aigora

Sensometrics 2026 Valencia

Is more thinking always better?

Think again.

Less reasoning, better sensory predictions in sensometrics.

John Ennis · Thierry Worch · Benjamin Mahieu · Aigora · FrieslandCampina · Oniris Nantes

Sensometrics 2026 · Valencia · Thu 21 May, 10:00

Sensory and hedonic judgment is a System 1 task.

Reasoning-heavy LLMs are System 2 tools. *The mismatch costs you.*

The iron triangle of consumer research.

Research always pays somewhere.

Speed, cost, and evidence quality all matter.
Push one corner too hard and another corner moves.

LLMs change where the trade-off lands.

They don't remove the trade-off.
They move it into model choice, structured output and validation.

But which LLM?

The industry often assumes more reasoning means better judgment.
We tested that assumption.

Mahieu et al. (2022): a home-use cooked ham study.

1 **Home-use purchase**

Consumers bought hams from the 30-product list

2 **TimeSens check-in**

EAN entry and package photos before and after opening

3 **Free-comment tasks**

Visual, texture and flavor, always in that order

4 **Liking and JAR**

0 to 10 liