	-.021	.039	.312	-.007	-.028	-.078	.140	-.021	-.052	.049	.026
Q43	.114	.443	.054	.073	.058	-.126	.554	-.096	.019	.020	-.166	.052	.015	.018
Q44	.062	.018	-.098	.182	.077	-.109	.552	.069	.306	-.043	.311	-.101	-.030	.025
Q45	.091	.022	-.093	.174	.092	-.120	.532	.081	.308	-.068	.318	-.120	-.045	.015
Q46	.136	.474	.050	.032	.064	-.104	.515	-.116	.019	-.002	-.176	.009	.044	.069
Q47	.055	.028	-.161	.162	.110	-.079	.503	.071	.240	-.026	.335	-.110	-.162	.075
Q48	-.142	.041	.092	.081	.016	-.136	-.497	-.083	-.089	.270	-.180	.138	-.245	.057
Q49	.156	.270	-.016	.152	.025	-.004	-.421	.031	-.029	.078	.111	.032	.399	.187
Q50	.053	.031	.072	.184	.092	-.282	-.376	-.233	-.059	.201	-.206	.069	-.137	.002
Q51	.047	.024	.070	.193	.109	-.267	-.360	-.220	-.064	.251	-.219	.088	-.092	-.028
Q52	.200	.123	.066	-.036	.049	.174	.352	-.069	.029	-.021	-.032	-.067	.188	-.056
Q53	-.170	.017	.099	.198	.025	-.060	-.308	-.093	-.054	.189	-.166	.076	-.012	.017
Q54	-.152	.156	.138	.081	-.019	.060	-.222	.023	-.036	.131	-.117	-.007	.157	.011
Q55	-.027	-.009	-.027	.028	.022	-.078	.008	.895	.007	.049	-.014	.026	-.046	.005
Q56	-.049	-.014	-.027	.074	-.027	-.068	.030	.881	-.010	.029	-.056	.008	.022	.025
Q57	-.032	-.036	-.044	.056	-.024	-.060	-.011	.859	-.000	-.011	-.026	.054	-.107	.055
Q58	.089	.048	.194	-.212	.454	-.044	-.103	.483	-.041	-.200	-.056	-.132	-.034	-.177
Q59	.135	.145	.089	-.004	-.004	-.042	.030	.237	-.027	.001	.045	-.076	.234	.090
Q60	-.010	-.041	-.010	.020	-.052	.001	.023	-.003	.806	-.007	-.031	-.027	.010	.029

Note. Bold indicates $|\text{loading}| \geq .30$, evaluated before rounding. F1 = Resinous/Smoky; F2 = Animalic; F3 = Expensive; F4 = Intense/Powerful; F5 = Earthy; F6 = Gourmand/Sweet; F7 = Masculine; F8 = Aquatic/Ozonic; F9 = Spicy/Woody; F10 = Carnation/Aldehydic; F11 = Minty; F12 = Fruity; F13 = Vintage/Waxy; F14 = US Mainstream. Pattern coefficients are reproduced in the keyed orientation of the fitted model; original question wording is in Table S1.

Table S2 (continued)*Complete Oblimin-Rotated Factor Pattern Matrix*

Row	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12	F13	F14
Q61	-.017	-.039	-.019	.015	-.054	.027	.034	.005	.761	-.024	-.055	-.063	.108	.012
Q62	-.032	.411	.027	-.102	.074	.052	-.224	-.077	.568	-.159	.020	-.065	-.038	-.029
Q63	.024	.447	.052	-.144	.120	.054	-.151	-.095	.517	-.016	.071	-.057	-.105	-.022
Q64	.207	-.168	.087	-.001	-.036	-.172	.054	.064	.437	.217	.101	.188	-.038	-.041
Q65	.240	-.159	.027	.004	.000	-.178	-.013	.066	.434	.203	.074	.190	-.053	-.025
Q66	.068	-.079	-.005	.132	-.088	-.079	-.141	.026	.278	.218	-.032	.074	-.000	-.095
Q67	.049	.113	-.046	-.005	.222	.031	.001	-.097	-.019	.671	.055	-.039	.029	.006
Q68	.070	.071	-.091	-.008	.168	-.011	-.018	-.053	.041	.628	.012	.005	.117	-.020
Q69	.041	.090	.106	-.045	.038	-.032	-.040	.171	-.089	.600	.025	-.053	.101	.081
Q70	.134	.119	.056	.013	.014	-.019	-.043	.110	-.015	.568	-.000	-.065	.113	.116
Q71	.009	-.068	.059	.247	-.035	.116	.105	.091	-.034	-.275	-.093	.002	-.233	-.120
Q72	-.056	.001	.044	.001	.010	-.085	-.092	.165	.079	.271	-.123	.058	.003	-.019
Q73	-.067	-.041	.087	.020	-.009	-.082	.037	-.044	.004	.041	.817	.030	-.097	-.056
Q74	-.044	-.054	.046	.001	-.027	-.068	-.033	-.041	-.011	.007	.799	-.007	.007	-.031
Q75	.097	.020	.134	-.088	.114	-.056	.016	-.039	-.044	.029	.474	-.100	-.050	-.112
Q76	-.137	-.089	-.125	.081	.136	-.030	.217	-.119	.143	.113	.363	-.194	-.033	.090
Q77	.204	.188	-.023	.229	-.073	-.131	.020	-.112	.167	-.160	-.235	-.056	-.060	-.143
Q78	-.052	-.007	-.011	.004	.023	-.009	.077	.012	.013	-.050	.014	.919	.078	-.017
Q79	.019	.050	.001	-.024	.031	.094	.017	.038	-.023	-.032	.006	.847	-.201	.015
Q80	-.068	.062	.002	-.055	-.020	.100	-.071	-.004	-.044	.061	-.121	.581	.105	-.090
Q81	-.047	-.081	-.024	.043	.019	-.206	-.038	.076	.072	.021	.284	.428	.103	.140
Q82	-.087	-.100	.005	-.007	.018	-.069	.062	-.145	.010	.113	-.081	-.101	.881	-.081
Q83	.003	.144	.163	-.136	.008	.221	.012	.030	.041	.196	-.074	-.076	.455	.042
Q84	-.096	-.041	.177	-.033	.005	-.004	.027	-.012	.096	.232	-.188	-.111	.326	.042
Q85	.017	-.169	.126	.033	.265	-.178	.197	.011	.240	.085	.020	.042	.321	.013
Q86	-.004	-.007	.306	-.014	.014	.025	.066	.031	.021	-.015	-.045	-.009	-.071	.851
Q87	-.036	-.020	-.033	-.132	.013	.028	.054	.032	.003	.039	-.071	.024	-.017	.809
Q88	.266	.322	.195	-.075	.106	-.005	-.018	.078	.037	-.040	-.009	-.111	.004	-.393
Q89	.042	.129	-.047	-.034	.008	-.066	-.208	.111	-.062	.248	.202	-.103	.198	.340
Q90	.074	.035	.014	-.148	-.095	-.042	-.071	.054	-.063	-.120	.164	-.112	.178	.192

Note. Bold indicates $|\text{loading}| \geq .30$, evaluated before rounding. F1 = Resinous/Smoky; F2 = Animalic; F3 = Expensive; F4 = Intense/Powerful; F5 = Earthy; F6 = Gourmand/Sweet; F7 = Masculine; F8 = Aquatic/Ozonic; F9 = Spicy/Woody; F10 = Carnation/Aldehydic; F11 = Minty; F12 = Fruity; F13 = Vintage/Waxy; F14 = US Mainstream. Pattern coefficients are reproduced in the keyed orientation of the fitted model; original question wording is in Table S1.

Response Availability and Repeated Item Agreement

Random200 included 18,000 potential ratings and Famous50 included 4,500. Substantive response availability increased in both cohorts (Table S3). Missing-to-observed transitions indicate increased coverage, not independent evidence of correctness. All pairwise comparisons use the same fragrances within the chosen cohort; matched samples hold the individual pair sample fixed across versions, whereas available samples use every complete pair within a version.

Table S3

Response Availability and Changes in the Two Evaluation Cohorts

Characteristic	Random200	Famous50
Fragrances	200	50
Rating columns	90	90
Substantive ratings before correction	15,101 (83.9%)	3,121 (69.4%)
Substantive ratings after correction	17,229 (95.7%)	4,422 (98.3%)
Missing ratings before correction	2,899	1,379
Missing ratings after correction	771	78
Previously missing ratings filled	2,194	1,310
Observed ratings removed	66	9
Observed ratings changed	2,033	633

Note. Availability percentages use all potential rating cells within the cohort. Missing ratings are those excluded by the substantive-response coding used in the reliability outputs. Changed observed ratings are observed in both versions but have different values; filling and removal are separate categories.

Table S4*Median Agreement Across Repeated Question Pairs*

Cohort and sample	Statistic	Pairs	Original	Corrected
Random200 matched	Polychoric r	30	.790	.944
Random200 matched	ICC(A,1)	30	.580	.869
Random200 matched	Exact agreement (%)	31	80.8	85.4
Random200 matched	Mean absolute difference	31	.461	.187
Random200 available	Polychoric r	30	.780	.937
Random200 available	ICC(A,1)	31	.560	.843
Random200 available	Exact agreement (%)	31	80.4	75.5
Random200 available	Mean absolute difference	31	.465	.273
Famous50 matched	Polychoric r	29	.893	.991
Famous50 matched	ICC(A,1)	30	.663	.914
Famous50 matched	Exact agreement (%)	31	78.0	85.0
Famous50 matched	Mean absolute difference	31	.565	.167
Famous50 available	Polychoric r	29	.893	.970
Famous50 available	ICC(A,1)	31	.656	.885
Famous50 available	Exact agreement (%)	31	76.3	77.1
Famous50 available	Mean absolute difference	31	.625	.271

Note. Each row summarizes the same set of pairs with an estimable statistic in both versions. Medians are unweighted across pairs. Exact agreement and mean absolute difference include all 31 pairs; coefficient rows can exclude constant-variable pairs. A difference between medians is not the median paired change. Three pairs concern the same triplicated marketing question and are statistically dependent.

The median matched polarity agreement increased from 88.7% to 97.9% in Random200 and from 88.2% to 100.0% in Famous50. These are pair-level medians, not pooled percentages.

Table S5*Sample Sizes and Estimation Diagnostics*

Cohort and sample	Version	Pair n range (Mdn)	Sparse ^a	Boundary ^b	Estimable r / ICC
Random200 matched	Original	74-163 (140)	28	0	31 / 31
Random200 matched	Corrected	74-163 (140)	22	3	30 / 30
Random200 available	Original	75-164 (140)	28	0	31 / 31
Random200 available	Corrected	110-200 (190)	15	1	30 / 31
Famous50 matched	Original	16-41 (24)	31	3	31 / 31
Famous50 matched	Corrected	16-41 (24)	26	10	29 / 30
Famous50 available	Original	16-41 (24)	31	3	31 / 31
Famous50 available	Corrected	42-50 (49)	19	5	29 / 31

Note. There are 31 pairs per row. ^a At least one occupied marginal category has fewer than five observations. ^b Polychoric estimates flagged near the numerical boundary. Detailed warnings, ICC confidence limits, and signed differences are in the archive. Constant ratings can yield perfect exact agreement while a correlation is undefined; undefined values were not converted to one.

For matched Random200 data, corrected petrichor-prominence ratings were constant and identical, producing 100% exact agreement but undefined polychoric and ICC estimates. The original petrichor pair had a negative polychoric estimate ($-.886$) despite 92.1% exact agreement, illustrating the instability and limited information in very sparse, highly imbalanced tables. Famous50 was especially sparse: every original matched pair had a sparse marginal category, and 10 corrected matched pairs had near-boundary polychoric estimates. Strong coefficients in these tables should therefore be read with their sample sizes and diagnostics.

Table S6*Repeated-Item Agreement in Random200 Matched Complete Cases*

Pair	Item	n	Exact % (O / C)	Polychoric r (O / C)	ICC(A,1) (O / C)	MAD (O / C)
P01	Department-store availability	156	60.3 / 96.2	.617 / .946	.324 / .777	.724 / .045
P02	Fruity notes	150	62.0 / 80.0	.810 / .978	.718 / .954	.647 / .227
P03	Rose notes	151	84.8 / 88.7	.968 / .990	.878 / .972	.291 / .132
P04	Animalic notes	111	64.9 / 73.0	.799 / .937	.665 / .889	.640 / .324
P05	Leather prominence	150	87.3 / 92.0	.904 / .977	.769 / .891	.260 / .127
P06	Cardamom notes	150	93.3 / 82.7	.969 / .902	.850 / .924	.160 / .187
P07	Marketed to men (1 vs. 2)	110	84.5 / 94.5	.949 / .996	.869 / .967	.291 / .082
P08	Marketed to men (1 vs. 3)	123	79.7 / 98.4	.931 / >.999	.861 / .997	.350 / .016
P09	Marketed to men (2 vs. 3)	107	86.9 / 94.4	.983 / .997	.945 / .978	.178 / .075
P10	Petrichor prominence	140	92.1 / 100.0	-.886 / —	-.012 / —	.136 / .000
P11	Musk prominence	107	57.0 / 62.6	.765 / .840	.637 / .617	.785 / .626
P12	Fecal notes	151	90.1 / 88.7	.488 / .651	.134 / .320	.132 / .146
P13	Price < \$30/oz	130	67.7 / 85.4	.833 / .992	.657 / .963	.738 / .169
P14	Vetiver notes	151	80.8 / 78.8	.793 / .918	.578 / .916	.503 / .245
P15	Mossy notes	136	86.0 / 88.2	.852 / .947	.700 / .879	.294 / .176
P16	Ozonic notes	130	86.9 / 90.8	.761 / .977	.607 / .947	.285 / .092
P17	Beast-mode projection	94	45.7 / 46.8	.660 / .830	.317 / .465	1.181 / .957
P18	Gourmand notes	144	41.0 / 68.1	.787 / .927	.376 / .794	1.410 / .562
P19	Carnation notes	141	92.9 / 75.9	.674 / .785	.315 / .661	.241 / .262
P20	Price > \$100/oz	119	73.9 / 94.1	.710 / >.999	.426 / .943	.706 / .084
P21	Resinous notes	134	56.7 / 70.9	.794 / .943	.439 / .817	1.201 / .515
P22	Smoky notes	137	76.6 / 86.9	.770 / .981	.582 / .924	.562 / .182
P23	Longevity ≥ 8 hr	100	50.0 / 56.0	.590 / .696	.550 / .646	.820 / .680
P24	Sillage/projection	74	43.2 / 43.2	.493 / .577	.369 / .427	1.041 / .973
P25	Cedar notes	141	83.7 / 83.0	.900 / .974	.726 / .954	.461 / .199
P26	Mint prominence	163	95.7 / 96.9	.850 / .987	.521 / .832	.135 / .055
P27	Strangeness	112	28.6 / 33.9	.391 / .845	.143 / .360	1.366 / 1.036
P28	Patchouli prominence	158	77.2 / 86.7	.826 / .941	.654 / .859	.544 / .196
P29	Aldehydic notes	141	90.8 / 70.2	.662 / .624	.365 / .693	.312 / .348
P30	Cumin prominence	145	95.2 / 97.9	.559 / >.999	.113 / .410	.103 / .048
P31	Hay notes	157	97.5 / 80.3	-.853 / -.155	-.009 / -.047	.038 / .197

Note. O = original; C = corrected. MAD = mean absolute difference on the 0, 1, 3, 4 response metric. A dash indicates an undefined coefficient. Values above .999 are displayed as >.999 rather than rounded to perfect correlation. Pair identifiers and occurrence numbers correspond to the original instrument positions in pair_map.csv; they are distinct from the Q identifiers of the loading workbook. Full ICC confidence intervals, signed differences, polarity agreement, and diagnostics are in agreement_statistics.csv.

Table S7*Repeated-Item Agreement in Famous50 Matched Complete Cases*

Pair	Item	n	Exact % (O / C)	Polychoric r (O / C)	ICC(A,1) (O / C)	MAD (O / C)
P01	Department-store availability	38	76.3 / 78.9	.436 / .932	.089 / .811	.632 / .342
P02	Fruity notes	18	55.6 / 77.8	.780 / .969	.668 / .965	.667 / .222
P03	Rose notes	38	81.6 / 92.1	.949 / .992	.831 / .950	.421 / .158
P04	Animalic notes	16	68.8 / 75.0	.821 / .991	.675 / .914	.625 / .312
P05	Leather prominence	24	83.3 / 95.8	.952 / .998	.796 / .980	.292 / .042
P06	Cardamom notes	24	91.7 / 91.7	.921 / >.999	.771 / .866	.208 / .167
P07	Marketed to men (1 vs. 2)	35	85.7 / 100.0	.940 / >.999	.810 / 1.000	.400 / .000
P08	Marketed to men (1 vs. 3)	40	95.0 / 100.0	.984 / >.999	.891 / 1.000	.200 / .000
P09	Marketed to men (2 vs. 3)	38	86.8 / 100.0	.973 / >.999	.927 / 1.000	.211 / .000
P10	Petrichor prominence	36	88.9 / 94.4	.999 / —	.389 / .000	.306 / .056
P11	Musk prominence	20	30.0 / 70.0	.347 / .923	.221 / .357	1.500 / .500
P12	Fecal notes	20	85.0 / 75.0	-.942 / .796	-.075 / .333	.150 / .450
P13	Price < \$30/oz	29	79.3 / 93.1	.912 / .999	.768 / .989	.552 / .069
P14	Vetiver notes	23	78.3 / 78.3	.892 / .977	.781 / .965	.478 / .217
P15	Mossy notes	17	64.7 / 64.7	.831 / .931	.748 / .943	.647 / .353
P16	Ozonic notes	19	100.0 / 94.7	>.999 / >.999	1.000 / .940	.000 / .105
P17	Beast-mode projection	18	33.3 / 38.9	.974 / .980	.349 / .891	1.667 / .611
P18	Gourmand notes	25	56.0 / 76.0	.889 / .963	.603 / .873	.920 / .400
P19	Carnation notes	23	78.3 / 82.6	.940 / .965	.685 / .853	.565 / .348
P20	Price > \$100/oz	41	78.0 / 95.1	.831 / >.999	.761 / .964	.293 / .049
P21	Resinous notes	19	26.3 / 57.9	.997 / .942	.221 / .745	2.000 / .737
P22	Smoky notes	20	75.0 / 90.0	.452 / .990	.236 / .982	.700 / .100
P23	Longevity ≥ 8 hr	24	54.2 / 91.7	.512 / .999	.389 / .913	.667 / .125
P24	Sillage/projection	20	45.0 / 85.0	.792 / .999	.261 / .913	.750 / .150
P25	Cedar notes	25	68.0 / 72.0	.893 / .946	.759 / .792	.640 / .520
P26	Mint prominence	27	96.3 / 100.0	.999 / >.999	.658 / 1.000	.148 / .000
P27	Strangeness	18	61.1 / 66.7	.822 / .994	.493 / .534	.778 / .556
P28	Patchouli prominence	26	50.0 / 69.2	.978 / .984	.559 / .893	1.077 / .423
P29	Aldehydic notes	20	75.0 / 65.0	.626 / .801	.421 / .885	.800 / .350
P30	Cumin prominence	22	95.5 / 100.0	.999 / >.999	.656 / 1.000	.182 / .000
P31	Hay notes	27	88.9 / 100.0	-.943 / —	-.044 / —	.185 / .000

Note. O = original; C = corrected. MAD = mean absolute difference on the 0, 1, 3, 4 response metric. A dash indicates an undefined coefficient. Values above .999 are displayed as >.999 rather than rounded to perfect correlation. Pair identifiers and occurrence numbers correspond to the original instrument positions in pair_map.csv; they are distinct from the Q identifiers of the loading workbook. Full ICC confidence intervals, signed differences, polarity agreement, and diagnostics are in agreement_statistics.csv.

Table S8*Repeated-Item Agreement in Random200 All Available Complete Cases*

Pair	Item	n (O / C)	Exact % (O / C)	Polychoric r (O / C)	ICC(A,1) (O / C)	MAD (O / C)
P01	Department-store availability	156 / 200	60.3 / 96.5	.617 / .965	.324 / .812	.724 / .045
P02	Fruity notes	150 / 195	62.0 / 75.4	.810 / .965	.718 / .945	.647 / .277
P03	Rose notes	152 / 190	84.9 / 80.5	.969 / .967	.880 / .961	.289 / .211
P04	Animalic notes	112 / 178	64.3 / 62.4	.792 / .886	.654 / .838	.661 / .472
P05	Leather prominence	151 / 197	86.8 / 89.3	.894 / .973	.741 / .887	.285 / .152
P06	Cardamom notes	151 / 192	93.4 / 75.5	.969 / .837	.850 / .877	.159 / .271
P07	Marketed to men (1 vs. 2)	112 / 191	83.9 / 93.2	.941 / .995	.850 / .967	.321 / .089
P08	Marketed to men (1 vs. 3)	124 / 195	79.0 / 94.4	.921 / .997	.841 / .970	.379 / .082
P09	Marketed to men (2 vs. 3)	108 / 193	87.0 / 93.3	.983 / .996	.946 / .966	.176 / .093
P10	Petrichor prominence	140 / 199	92.1 / 99.5	-.886 / —	-.012 / .000	.136 / .005
P11	Musk prominence	107 / 180	57.0 / 56.1	.765 / .806	.637 / .611	.785 / .694
P12	Fecal notes	151 / 189	90.1 / 85.7	.488 / .715	.134 / .538	.132 / .169
P13	Price < \$30/oz	132 / 180	67.4 / 85.6	.834 / .993	.655 / .969	.750 / .161
P14	Vetiver notes	153 / 186	80.4 / 71.5	.772 / .850	.560 / .895	.523 / .312
P15	Mossy notes	136 / 191	86.0 / 75.4	.852 / .857	.700 / .843	.294 / .293
P16	Ozonic notes	130 / 185	86.9 / 84.3	.761 / .932	.607 / .915	.285 / .157
P17	Beast-mode projection	99 / 139	44.4 / 49.6	.635 / .734	.290 / .511	1.253 / .820
P18	Gourmand notes	144 / 192	41.0 / 68.8	.787 / .927	.376 / .801	1.410 / .547
P19	Carnation notes	142 / 194	93.0 / 62.4	.675 / .412	.315 / .652	.239 / .392
P20	Price > \$100/oz	122 / 183	72.1 / 91.8	.690 / .983	.392 / .930	.770 / .098
P21	Resinous notes	135 / 183	57.0 / 67.2	.796 / .947	.441 / .841	1.193 / .497
P22	Smoky notes	137 / 187	76.6 / 78.1	.770 / .967	.582 / .908	.562 / .273
P23	Longevity ≥ 8 hr	101 / 123	50.5 / 58.5	.599 / .720	.556 / .675	.812 / .626
P24	Sillage/projection	75 / 110	42.7 / 58.2	.491 / .670	.362 / .561	1.053 / .709
P25	Cedar notes	142 / 184	83.1 / 73.9	.899 / .941	.729 / .934	.465 / .293
P26	Mint prominence	164 / 195	95.7 / 96.4	.851 / .991	.521 / .875	.134 / .056
P27	Strangeness	113 / 166	28.3 / 37.3	.392 / .766	.143 / .387	1.363 / .982
P28	Patchouli prominence	158 / 195	77.2 / 86.7	.826 / .952	.654 / .870	.544 / .190
P29	Aldehydic notes	142 / 187	90.1 / 62.0	.629 / .534	.339 / .641	.338 / .422
P30	Cumin prominence	145 / 198	95.2 / 98.5	.559 / >.999	.113 / .719	.103 / .035
P31	Hay notes	158 / 191	97.5 / 68.6	-.849 / -.303	-.009 / -.105	.038 / .314

Note. O = original; C = corrected. MAD = mean absolute difference on the 0, 1, 3, 4 response metric. A dash indicates an undefined coefficient. Values above .999 are displayed as >.999 rather than rounded to perfect correlation. Pair identifiers and occurrence numbers correspond to the original instrument positions in pair_map.csv; they are distinct from the Q identifiers of the loading workbook. Full ICC confidence intervals, signed differences, polarity agreement, and diagnostics are in agreement_statistics.csv.

Table S9*Repeated-Item Agreement in Famous50 All Available Complete Cases*

Pair	Item	n (O / C)	Exact % (O / C)	Polychoric r (O / C)	ICC(A,1) (O / C)	MAD (O / C)
P01	Department-store availability	38 / 50	76.3 / 78.0	.436 / .953	.089 / .857	.632 / .340
P02	Fruity notes	19 / 48	52.6 / 72.9	.653 / .933	.519 / .897	.842 / .354
P03	Rose notes	38 / 50	81.6 / 84.0	.949 / .986	.831 / .946	.421 / .220
P04	Animalic notes	16 / 47	68.8 / 55.3	.821 / .937	.675 / .760	.625 / .638
P05	Leather prominence	24 / 50	83.3 / 96.0	.952 / .998	.796 / .979	.292 / .040
P06	Cardamom notes	24 / 48	91.7 / 77.1	.921 / .866	.771 / .779	.208 / .271
P07	Marketed to men (1 vs. 2)	35 / 50	85.7 / 100.0	.940 / >.999	.810 / 1.000	.400 / .000
P08	Marketed to men (1 vs. 3)	40 / 50	95.0 / 100.0	.984 / >.999	.891 / 1.000	.200 / .000
P09	Marketed to men (2 vs. 3)	38 / 50	86.8 / 100.0	.973 / >.999	.927 / 1.000	.211 / .000
P10	Petrichor prominence	36 / 50	88.9 / 94.0	.999 / —	.389 / .000	.306 / .060
P11	Musk prominence	20 / 49	30.0 / 69.4	.347 / .613	.221 / .226	1.500 / .490
P12	Fecal notes	21 / 47	85.7 / 63.8	-.949 / .682	-.071 / .620	.143 / .489
P13	Price < \$30/oz	29 / 49	79.3 / 87.8	.912 / .972	.768 / .932	.552 / .184
P14	Vetiver notes	24 / 46	75.0 / 76.1	.817 / .972	.696 / .963	.625 / .239
P15	Mossy notes	17 / 48	64.7 / 68.8	.831 / .924	.748 / .867	.647 / .458
P16	Ozonic notes	19 / 49	100.0 / 71.4	>.999 / .993	1.000 / .852	.000 / .367
P17	Beast-mode projection	18 / 46	33.3 / 52.2	.974 / .926	.349 / .873	1.667 / .522
P18	Gourmand notes	25 / 48	56.0 / 77.1	.889 / .961	.603 / .844	.920 / .417
P19	Carnation notes	23 / 48	78.3 / 77.1	.940 / .976	.685 / .908	.565 / .312
P20	Price > \$100/oz	41 / 50	78.0 / 96.0	.831 / >.999	.761 / .977	.293 / .040
P21	Resinous notes	19 / 48	26.3 / 58.3	.997 / .853	.221 / .694	2.000 / .750
P22	Smoky notes	21 / 49	76.2 / 77.6	.468 / .972	.242 / .939	.667 / .245
P23	Longevity ≥ 8 hr	24 / 44	54.2 / 86.4	.512 / .981	.389 / .889	.667 / .205
P24	Sillage/projection	20 / 42	45.0 / 83.3	.792 / .970	.261 / .947	.750 / .167
P25	Cedar notes	25 / 48	68.0 / 75.0	.893 / .891	.759 / .851	.640 / .375
P26	Mint prominence	27 / 50	96.3 / 96.0	.999 / .991	.658 / .986	.148 / .040
P27	Strangeness	18 / 50	61.1 / 64.0	.822 / .928	.493 / .597	.778 / .600
P28	Patchouli prominence	26 / 50	50.0 / 72.0	.978 / .963	.559 / .885	1.077 / .360
P29	Aldehydic notes	20 / 48	75.0 / 60.4	.626 / .713	.421 / .799	.800 / .479
P30	Cumin prominence	22 / 50	95.5 / 100.0	.999 / >.999	.656 / 1.000	.182 / .000
P31	Hay notes	27 / 50	88.9 / 90.0	-.943 / —	-.044 / .000	.185 / .100

Note. O = original; C = corrected. MAD = mean absolute difference on the 0, 1, 3, 4 response metric. A dash indicates an undefined coefficient. Values above .999 are displayed as >.999 rather than rounded to perfect correlation. Pair identifiers and occurrence numbers correspond to the original instrument positions in pair_map.csv; they are distinct from the Q identifiers of the loading workbook. Full ICC confidence intervals, signed differences, polarity agreement, and diagnostics are in agreement_statistics.csv.

Detailed Accuracy Audit Results

The audit selected six originally entered cells from each of the 200 Random200 fragrances, retaining duplicate columns as separate sampling units. Inverse-probability weights summed to 15,101 originally entered ratings before restrictions. A reference judgment was available for 433 of 1,200 targets; most judgments were provisional, particularly inferences of absent notes. Five unresolved targets and one previously assessable target became missing after correction. The paired analysis therefore contains 432 cells, with a further 762 answered targets remaining unresolved. Eighty-eight fragrances had no assessable reference judgment.

The full-table Random200 counts in the audit agree with the reliability input-quality summary. Across the original 18,000 potential cells, 13,002 entered ratings were unchanged, 1,023 switched No/Yes direction, 1,010 changed category within direction, 2,194 blanks were filled, 66 responses became blank, and 705 remained blank. Thus, fewer blanks and improved agreement on the fixed benchmark represent different outcomes. The 2,194 filled cells made up 12.7% of corrected entered ratings and were not sampled by the original audit.

Table S10

Audit Reference Coverage and Paired Outcome Counts

Code	Meaning	Frozen cells	Scored in corrected file
Y	Direct/strong affirmative reference	52	52
N	Direct/strong negative reference	18	18
y	Provisional affirmative inference	20	20
n	Provisional negative inference	343	342
U	Reference unresolved	767	0

Note. Y/N = direct affirmative/negative evidence; y/n = provisional affirmative/negative inference; U = unresolved reference. The letter n in the code column denotes a provisional negative reference, not a sample size. The one excluded provisional-negative reference had a response that became missing.

Among the original 433 assessable cells, 388 agreed in both versions, 30 changed from disagreement to agreement, 13 disagreed in both, one changed from agreement to disagreement, and one prior disagreement became missing. The last case is excluded from both sides of the 432-cell paired comparison and is not credited as a correction.

Table S10 (continued)

Raw Directional Confusion Counts on the 432 Paired Audit Cells

Reference direction	Original No	Original Yes	Corrected No	Corrected Yes
No	338	22	353	7
Yes	21	51	7	65
Total	359	73	360	72

Note. Counts are unweighted. Original agreement = $(338 + 51)/432 = 90.0\%$; corrected agreement = $(353 + 65)/432 = 96.8\%$. The corresponding design-weighted estimates are 90.7% and 96.9%. Both apparent false negatives and apparent false positives decreased. References were frozen across versions.

Table S11*Audit Agreement and Evidence Coverage by Question Family*

Question family	Sampled / paired	Coverage (%)	Original (%)	Corrected (%)	Gain (pp)
Note / accord presence	460 / 222	49.1%	88.7%	96.6%	+8.0
Note prominence	346 / 157	46.2%	93.2%	98.4%	+5.1
Marketing audience	43 / 24	54.7%	95.1%	100.0%	+4.9
Date	13 / 8	63.0%	100.0%	100.0%	+0.0
Commercial / concentration	87 / 9	10.8%	100.0%	100.0%	+0.0
Performance / evolution	69 / 8	12.1%	60.9%	60.9%	+0.0
Subtle odor descriptors	156 / 4	2.4%	100.0%	100.0%	+0.0
Consumer strangeness	26 / 0	0.0%	Not estimable	Not estimable	—

Note. Coverage is the weighted fraction of the family's original audit targets currently assessable. Agreement retains original weights and uses the same paired cells within each family. Gains use unrounded estimates; a rounded difference can differ by 0.1 point. Commercial/concentration pools price, availability, and product-form questions. Small assessed samples restrict generalization. Overall weighted coverage of 36.8% uses the still-answered targets as its denominator.

Note presence and prominence together contributed 379 of the 432 paired judgments and 28 of the 30 resolved disagreements. Two additional resolutions concerned repeated marketing-audience questions for the same fragrance, Lonesome Rider, and should not be treated as independent fragrance-level confirmations. The initial evidence-retrieval stratum contained 368 paired judgments (90.1% to 96.6% agreement), and the targeted-retry stratum contained 64 (94.0% to 98.7%). These are retrieval strata rather than objective measures of fragrance obscurity.

Table S12*Paired Audit Agreement by Original Response Category*

Original response group	Paired n	Original agreement	Corrected agreement	Gain (pp)
0	347	95.4%	98.8%	+3.4
1	12	66.1%	83.6%	+17.5
3	24	42.7%	77.2%	+34.5
4	49	84.3%	95.6%	+11.3

Note. Cells remain grouped by their original response, so each row compares the same cells across versions. Percentages are design-weighted. Agreement concerns No/Yes direction, not whether the probably or definitely confidence category is correct.

When cells were instead grouped by their corrected response, agreement was 98.5% for code 0 (334 cells), 93.5% for code 1 (26), 67.7% for code 3 (15), and 96.9% for code 4 (57). These groups differ from the original-response strata and are not paired category-level effects. Response-category standardization produced the audit's working extrapolations of 87.6% for the original file and 95.6% for the corrected file. These figures require unverified assumptions about unassessed and newly filled cells and are not presented as validated full-file accuracy.

Audit Sensitivity and Source Uncertainty

The audit's scenario analyses quantify the dependence of broader conclusions on unknown outcomes. Let u denote accuracy among unresolved retained cells and f accuracy among newly filled cells. Conditional on accepting the frozen reference labels, the audit used retained-cell accuracy = $0.35701 + 0.63168 \times u$ and full corrected-file accuracy = $0.87266 \times$ retained-cell accuracy + $0.12734 \times f$. The first-stage quantities are design-weighted estimates; the second-stage proportions reflect 15,035 retained entered cells and 2,194 newly filled cells. Neither u nor f was measured.

Table S13

Hypothetical Sensitivity to Unassessed Outcomes and Reference Error

Assumed accuracy for both u and f	Implied retained-domain accuracy	Implied full corrected-file accuracy
50.0%	67.3%	65.1%
60.0%	73.6%	71.9%
70.0%	79.9%	78.7%
80.0%	86.2%	85.4%
90.0%	92.6%	92.2%

Note. The same assumed accuracy is assigned to u and f in each row. These are hypothetical scenarios, not estimates or confidence bounds. With both quantities unrestricted from 0 to 1, the conditional plug-in range is 31.2%–99.0%.

Table S13 (continued)

Sensitivity to Incorrect Provisional Negative References

Provisional n references assumed wrong	Original paired agreement	Corrected agreement
0.0%	90.7%	96.9%
5.0%	87.2%	93.1%
10.0%	83.6%	89.3%
20.0%	76.6%	81.6%

Note. A uniformly selected fraction of provisional negative references is assumed to be wrong; both apparent agreements lost and apparent errors removed are counted. This stress test is hypothetical and does not estimate the actual rate of reference error.

Verification confirmed identical frozen references and sampled identities across audits. The revised review states that correction preceded both audits, superseding earlier chronology caveats in the archived audit materials. An initial preview of 11 fragrance rows limited masking. It also identified source dependence: the corrector and auditor both used web-search-assisted judgments, often drawing on the same fragrance databases. The review questioned one direct-tier fruity reference for *Dancing With Strangers* and several provisional judgments. The paired statistics retain those frozen decisions to preserve comparability. Direct

evidence therefore means explicit support in a source, not a gold-standard sensory assessment; provisional absence inferred from an incomplete note list is weaker still.

After excluding the 11 preview-exposed rows, 1,134 sampled targets remained among 189 fragrances. The paired comparison contained 403 assessable cells from 105 fragrances. Weighted directional agreement increased from 90.3% to 97.1%, a gain of 6.8 percentage points (conditional 95% fragrance-cluster bootstrap interval [4.6, 9.4]). These results retain the original sampling weights and use all 189 remaining rows as the bootstrap frame, including the 84 without paired assessable cells. Improvement therefore persisted when the preview-exposed rows were removed. This sensitivity analysis does not resolve dependence on shared sources, reference error, or nonrandom evidence availability. Full-precision estimates and the 189-row resampling frame are archived in `derived/masking_sensitivity`.

These limitations are distinct from sampling uncertainty. The conditional bootstrap intervals do not include reference error, unresolved evidence, or the outcomes of newly filled cells. The audit neither involved smelling fragrances nor established the accuracy of the probably/definitely distinction. The revised review's extrapolations from correction behavior are not treated as independent accuracy estimates because they assume that error-detection behavior transfers from assessable to unassessable cells.

Supplementary Archive

`FragranceDive_Supplementary_Data.zip` accompanies the article directly. Its original 167-file component contains the R workflow, factor workbook and CSVs, all agreement tables and confusion counts, cohort membership and diagnostics, original figures, and both complete audit packages. The expanded archive adds the sampling script, interfactor correlations, correction API response/error records, the masking sensitivity analysis, and explicit reuse licenses. Data and questionnaire materials are dedicated to the public domain under CC0 1.0 Universal; scripts are distributed under the MIT License.

`archive_inventory/ORIGINAL_ARCHIVE_INVENTORY.md` and the associated CSV enumerate the originally supplied package; a separate response inventory identifies the added JSON records. The current SHA-256 manifest covers the distributed files.

Rerunning both original audit scripts reproduced all 34 result files byte for byte; separate checks confirmed sampling identities, weighted scoring, and conditional bootstrap intervals. The added masking analysis reproduces the main paired result before applying the exclusion. Numerical reproduction does not re-adjudicate evidence. The article archive alone does not refit the full reference model or regenerate the scatterplots: those tasks additionally require the data released through Zenodo and the documented preprocessing, input-selection, and scoring steps. This is a documented, staged workflow rather than a fully automated one-execution pipeline. Historical audit files are preserved for provenance and checksum verification; their superseded corpus-size and chronology statements do not replace the corrected study description.

Illustrative API Response Excerpts

The following excerpts reproduce selected fields from two result objects in `llm_20260907_221353_943846.json`. Full rationale text, source lists, and the API envelope remain in `correction_api_responses/raw`. The examples illustrate the response format and acceptance threshold; they were not randomly selected and do not independently verify the fragrance claims.

```

{
  "fragrance": "Bavaria The Gemstone Hawk (Fragrance World)",
  "row_index": 92,
  "current_value": 0,
  "recommended_value": 4,
  "confidence": 0.93,
  "needs_change": true
}

{
  "fragrance": "Ingredients 07 2012 (Maison Crivelli)",
  "row_index": 148,
  "current_value": 0,
  "recommended_value": 2,
  "confidence": 0.62,
  "needs_change": true
}

```

For Bavaria The Gemstone Hawk, the returned recommendation changes 0 to 4 with confidence .93, meeting the operational .70 confidence threshold. For Ingredients 07 2012, the returned recommendation is Unknown (2) with confidence .62; because this is below the threshold, it would not be applied. In the analysis files, accepted Unknown outcomes are represented as blanks rather than retained as category 2. A model's needs_change flag alone was insufficient to authorize a change. These excerpts describe returned recommendations, not a reconstructed item-level application log.

Table S14*Oblique Factor Correlation Matrix (Φ)*

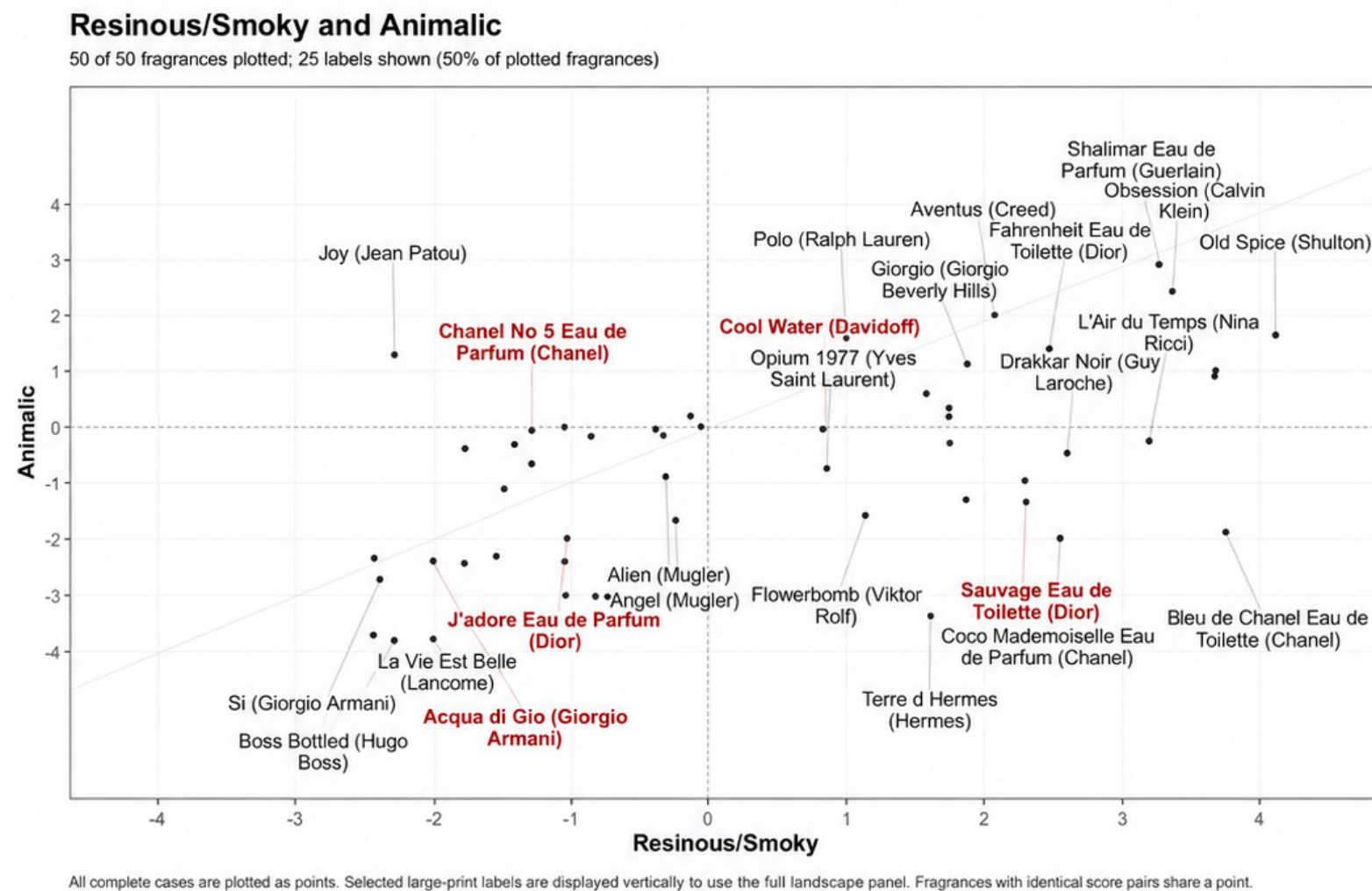
	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12	F13	F14
F1	1.00													
F2	.35	1.00												
F3	.31	.26	1.00											
F4	.32	.20	.14	1.00										
F5	.19	.16	.20	.09	1.00									
F6	.03	-.06	.02	.04	-.17	1.00								
F7	.16	.04	-.03	.07	.09	-.08	1.00							
F8	-.16	-.09	-.05	-.20	.05	-.19	.02	1.00						
F9	.21	.06	.02	.16	.18	-.12	.29	-.06	1.00					
F10	-.08	.06	.11	.06	.14	-.15	-.12	-.07	-.02	1.00				
F11	.01	-.07	-.08	-.06	.12	-.10	.13	.24	.23	-.07	1.00			
F12	-.29	-.21	-.19	-.02	-.11	.10	-.18	.02	-.17	.05	-.08	1.00		
F13	.10	.08	.08	-.03	.10	.06	-.08	-.01	-.03	.05	.08	-.10	1.00	
F14	-.13	-.10	-.06	-.02	-.05	.00	-.10	.07	-.05	.08	.09	.17	.07	1.00

Note. The lower triangle is shown; the matrix is symmetric with a unit diagonal. Coefficients are reproduced at the two-decimal precision of frag_interfactor_correlations.csv, without reconstructing additional precision. Φ describes correlations among the fitted oblique factors, not correlations among the -4 to +4 display scores. F1-F14 follow the same factor order and orientation as the pattern matrix in Table S2.

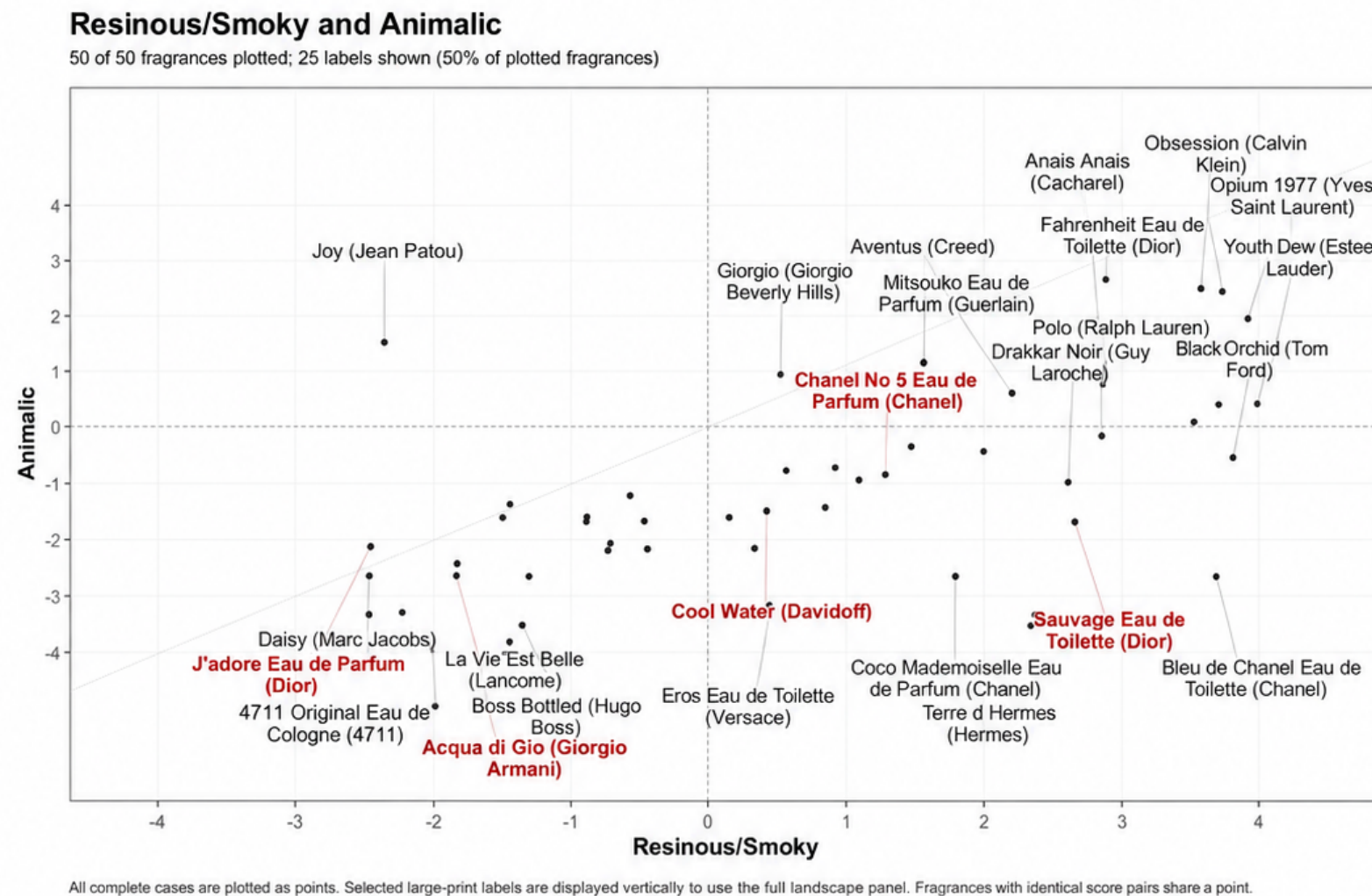
F1 = Resinous/Smoky; F2 = Animalic; F3 = Expensive; F4 = Intense/Powerful; F5 = Earthy; F6 = Gourmand/Sweet; F7 = Masculine; F8 = Aquatic/Ozonic; F9 = Spicy/Woody; F10 = Carnation/Aldehydic; F11 = Minty; F12 = Fruity; F13 = Vintage/Waxy; F14 = US Mainstream.

Figure S1*Resinous/Smoky and Animalic Factor Scores Before and After Correction*

A. Original ratings



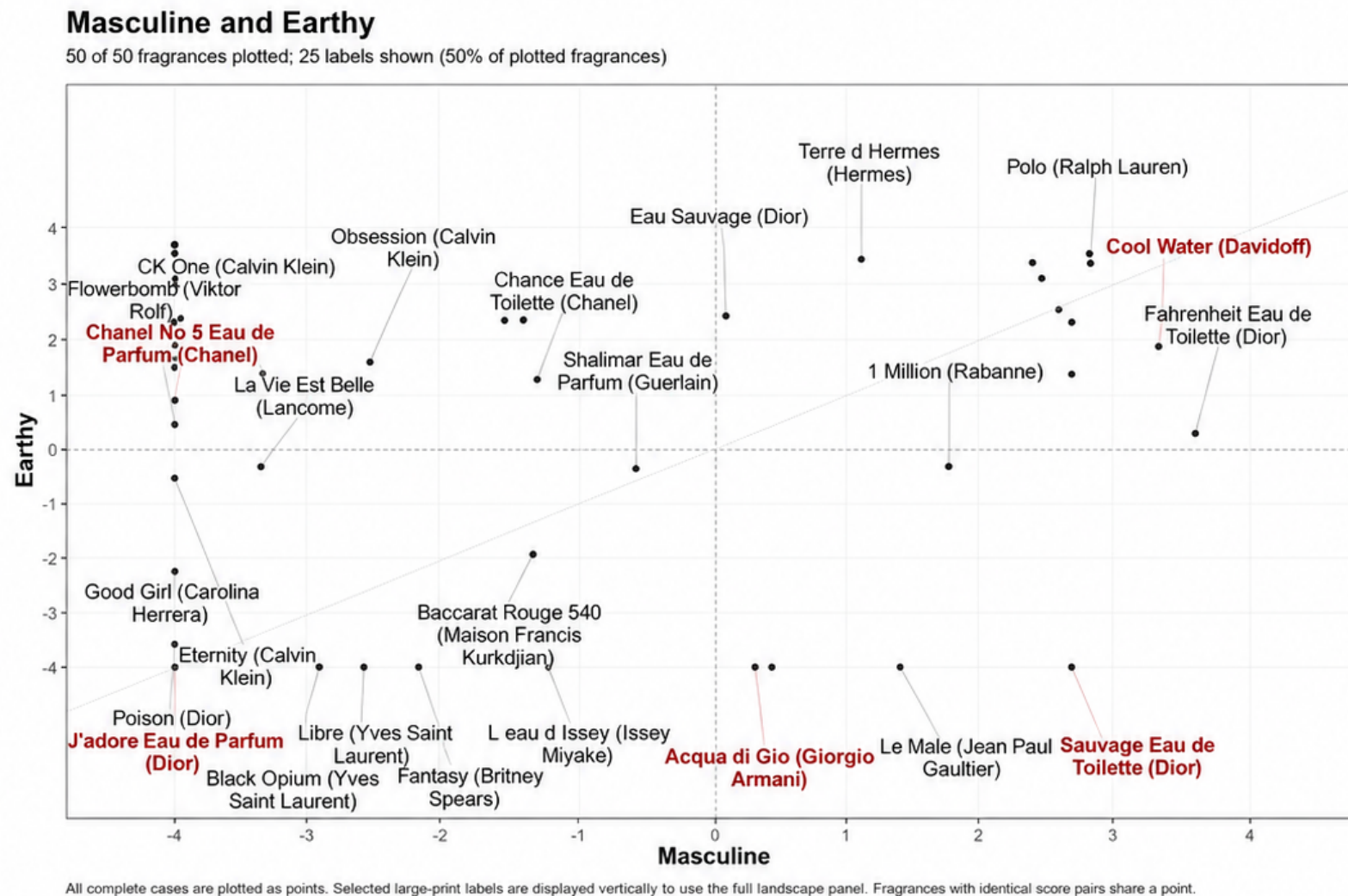
Note. All 50 fragrances are plotted on the fixed reference metric; 25 labels are shown. Leader lines connect displaced labels to points. Red labels emphasize selected examples and do not mark statistical significance. The diagonal is an identity guide, not a fitted regression.

Figure S1 (continued)*Resinous/Smoky and Animalic Factor Scores Before and After Correction***B. Corrected ratings**

Note. All 50 fragrances are plotted on the fixed reference metric; 25 labels are shown. Leader lines connect displaced labels to points. Red labels emphasize selected examples and do not mark statistical significance. The diagonal is an identity guide, not a fitted regression.

Figure S2*Masculine and Earthy Factor Scores Before and After Correction*

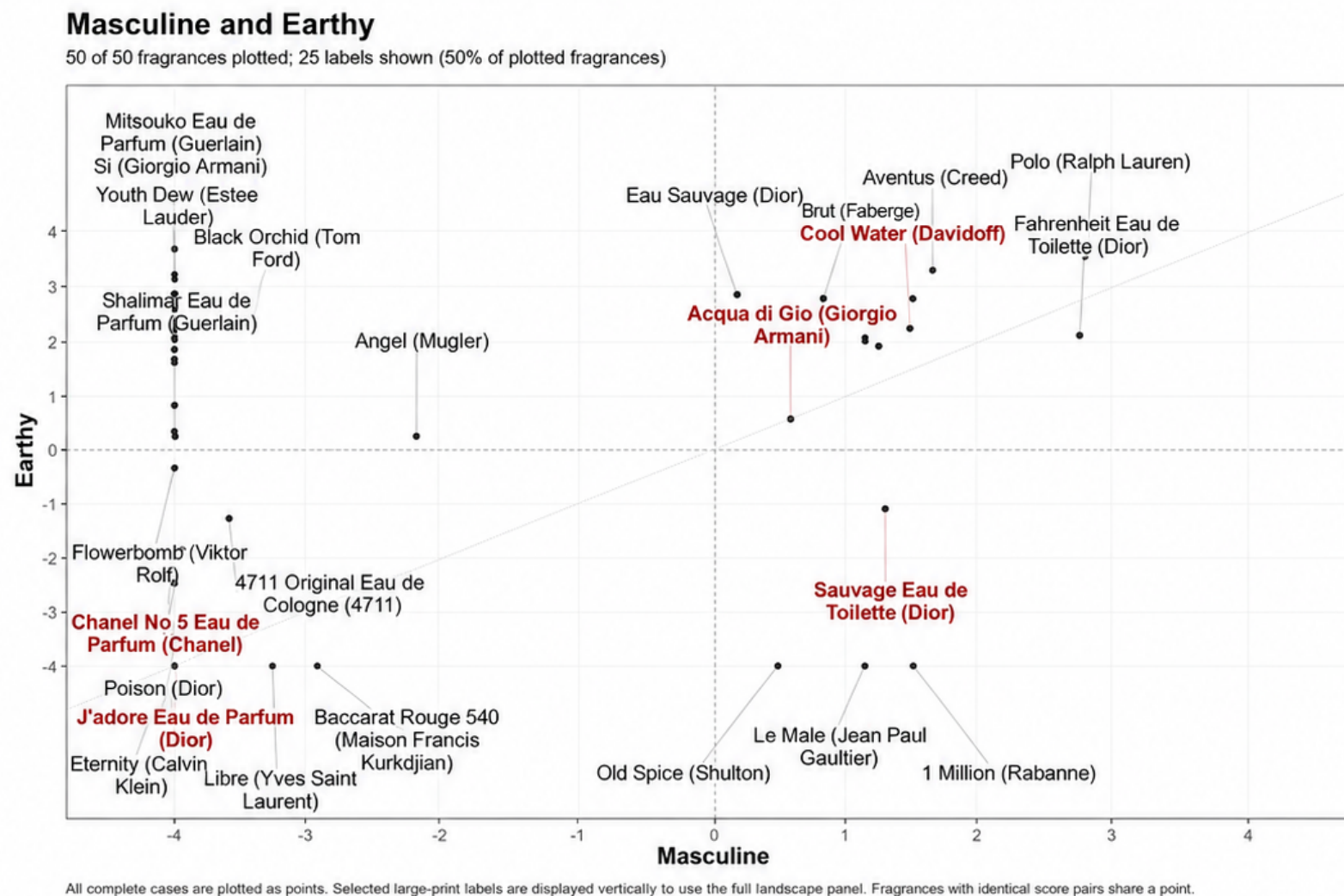
A. Original ratings



Note. All 50 fragrances are plotted on the fixed reference metric; 25 labels are shown. Leader lines connect displaced labels to points. Red labels emphasize selected examples and do not mark statistical significance. The diagonal is an identity guide, not a fitted regression.

Figure S2 (continued)*Masculine and Earthy Factor Scores Before and After Correction*

B. Corrected ratings

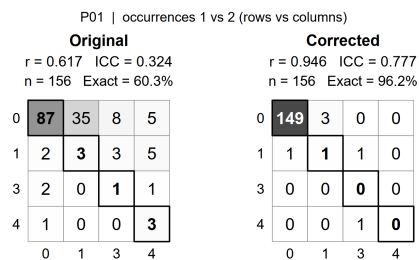
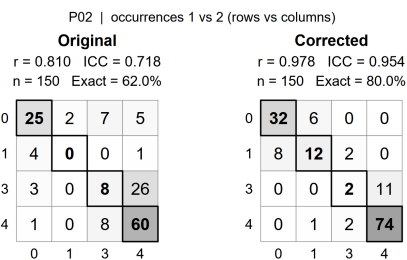
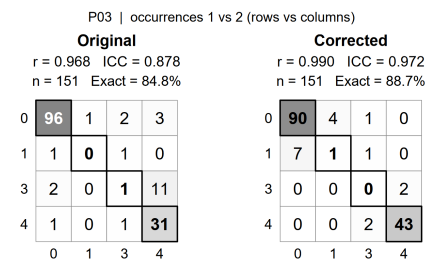
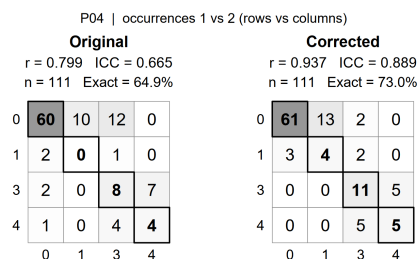
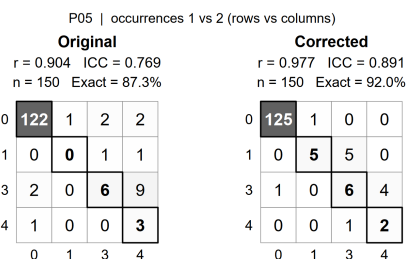
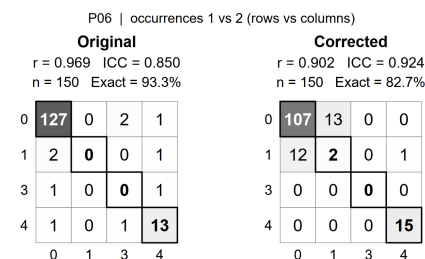
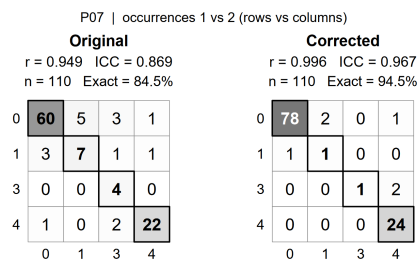
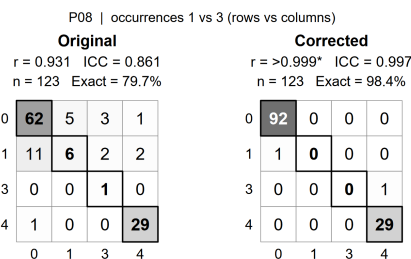
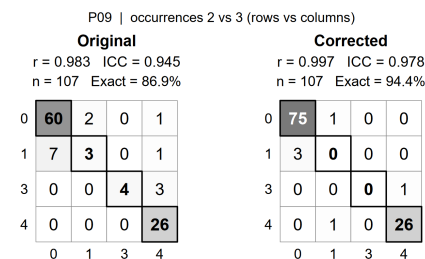


Note. All 50 fragrances are plotted on the fixed reference metric; 25 labels are shown. Leader lines connect displaced labels to points. Red labels emphasize selected examples and do not mark statistical significance. The diagonal is an identity guide, not a fitted regression.

Figure S3*All Random200 Repeated Item Matrices Using Matched Complete Cases***Repeated-item agreement: all 31 pairs**

Page 1 / 4

Matched complete cases: the same fragrances on both sides of each pair

Can this fragrance be found at most U.S. department stores?**Does this fragrance have fruity notes?****Does this fragrance have rose notes?****Does this fragrance have animalic notes?****Is leather a leading note in this fragrance?****Does this fragrance have cardamom notes?****Is this fragrance marketed mostly to men?****Is this fragrance marketed mostly to men?****Is this fragrance marketed mostly to men?**

Cells = counts; shading = share of complete pairs. All panels use the same scale. Diagonal borders = exact agreement.

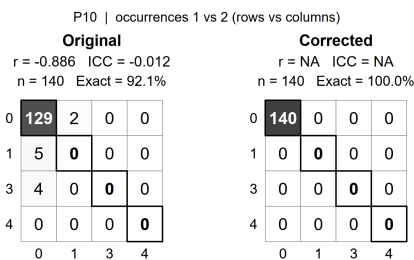
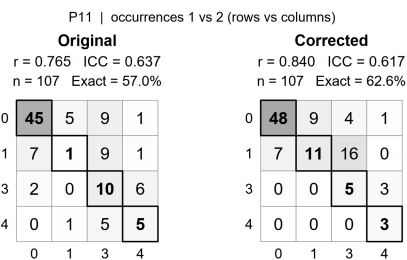
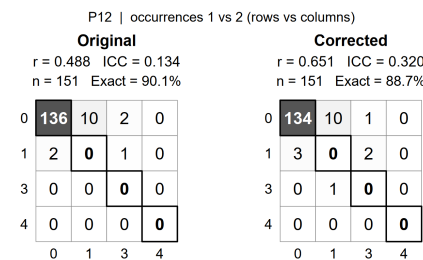
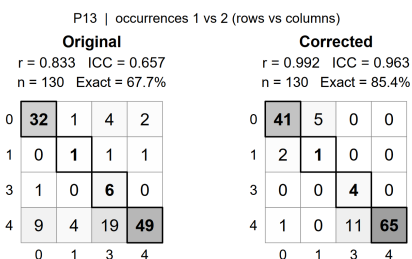
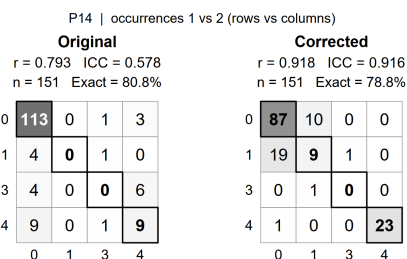
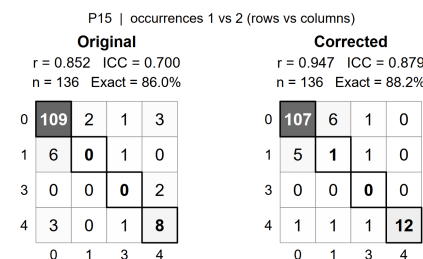
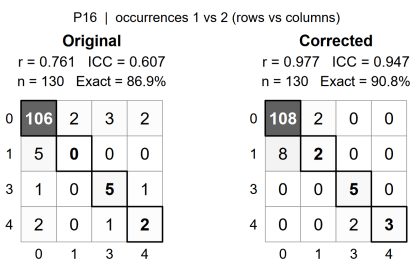
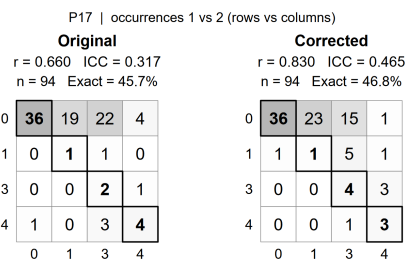
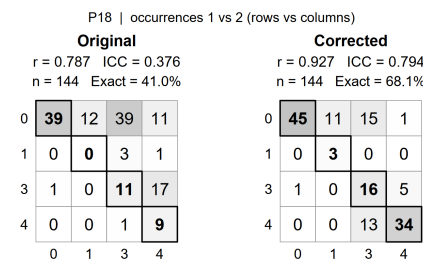
 r = polychoric r ; ICC = ICC(A,1). NA = not estimable; * = near-boundary r . Codes: 0 definitely no; 1 probably no; 3 probably yes; 4 definitely yes.

Note. Each adjacent original/corrected pair represents one repeated question. The same fragrances enter both versions of each pair. Counts and whole-matrix shading are reproduced from the analysis exports. Pair IDs identify the same comparisons as Tables S6–S9.

Figure S3 (continued)*All Random200 Repeated Item Matrices Using Matched Complete Cases***Repeated-item agreement: all 31 pairs**

Page 2 / 4

Matched complete cases: the same fragrances on both sides of each pair

Is petrichor a prominent note in this fragrance?**Is musk a leading note in this fragrance?****Does this fragrance have fecal notes?****Does this fragrance cost less than \$30 per ounce?****Does this fragrance have vetiver notes?****Does this fragrance have mossy notes?****Does this fragrance have ozonic notes?****Is this fragrance considered a "beast mode" projector?****Does this fragrance have gourmand notes?**

Cells = counts; shading = share of complete pairs. All panels use the same scale. Diagonal borders = exact agreement.

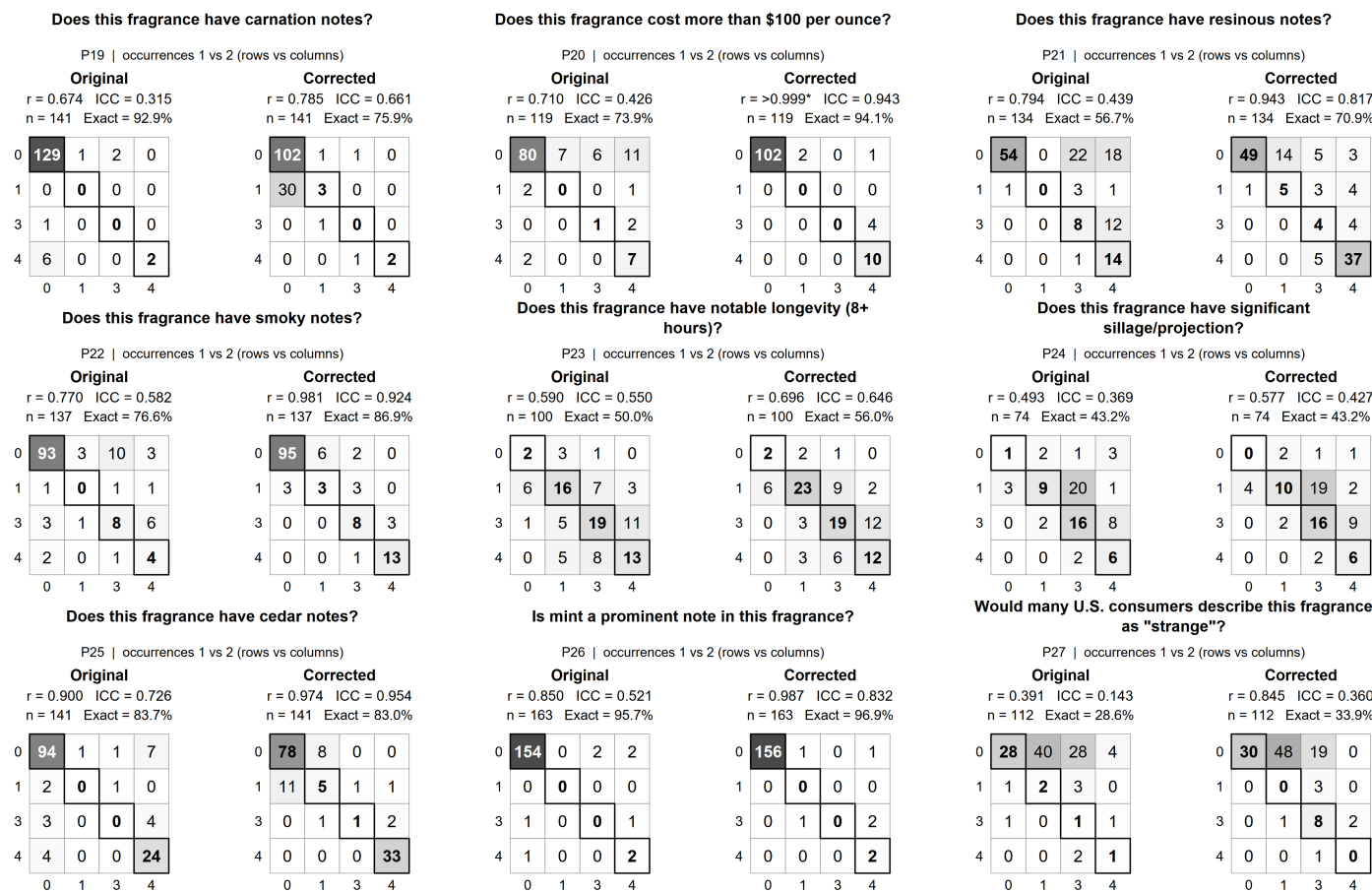
r = polychoric r; ICC = ICC(A,1). NA = not estimable; * = near-boundary r. Codes: 0 definitely no; 1 probably no; 3 probably yes; 4 definitely yes.

Note. Each adjacent original/corrected pair represents one repeated question. The same fragrances enter both versions of each pair. Counts and whole-matrix shading are reproduced from the analysis exports. Pair IDs identify the same comparisons as Tables S6–S9.

Figure S3 (continued)*All Random200 Repeated Item Matrices Using Matched Complete Cases***Repeated-item agreement: all 31 pairs**

Page 3 / 4

Matched complete cases: the same fragrances on both sides of each pair



Cells = counts; shading = share of complete pairs. All panels use the same scale. Diagonal borders = exact agreement.

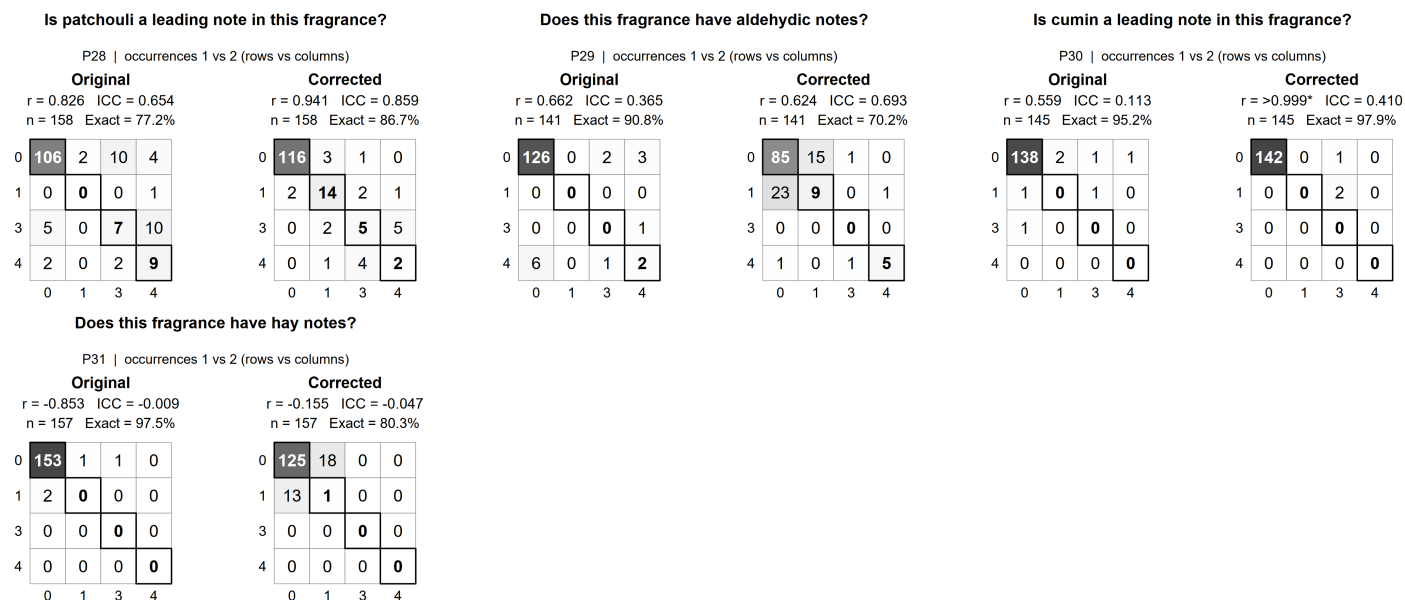
 r = polychoric r ; ICC = ICC(A,1). NA = not estimable; * = near-boundary r . Codes: 0 definitely no; 1 probably no; 3 probably yes; 4 definitely yes.

Note. Each adjacent original/corrected pair represents one repeated question. The same fragrances enter both versions of each pair. Counts and whole-matrix shading are reproduced from the analysis exports. Pair IDs identify the same comparisons as Tables S6–S9.

Figure S3 (continued)*All Random200 Repeated Item Matrices Using Matched Complete Cases***Repeated-item agreement: all 31 pairs**

Page 4 / 4

Matched complete cases: the same fragrances on both sides of each pair



Cells = counts; shading = share of complete pairs. All panels use the same scale. Diagonal borders = exact agreement.

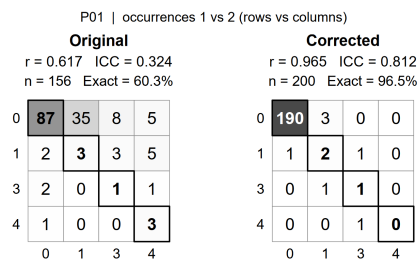
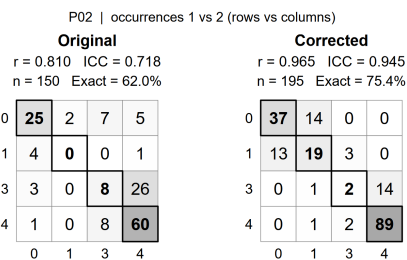
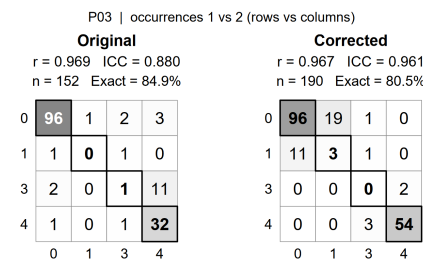
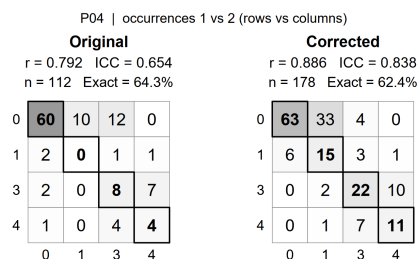
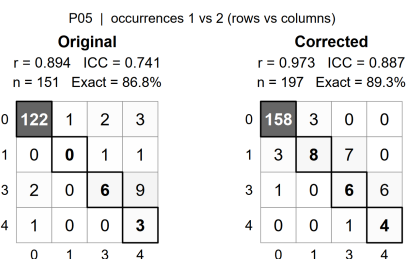
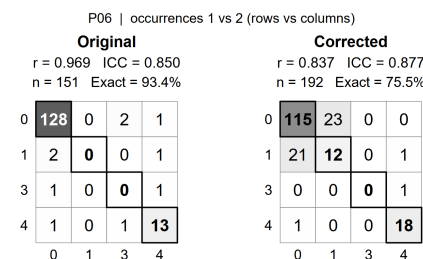
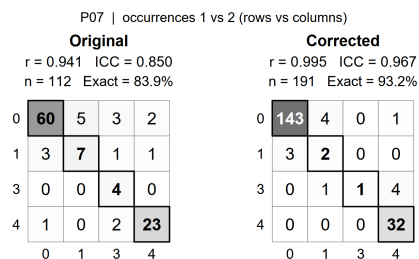
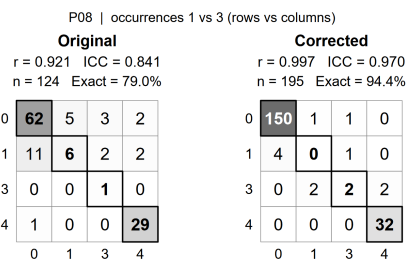
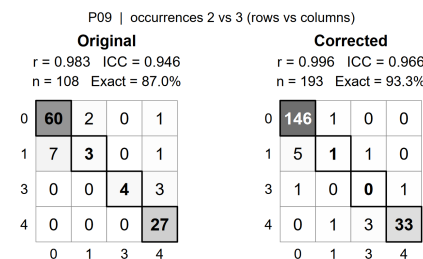
 r = polychoric r ; ICC = $ICC(A,1)$. NA = not estimable; * = near-boundary r . Codes: 0 definitely no; 1 probably no; 3 probably yes; 4 definitely yes.

Note. Each adjacent original/corrected pair represents one repeated question. The same fragrances enter both versions of each pair. Counts and whole-matrix shading are reproduced from the analysis exports. Pair IDs identify the same comparisons as Tables S6–S9.

Figure S4*All Random200 Repeated Item Matrices Using Available Complete Cases***Repeated-item agreement: all 31 pairs**

Page 1 / 4

All available complete pairs within each dataset; sample sizes can differ

Can this fragrance be found at most U.S. department stores?**Does this fragrance have fruity notes?****Does this fragrance have rose notes?****Does this fragrance have animalic notes?****Is leather a leading note in this fragrance?****Does this fragrance have cardamom notes?****Is this fragrance marketed mostly to men?****Is this fragrance marketed mostly to men?****Is this fragrance marketed mostly to men?**

Cells = counts; shading = share of complete pairs. All panels use the same scale. Diagonal borders = exact agreement.

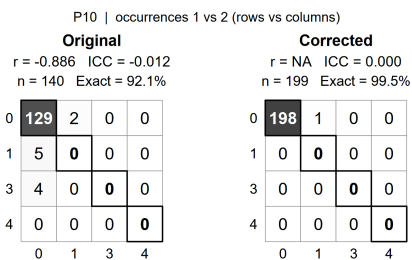
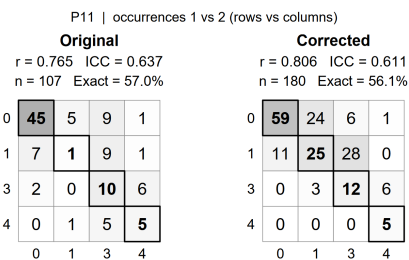
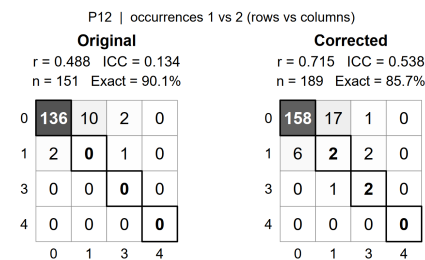
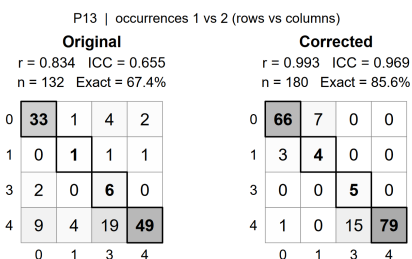
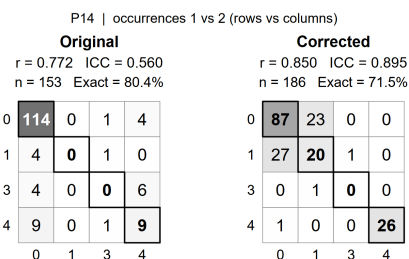
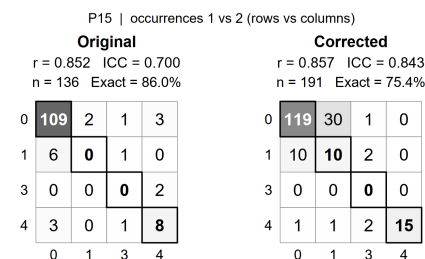
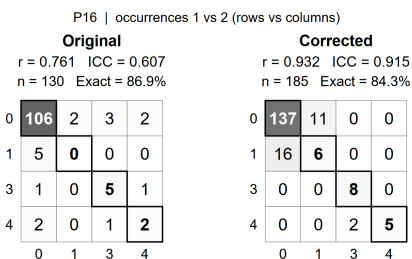
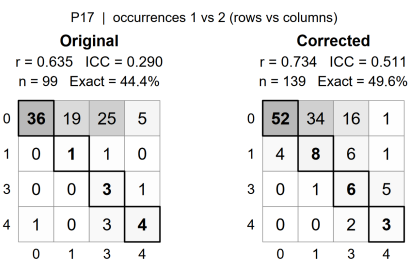
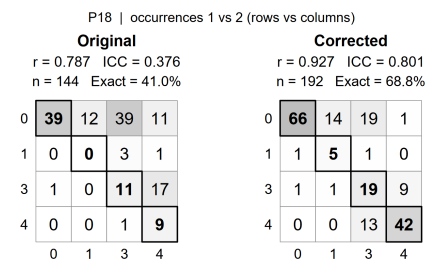
 r = polychoric r ; ICC = $ICC(A,1)$. NA = not estimable; * = near-boundary r . Codes: 0 definitely no; 1 probably no; 3 probably yes; 4 definitely yes.

Note. Each adjacent original/corrected pair represents one repeated question. Sample sizes may differ between versions. Counts and whole-matrix shading are reproduced from the analysis exports. Pair IDs identify the same comparisons as Tables S6–S9.

Figure S4 (continued)*All Random200 Repeated Item Matrices Using Available Complete Cases***Repeated-item agreement: all 31 pairs**

Page 2 / 4

All available complete pairs within each dataset; sample sizes can differ

Is petrichor a prominent note in this fragrance?**Is musk a leading note in this fragrance?****Does this fragrance have fecal notes?****Does this fragrance cost less than \$30 per ounce?****Does this fragrance have vetiver notes?****Does this fragrance have mossy notes?****Does this fragrance have ozonic notes?****Is this fragrance considered a "beast mode" projector?****Does this fragrance have gourmand notes?**

Cells = counts; shading = share of complete pairs. All panels use the same scale. Diagonal borders = exact agreement.

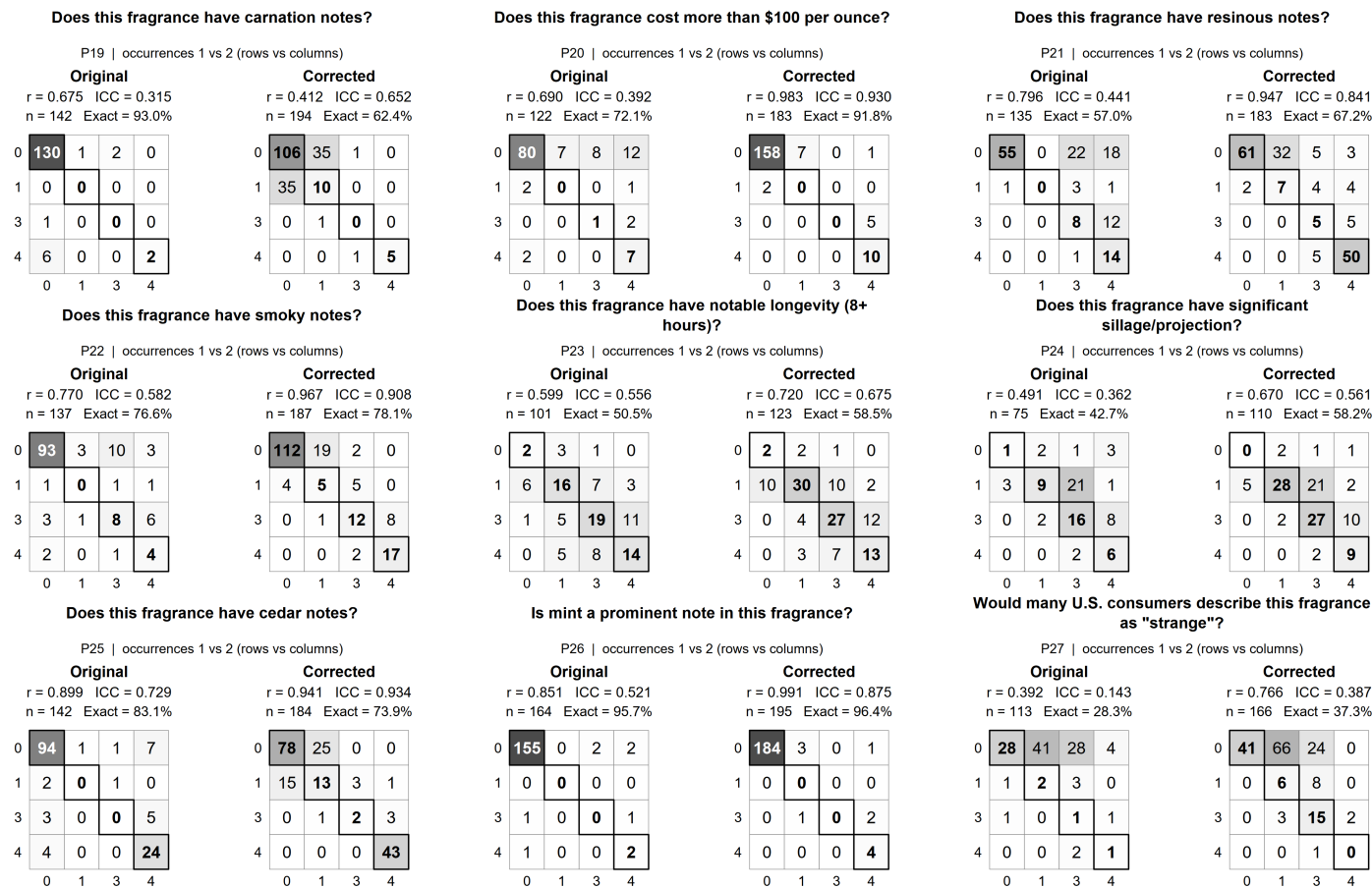
r = polychoric r; ICC = ICC(A,1). NA = not estimable; * = near-boundary r. Codes: 0 definitely no; 1 probably no; 3 probably yes; 4 definitely yes.

Note. Each adjacent original/corrected pair represents one repeated question. Sample sizes may differ between versions. Counts and whole-matrix shading are reproduced from the analysis exports. Pair IDs identify the same comparisons as Tables S6–S9.

Figure S4 (continued)*All Random200 Repeated Item Matrices Using Available Complete Cases***Repeated-item agreement: all 31 pairs**

Page 3 / 4

All available complete pairs within each dataset; sample sizes can differ



Cells = counts; shading = share of complete pairs. All panels use the same scale. Diagonal borders = exact agreement.

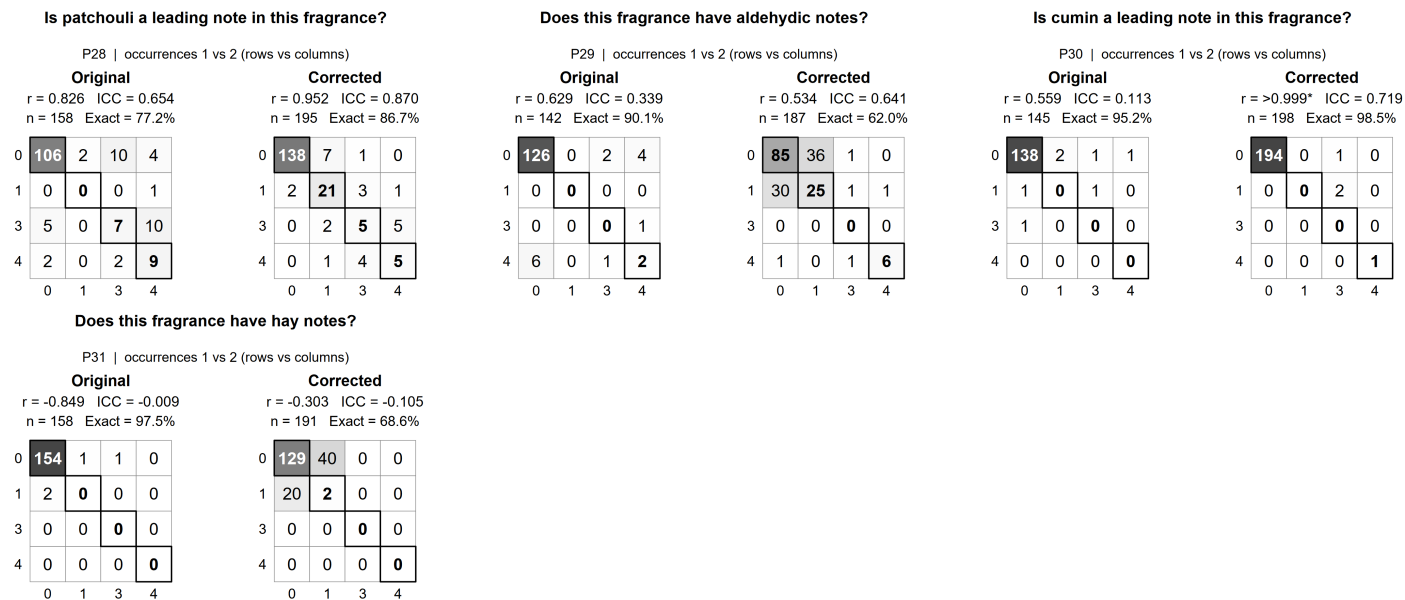
r = polychoric r; ICC = ICC(A,1). NA = not estimable; * = near-boundary r. Codes: 0 definitely no; 1 probably no; 3 probably yes; 4 definitely yes.

Note. Each adjacent original/corrected pair represents one repeated question. Sample sizes may differ between versions. Counts and whole-matrix shading are reproduced from the analysis exports. Pair IDs identify the same comparisons as Tables S6–S9.

Figure S4 (continued)*All Random200 Repeated Item Matrices Using Available Complete Cases***Repeated-item agreement: all 31 pairs**

Page 4 / 4

All available complete pairs within each dataset; sample sizes can differ



Cells = counts; shading = share of complete pairs. All panels use the same scale. Diagonal borders = exact agreement.

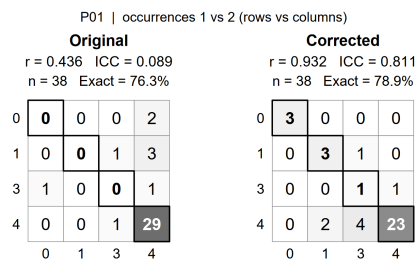
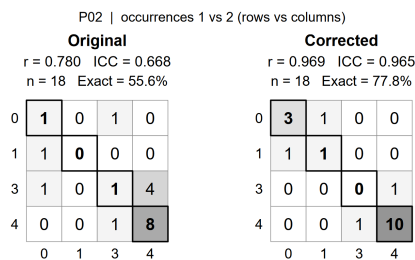
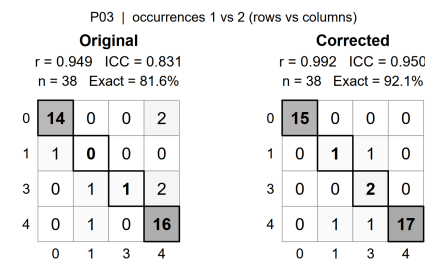
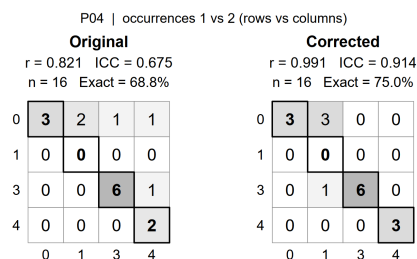
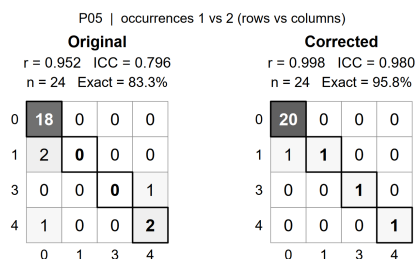
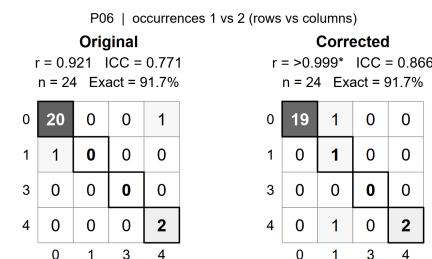
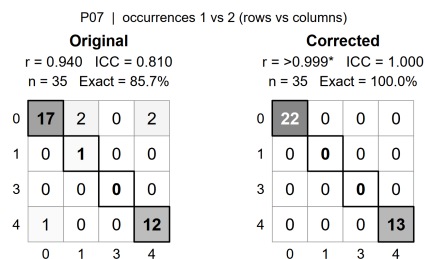
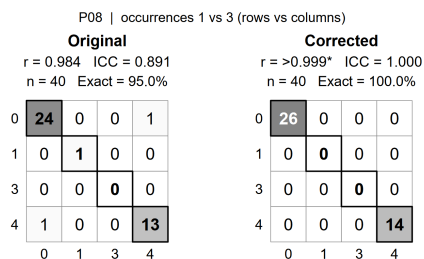
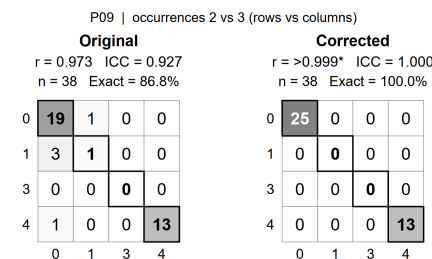
 r = polychoric r ; ICC = ICC(A,1). NA = not estimable; * = near-boundary r . Codes: 0 definitely no; 1 probably no; 3 probably yes; 4 definitely yes.

Note. Each adjacent original/corrected pair represents one repeated question. Sample sizes may differ between versions. Counts and whole-matrix shading are reproduced from the analysis exports. Pair IDs identify the same comparisons as Tables S6–S9.

Figure S5*All Famous50 Repeated Item Matrices Using Matched Complete Cases***Repeated-item agreement: all 31 pairs**

Matched complete cases: the same fragrances on both sides of each pair

Page 1 / 4

Can.this.fragrance.be.found.at.most.U.S..department.stores.**Does.this.fragrance.have.fruity.notes.****Does.this.fragrance.have.rose.notes.****Does.this.fragrance.have.animalic.notes.****Is.leather.a.leading.note.in.this.fragrance.****Does.this.fragrance.have.cardamom.notes.****Is.this.fragrance.marketed.mostly.to.men.****Is.this.fragrance.marketed.mostly.to.men.****Is.this.fragrance.marketed.mostly.to.men.**

Cells = counts; shading = share of complete pairs. All panels use the same scale. Diagonal borders = exact agreement.

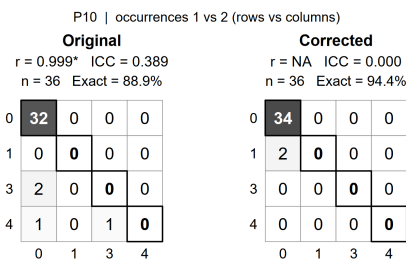
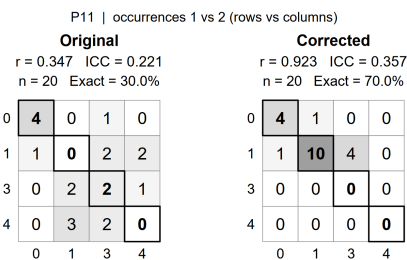
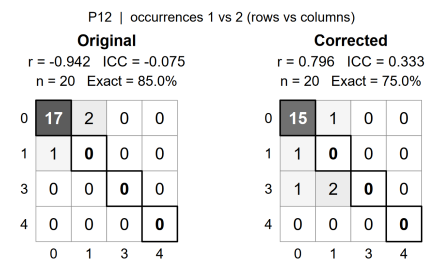
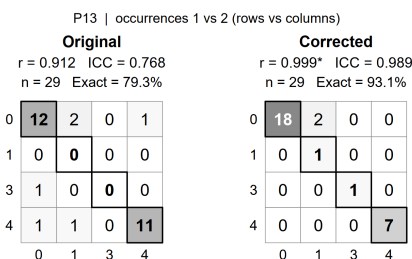
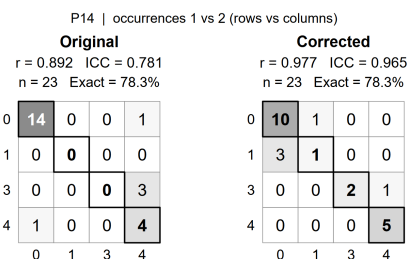
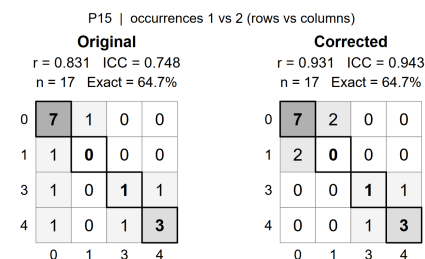
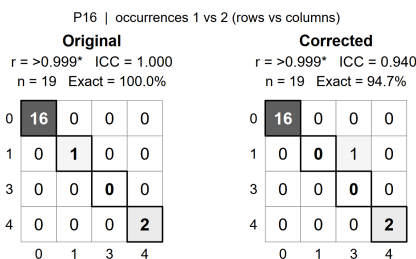
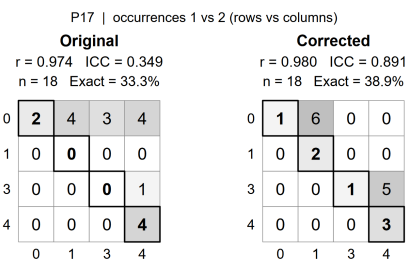
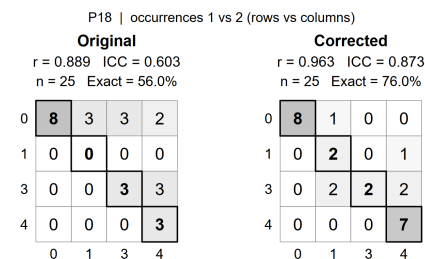
 r = polychoric r ; ICC = ICC(A,1). NA = not estimable; * = near-boundary r . Codes: 0 definitely no; 1 probably no; 3 probably yes; 4 definitely yes.

Note. Each adjacent original/corrected pair represents one repeated question. The same fragrances enter both versions of each pair. Counts and whole-matrix shading are reproduced from the analysis exports. Pair IDs identify the same comparisons as Tables S6–S9.

Figure S5 (continued)*All Famous50 Repeated Item Matrices Using Matched Complete Cases***Repeated-item agreement: all 31 pairs**

Page 2 / 4

Matched complete cases: the same fragrances on both sides of each pair

Is.petrichor.a.prominent.note.in.this.fragrance.**Is.musk.a.leading.note.in.this.fragrance.****Does.this.fragrance.have.fecal.notes.****Does.this.fragrance.cost.less.than..30.per.ounce.****Does.this.fragrance.have.vetiver.notes.****Does.this.fragrance.have.mossy.notes.****Does.this.fragrance.have.ozonic.notes.****Is.this.fragrance.considered.a..beast.mode..projector.****Does.this.fragrance.have.gourmand.notes.**

Cells = counts; shading = share of complete pairs. All panels use the same scale. Diagonal borders = exact agreement.

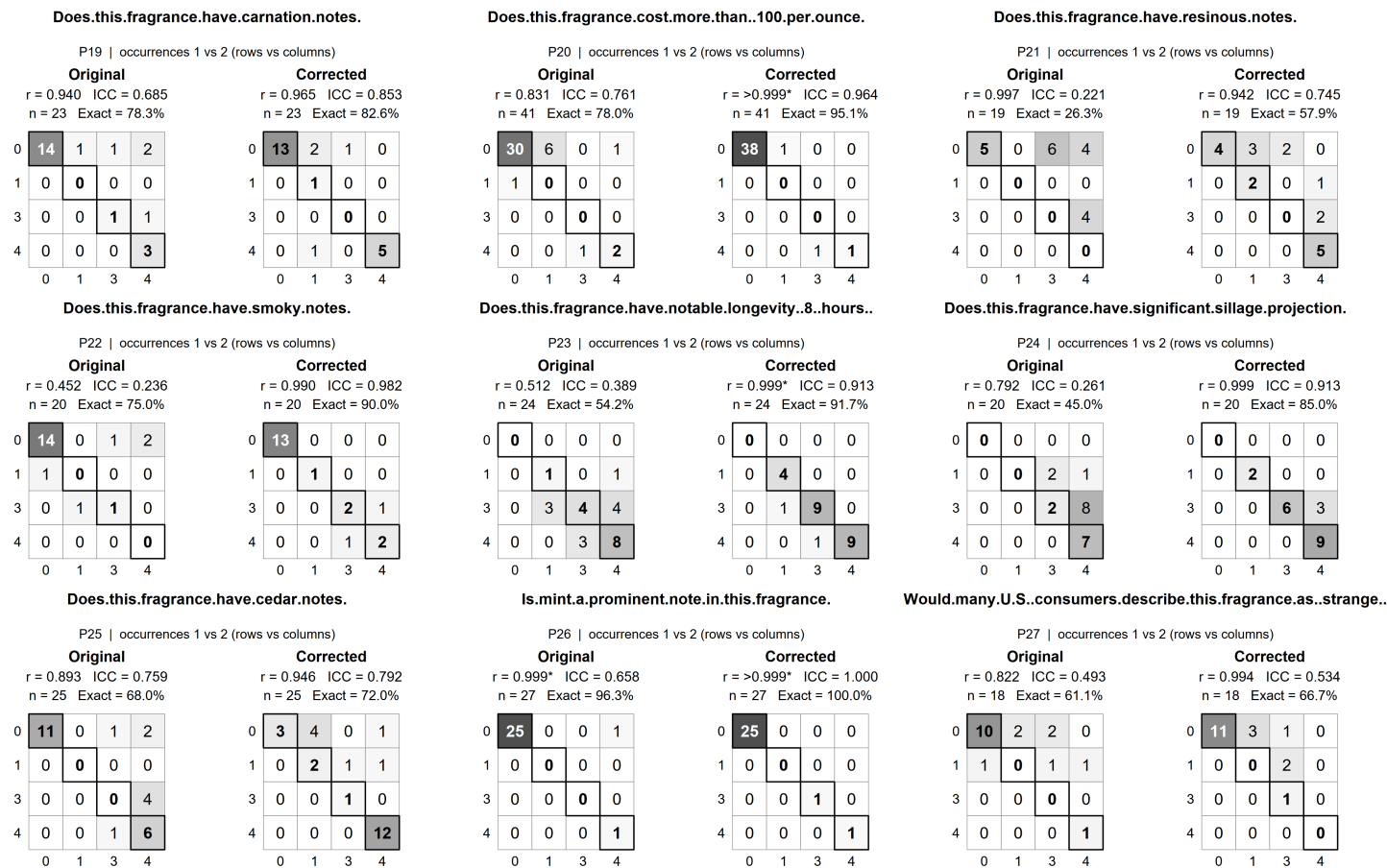
 r = polychoric r ; ICC = ICC(A,1). NA = not estimable; * = near-boundary r . Codes: 0 definitely no; 1 probably no; 3 probably yes; 4 definitely yes.

Note. Each adjacent original/corrected pair represents one repeated question. The same fragrances enter both versions of each pair. Counts and whole-matrix shading are reproduced from the analysis exports. Pair IDs identify the same comparisons as Tables S6–S9.

Figure S5 (continued)*All Famous50 Repeated Item Matrices Using Matched Complete Cases***Repeated-item agreement: all 31 pairs**

Page 3 / 4

Matched complete cases: the same fragrances on both sides of each pair



Cells = counts; shading = share of complete pairs. All panels use the same scale. Diagonal borders = exact agreement.

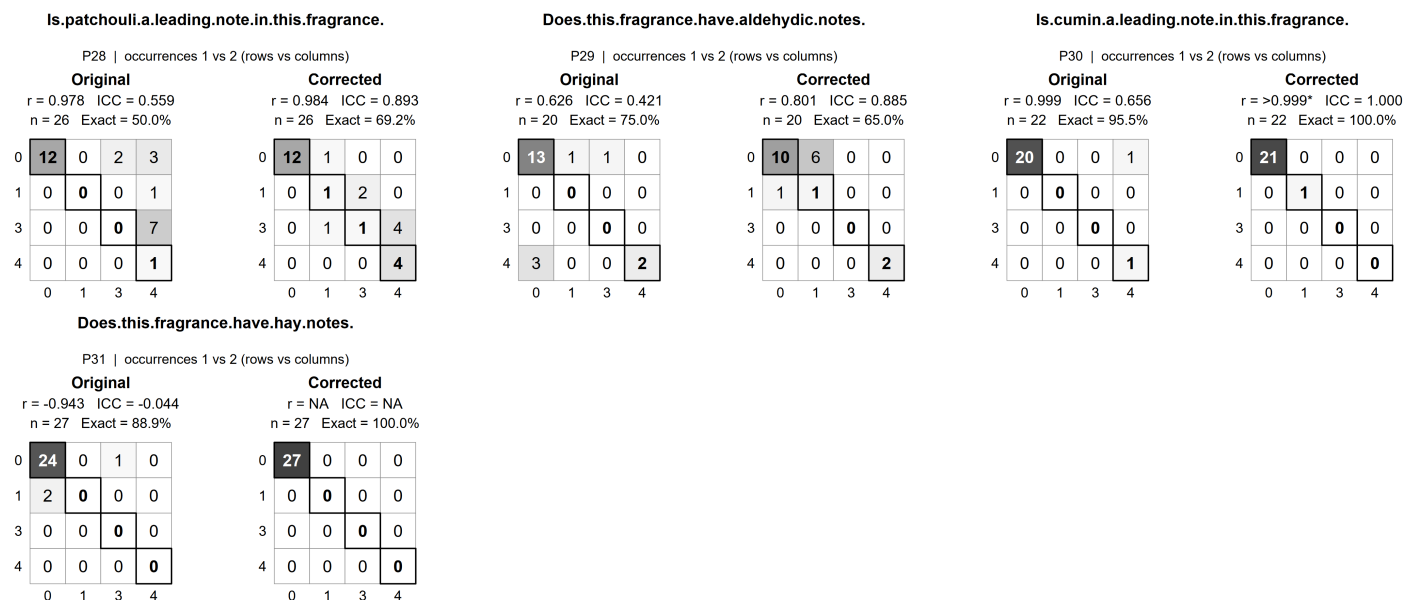
 r = polychoric r ; ICC = ICC(A,1). NA = not estimable; * = near-boundary r . Codes: 0 definitely no; 1 probably no; 3 probably yes; 4 definitely yes.

Note. Each adjacent original/corrected pair represents one repeated question. The same fragrances enter both versions of each pair. Counts and whole-matrix shading are reproduced from the analysis exports. Pair IDs identify the same comparisons as Tables S6–S9.

Figure S5 (continued)*All Famous50 Repeated Item Matrices Using Matched Complete Cases***Repeated-item agreement: all 31 pairs**

Page 4 / 4

Matched complete cases: the same fragrances on both sides of each pair



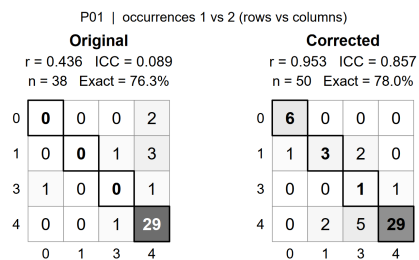
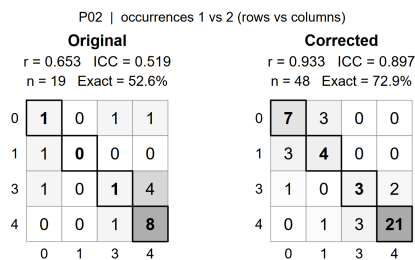
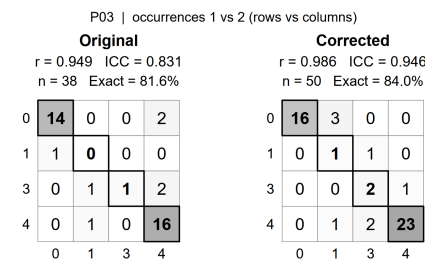
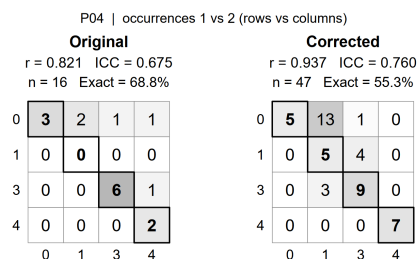
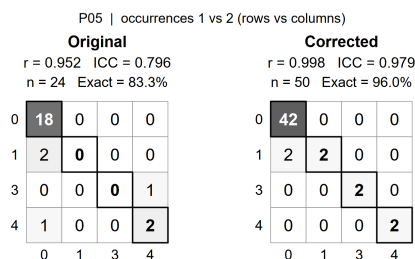
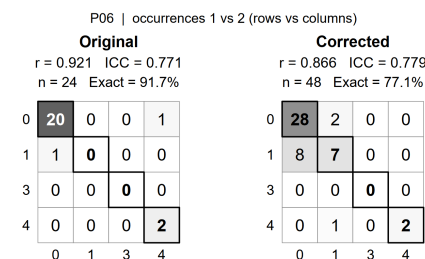
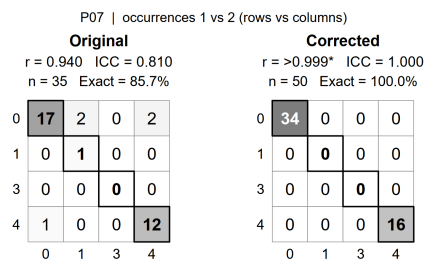
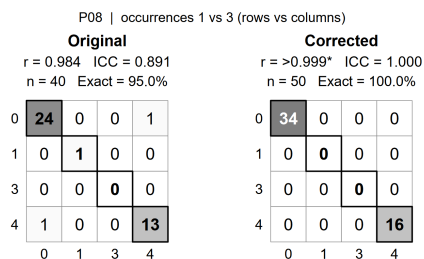
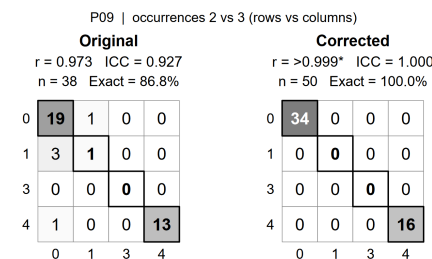
Cells = counts; shading = share of complete pairs. All panels use the same scale. Diagonal borders = exact agreement.
 r = polychoric r ; ICC = ICC(A,1). NA = not estimable; * = near-boundary r . Codes: 0 definitely no; 1 probably no; 3 probably yes; 4 definitely yes.

Note. Each adjacent original/corrected pair represents one repeated question. The same fragrances enter both versions of each pair. Counts and whole-matrix shading are reproduced from the analysis exports. Pair IDs identify the same comparisons as Tables S6–S9.

Figure S6*All Famous50 Repeated Item Matrices Using Available Complete Cases***Repeated-item agreement: all 31 pairs**

All available complete pairs within each dataset; sample sizes can differ

Page 1 / 4

Can.this.fragrance.be.found.at.most.U.S..department.stores.**Does.this.fragrance.have.fruity.notes.****Does.this.fragrance.have.rose.notes.****Does.this.fragrance.have.animalic.notes.****Is.leather.a.leading.note.in.this.fragrance.****Does.this.fragrance.have.cardamom.notes.****Is.this.fragrance.marketed.mostly.to.men.****Is.this.fragrance.marketed.mostly.to.men.****Is.this.fragrance.marketed.mostly.to.men.**

Cells = counts; shading = share of complete pairs. All panels use the same scale. Diagonal borders = exact agreement.

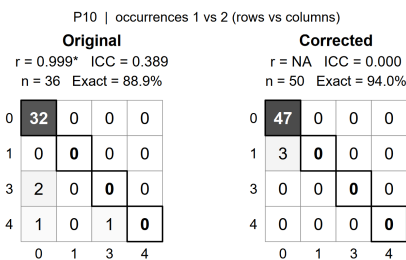
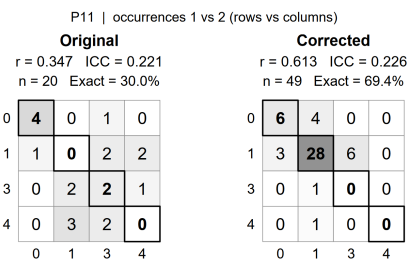
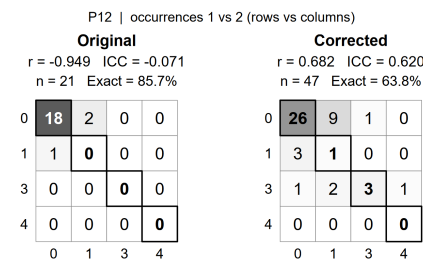
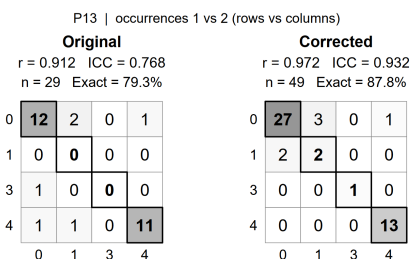
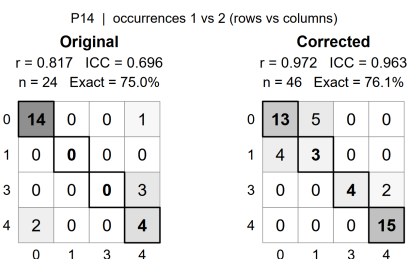
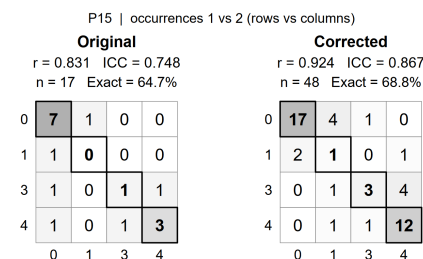
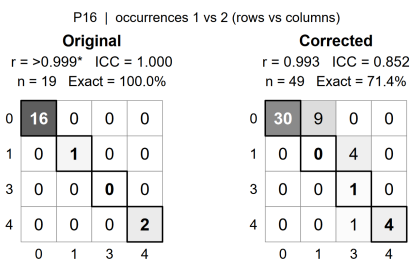
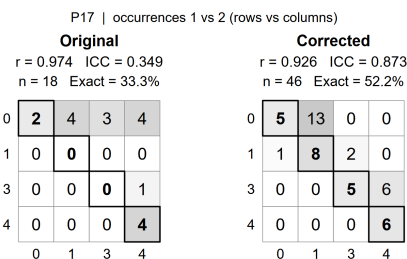
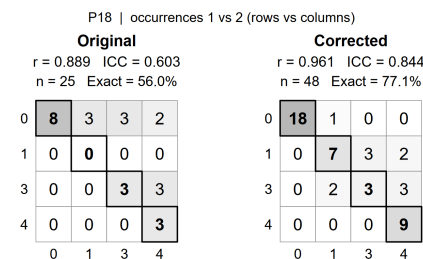
r = polychoric r; ICC = ICC(A,1). NA = not estimable; * = near-boundary r. Codes: 0 definitely no; 1 probably no; 3 probably yes; 4 definitely yes.

Note. Each adjacent original/corrected pair represents one repeated question. Sample sizes may differ between versions. Counts and whole-matrix shading are reproduced from the analysis exports. Pair IDs identify the same comparisons as Tables S6–S9.

Figure S6 (continued)*All Famous50 Repeated Item Matrices Using Available Complete Cases***Repeated-item agreement: all 31 pairs**

Page 2 / 4

All available complete pairs within each dataset; sample sizes can differ

Is.petrichor.a.prominent.note.in.this.fragrance.**Is.musk.a.leading.note.in.this.fragrance.****Does.this.fragrance.have.fecal.notes.****Does.this.fragrance.cost.less.than..30.per.ounce.****Does.this.fragrance.have.vetiver.notes.****Does.this.fragrance.have.mossy.notes.****Does.this.fragrance.have.ozonic.notes.****Is.this.fragrance.considered.a..beast.mode..projector.****Does.this.fragrance.have.gourmand.notes.**

Cells = counts; shading = share of complete pairs. All panels use the same scale. Diagonal borders = exact agreement.

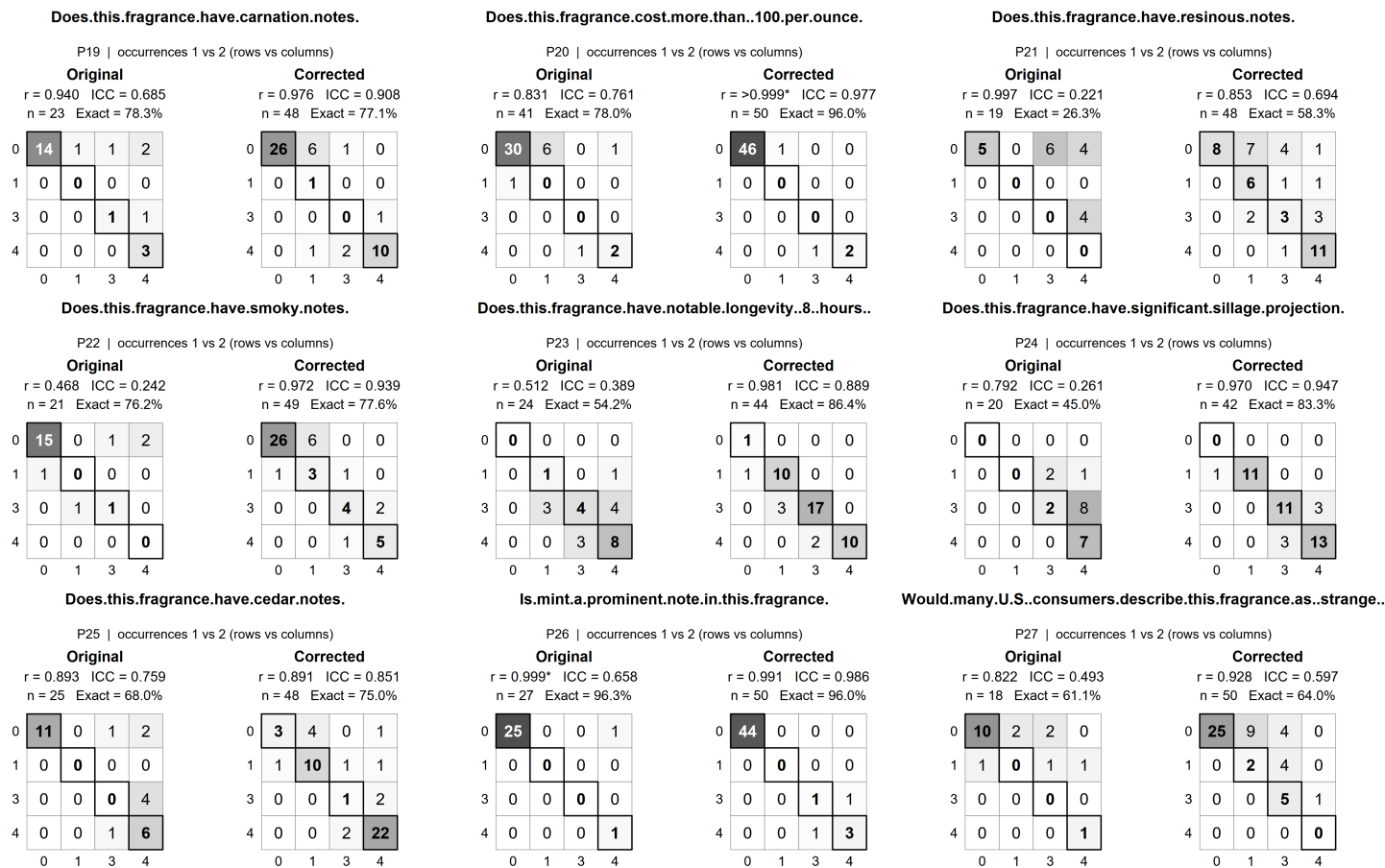
r = polychoric r; ICC = ICC(A,1). NA = not estimable; * = near-boundary r. Codes: 0 definitely no; 1 probably no; 3 probably yes; 4 definitely yes.

Note. Each adjacent original/corrected pair represents one repeated question. Sample sizes may differ between versions. Counts and whole-matrix shading are reproduced from the analysis exports. Pair IDs identify the same comparisons as Tables S6–S9.

Figure S6 (continued)*All Famous50 Repeated Item Matrices Using Available Complete Cases***Repeated-item agreement: all 31 pairs**

Page 3 / 4

All available complete pairs within each dataset; sample sizes can differ



Cells = counts; shading = share of complete pairs. All panels use the same scale. Diagonal borders = exact agreement.

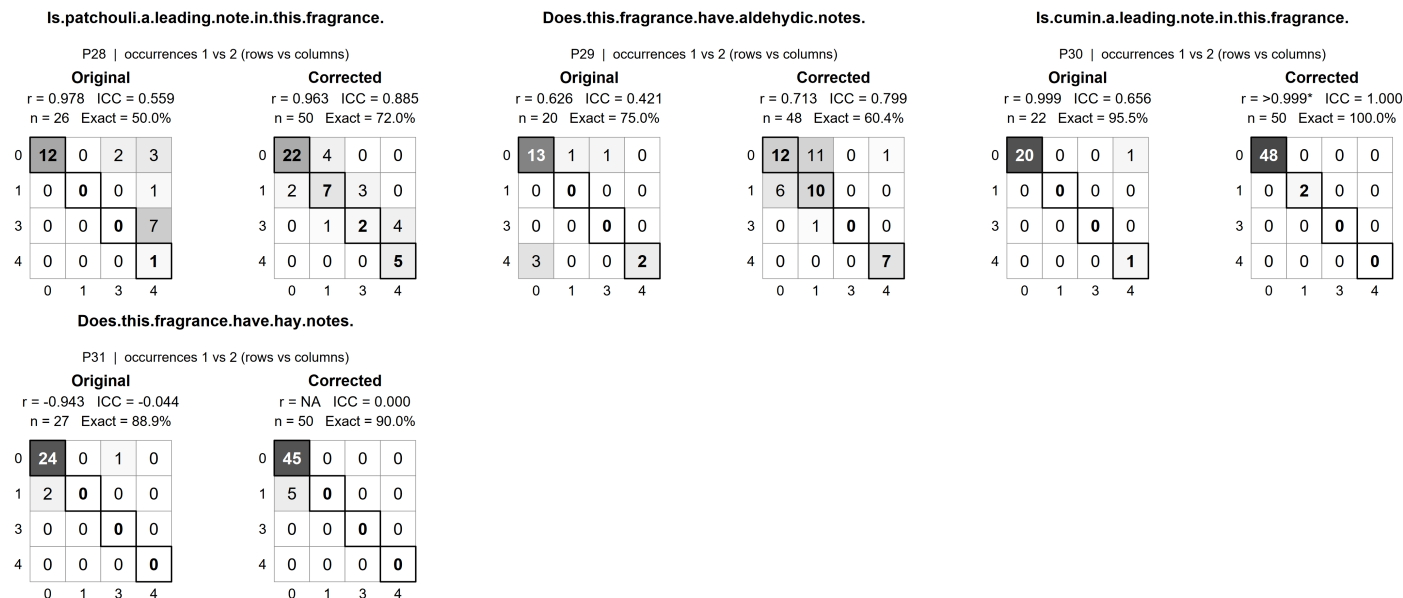
 r = polychoric r ; ICC = $ICC(A,1)$. NA = not estimable; * = near-boundary r . Codes: 0 definitely no; 1 probably no; 3 probably yes; 4 definitely yes.

Note. Each adjacent original/corrected pair represents one repeated question. Sample sizes may differ between versions. Counts and whole-matrix shading are reproduced from the analysis exports. Pair IDs identify the same comparisons as Tables S6–S9.

Figure S6 (continued)*All Famous50 Repeated Item Matrices Using Available Complete Cases***Repeated-item agreement: all 31 pairs**

Page 4 / 4

All available complete pairs within each dataset; sample sizes can differ



Cells = counts; shading = share of complete pairs. All panels use the same scale. Diagonal borders = exact agreement.
 r = polychoric r ; ICC = ICC(A,1). NA = not estimable; * = near-boundary r . Codes: 0 definitely no; 1 probably no; 3 probably yes; 4 definitely yes.

Note. Each adjacent original/corrected pair represents one repeated question. Sample sizes may differ between versions. Counts and whole-matrix shading are reproduced from the analysis exports. Pair IDs identify the same comparisons as Tables S6–S9.

Supplementary Discussion

Precedents for structured annotation

Structured entity annotation asks a model to answer a fixed battery of questions about an entity by synthesizing across sources rather than transcribing any single one of them. On this framing an individual response is an estimate rather than a retrieved fact, and the resulting matrix is evaluated in aggregate through external agreement, calibration, and the coherence of the dimensions it supports. Moore (2025) demonstrated the idea on 327 land animals and 202 well-known people, obtaining estimates that correlated with independently curated values for body size ($r = .85$) and metabolic rate ($r = -.74$) and a questionnaire matrix that yielded interpretable factors. Two subsequent resources applied the approach at scale using a web-search-enabled reasoning model and the same five-point response scale used here. LifeDive rated 35,856 animal, plant, and fungal species on 196 phenotype questions, recovering 14 interpretable dimensions and an audited directional accuracy of roughly 95% to 97%, with errors concentrated in responses the model had hedged (Moore, 2026a). PharmaDive rated 16,116 substances on 130 structural, pharmacological, and clinical attributes, recovering 11 interpretable dimensions and a pronounced error asymmetry in which negative ratings were 98% to 99% accurate while affirmative ratings accounted for most errors (Moore, 2026b). These precedents indicate that the approach can preserve analyzable structure across domains, and they identify the properties, particularly the distribution of errors across response options, that warrant checking in a new one.

Statistical work on imperfect annotations indicates what makes their use defensible. Angelopoulos, Bates, Fannjiang, Jordan, and Zrnic (2023) showed that abundant model predictions can be combined with a smaller set of gold-standard measurements to produce confidence intervals that remain valid whatever algorithm generated the predictions, provided the validated cases are sampled at random from the same population. Gligorić, Zrnic, Lee, Candès, and Jurafsky (2025) showed that verification effort can also be allocated selectively: using model confidence to choose which cases received human annotation reduced the human labels required by more than a quarter while preserving valid intervals, with safeguards preventing poorly calibrated confidence from degrading the estimate. Together, these results indicate that a machine-annotated corpus is most useful when paired with an explicitly sampled validation set, and that scarce review is best directed at the responses least well supported.

Correction mechanisms and agreement diagnostics

The procedure combines two operations with distinct roles. Multivariate prediction identifies answers that deserve attention because they depart from the rest of a fragrance's profile; web-assisted adjudication then evaluates the answer itself. This division is consistent with the rationale of CRITIC, in which access to external tools supported correction of LLM outputs across several tasks (Gou et al., 2024). Statistical unexpectedness is useful for allocating review effort, but unusual combinations may be defining features of a fragrance. A correction system must preserve those combinations when sources support them. The instruction to treat triage statistics as a reason for review, together with the option to retain the original answer or record Unknown, is therefore central to the method's intended behavior.

The present comparison evaluates the correction workflow as a whole in Random200 and Famous50, with comprehensive rather than selectively restricted review in Famous50; performance of correction across the full corpus and stability of a refitted factor structure

remain to be established. It does not isolate the contribution of outlier selection from the use of a different model, higher reasoning effort, a more focused one-question prompt, or additional searching. A randomized comparison of outlier-selected and randomly selected cells under the same adjudication conditions would establish whether statistical triage improves the number of verified errors corrected per unit of review effort. Such a comparison should hold the adjudicator, prompt, search access, and review budget constant while changing only the cell-selection rule, and should account separately for filled missing responses and for valid unusual combinations retained, so that greater efficiency in finding errors is distinguished from a general benefit of obtaining more information. Additional comparisons could assess whether withholding the current answer and predicted category reduces anchoring. Such tests would also reveal whether increasingly consistent data retain legitimate rare profiles or become overly concentrated around common combinations.

Confidence calibration and missing responses

The continuous confidence used to accept adjudications requires empirical calibration. Xiong et al. (2024) found that verbalized LLM confidence was often overconfident and that calibration varied across models, tasks, and elicitation strategies. Tian et al. (2023) reported the complementary result that verbalized confidence can be better calibrated than the same models' internal probabilities, and that calibration improves when a model is prompted to weigh alternative answers before committing to one. Both sets of findings come from short-form factual questions with known answers, so they establish that such judgments can carry usable signal without indicating how well confidence is calibrated for a fragrance attribute that may have no documented answer at all. Their sensitivity to prompt design applies here as well, because the adjudication prompt was fixed and its influence on the reported confidences was not varied. The threshold of .70 in the present workflow is therefore an operational rule, rather than evidence that accepted changes have at least a 70% probability of being correct. Future audits should estimate correctness across confidence levels and compare how different thresholds trade additional coverage against error. Original code-2 responses were converted to blanks before statistical analysis and were not retained as substantive or informative indicators. A future extension could preserve separate provenance for unsupported answers, inapplicable questions, and technical nonresponse, but that would be a new analysis rather than a feature of the present results. Such provenance may be particularly useful for obscure releases, where sparse documentation can affect the available response profile (Kandpal, Deng, Roberts, Wallace, & Raffel, 2023).

A targeted calibration study could separately test the ordinal response labels and the continuous confidence reported during adjudication. For each, assessors would need a probability sample covering affirmative and negative answers, common and unusual attributes, and different levels of documentation. Accuracy by confidence category should retain unresolved references rather than simply dropping them, because a score may appear well calibrated only where evidence is easiest to find. Different acceptance thresholds could then be compared on verified error reduction and added coverage. These analyses would assess the measurement value of confidence instead of assuming that a numerical confidence field is already calibrated.

Language and cultural context

The 14-factor solution should be understood as an empirical summary of this instrument and corpus. Odor arrangements depend on the stimuli, descriptors, assessors, and analytic procedures used to construct them (Kaeppler & Mueller, 2013). The descriptive vocabulary is a further variable. Majid and Burenhult (2014) found that speakers of Jahai named odors as consistently as they named colors, using dedicated abstract odor terms, whereas matched English speakers named odors far less consistently and described them mainly by their sources. English-language description of commercial fragrances is one such naming system, shaped by a particular perfumery tradition and marketplace. The dimensions recovered here are properties of that vocabulary as applied to these products, and a catalog described in another language or under different market conventions need not yield the same organization. Nikitina (2024) provides a practical starting point for multilingual extensions, but the lexicon's manual treatment of culturally specific terms highlights a problem that such extensions would need to address: translation alone does not establish equivalent attribute meanings. Candidate questions derived from these resources should be checked for comparable interpretation before their factor structures are compared across languages.

Product identity and chemical interpretation

The sampled product must also correspond to the release that was annotated. Concentrations, flankers, vintages, and reformulations can complicate that correspondence even when the name and house are known. Geier et al. (2024), for example, reported declining patch-test reactions to selected fragrance screening mixtures and allergens over 2012–2021 in a context of restrictions on particular fragrance ingredients. Their clinical findings do not measure perceptual change in commercial perfumes, but they illustrate why ingredient history and the date of a product record matter. A historically named fragrance cannot automatically be treated as an unchanged formulation. The present adjudication policy's emphasis on the specific release and current formulation reduces ambiguity, although annotations still depend on whether sources identify that formulation clearly. Historical associations of Vintage/Waxy consequently should not be interpreted as measured changes in perfume chemistry or as a longitudinal trajectory of scent. Chemical composition and sensitization also require data beyond the descriptive note profiles analyzed here.

The database is also complementary to work that models odor from chemical structure. Tom et al. (2025) extended a molecular odor representation to mixtures and reported that progress was limited chiefly by data, because publicly available mixture-perception datasets remain small and draw on a narrow set of compounds, while the larger mixture records held by fragrance houses are unavailable for research. The present resource covers tens of thousands of named products but contains no formulations, concentrations, or molecular identities, and its ratings describe what sources report rather than what a panel perceived. Connecting the two levels would require product-to-ingredient information that the present sources do not supply. The descriptive dimensions reported here should accordingly not be read as evidence about the chemistry that produces them.

Designing external validation

Finally, external validation should be designed around the analyses the database will support. A probability sample spanning original, newly filled, and unresolved cells would address the largest gaps in current accuracy assessment. Stratification by attribute domain and documentation level could improve precision while preserving known inclusion probabilities. Research on prediction-powered and design-based inference shows how smaller samples of validated observations can be combined with abundant model-generated labels for specified statistical targets (Angelopoulos et al., 2023; Egami, Hinck, Stewart, & Wei, 2023). Confidence-guided allocation can further reduce annotation effort when embedded in an appropriate sampling and estimation procedure (Gligorić et al., 2025). Applying those ideas here would require defining the target quantity and obtaining suitable reference measurements; they do not automatically validate factor scores or compensate for an incorrect reference standard.

The similarity analysis should first distinguish scent-related dimensions from price and market positioning. An odor-focused set of factor scores can be compared with blind sorting or pairwise similarity, while informed product judgments can be modeled using commercial information as additional predictors. Competing representations should be compared on held-out products, with factor weights selected independently of the final test. Simple note overlap and conventional family labels would provide useful benchmarks. This design would establish whether the richer representation adds predictive information, and whether that benefit concerns sensory relationships or informed product evaluation.

Supplementary References

- Angelopoulos, A. N., Bates, S., Fannjiang, C., Jordan, M. I., & Zrnic, T. (2023). Prediction-powered inference. *Science*, 382, 669–674. <https://doi.org/10.1126/science.adi6000>
- Anthropic. (2025). *Claude Haiku 4.5 (claude-haiku-4-5-20251001)* [Large language model]. <https://www.anthropic.com>
- Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5–32. <https://doi.org/10.1023/A:1010933404324>
- Csárdi, G. (2023). *dotenv: Load environment variables from .env* [R package]. <https://CRAN.R-project.org/package=dotenv>
- Dowle, M., & Srinivasan, A. (2025). *data.table: Extension of “data.frame”* [R package]. <https://CRAN.R-project.org/package=data.table>
- Egami, N., Hinck, M., Stewart, B. M., & Wei, H. (2023). Using imperfect surrogates for downstream inference: Design-based supervised learning for social science applications of large language models. *Advances in Neural Information Processing Systems*, 36, 68589–68601. <https://doi.org/10.52202/075280-3000>
- Fox, J. (2022). *polycor: Polychoric and polyserial correlations* [R package]. <https://CRAN.R-project.org/package=polycor>
- Gamer, M., Lemon, J., & Fellows, I. (2019). *irr: Various coefficients of interrater reliability and agreement* [R package]. <https://CRAN.R-project.org/package=irr>
- Geier, J., Schubert, S., Rieker-Schwiebacher, J., Brans, R., Weisshaar, E., Kränke, B., . . . Uter, W. (2024). Declining frequency of sensitization to fragrance mixes I and II: IVDK-data of the years 2012–2021. *Contact Dermatitis*, 90, 470–478. <https://doi.org/10.1111/cod.14493>
- Gligorić, K., Zrnic, T., Lee, C., Candès, E. J., & Jurafsky, D. (2025). Can unconfident LLM annotations be used for confident conclusions? In L. Chiruzzo, A. Ritter, & L. Wang (Eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)* (pp. 3514–3533). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2025.naacl-long.179>
- Gou, Z., Shao, Z., Gong, Y., Shen, Y., Yang, Y., Duan, N., & Chen, W. (2024). CRITIC: Large language models can self-correct with tool-interactive critiquing. In *The Twelfth International Conference on Learning Representations*. <https://openreview.net/forum?id=Sx038qxjek>
- Kaeppler, K., & Mueller, F. (2013). Odor classification: A review of factors influencing perception-based odor arrangements. *Chemical Senses*, 38, 189–209. <https://doi.org/10.1093/chemse/bjs141>
- Kandpal, N., Deng, H., Roberts, A., Wallace, E., & Raffel, C. (2023). Large language models struggle to learn long-tail knowledge. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, & J. Scarlett (Eds.), *Proceedings of the 40th International Conference on Machine Learning* (Vol. 202, pp. 15696–15707). PMLR. <https://proceedings.mlr.press/v202/kandpal23a.html>
- Majid, A., & Burenhult, N. (2014). Odors are expressible in language, as long as you speak the right language. *Cognition*, 130, 266–270. <https://doi.org/10.1016/j.cognition.2013.11.004>
- McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, 1, 30–46. <https://doi.org/10.1037/1082-989X.1.1.30>
- Moore, T. M. (2025). *Using large language models to obtain quantitative data* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/aqu9x_v1
- Moore, T. M. (2026a). *A language model generates seven million phenotype estimates across the tree of life* [Manuscript submitted for publication]. Department of Psychiatry, Perelman School of Medicine, University of Pennsylvania.

- Moore, T. M. (2026b). *PharmaDive: A language model generates an uncertainty-graded property matrix for more than 16,000 compounds* [Manuscript submitted for publication]. Department of Psychiatry, Perelman School of Medicine, University of Pennsylvania.
- Nikitina, L. (2024). Fragrance lexicon for analysis of consumer-generated perfume reviews in Russian and English. *MethodsX*, 12, Article 102627. <https://doi.org/10.1016/j.mex.2024.102627>
- Olsson, U. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika*, 44, 443–460. <https://doi.org/10.1007/BF02296207>
- Ooms, J. (2014). *The jsonlite package: A practical and consistent mapping between JSON data and R objects* (arXiv:1403.2805) [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.1403.2805>
- OpenAI. (2026). *ChatGPT* [Large language model]. <https://chatgpt.com>
- Revelle, W. (2025). *psych: Procedures for psychological, psychometric, and personality research* [R package]. Northwestern University. <https://CRAN.R-project.org/package=psych>
- Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, 34(4, Pt. 2), 1–97.
- Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86, 420–428. <https://doi.org/10.1037/0033-2909.86.2.420>
- Slowikowski, K. (2024). *ggrepel: Automatically position non-overlapping text labels with “ggplot2”* [R package]. <https://CRAN.R-project.org/package=ggrepel>
- Tian, K., Mitchell, E., Zhou, A., Sharma, A., Rafailov, R., Yao, H., . . . Manning, C. D. (2023). Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 5433–5442). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.emnlp-main.330>
- Tom, G., Ser, C. T., Rajaonson, E. M., Lo, S., Park, H. S., Lee, B. K., & Sanchez-Lengeling, B. (2025). Does this smell the same? Learning representations of olfactory mixtures using inductive biases. *Machine Learning: Science and Technology*, 6, Article 035063. <https://doi.org/10.1088/2632-2153/adfffc>
- Velicer, W. F. (1976). Determining the number of components from the matrix of partial correlations. *Psychometrika*, 41, 321–327. <https://doi.org/10.1007/BF02293557>
- Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis* (2nd ed.). Springer. <https://doi.org/10.1007/978-3-319-24277-4>
- Wickham, H. (2023). *httr: Tools for working with URLs and HTTP* [R package]. <https://CRAN.R-project.org/package=httr>
- Wilcox, R. R. (2021). *Introduction to robust estimation and hypothesis testing* (5th ed.). Academic Press.
- Wright, M. N., & Ziegler, A. (2017). ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, 77(1), 1–17. <https://doi.org/10.18637/jss.v077.i01>
- xAI. (2026a). *Grok 4.3* [Large language model]. <https://x.ai>
- xAI. (2026b). *Grok 4.6* [Large language model]. <https://x.ai>
- Xiong, M., Hu, Z., Lu, X., Li, Y., Fu, J., He, J., & Hooi, B. (2024). Can LLMs express their uncertainty? An empirical evaluation of confidence elicitation in LLMs. In *The Twelfth International Conference on Learning Representations*. <https://openreview.net/forum?id=gjeQKFxFpZ>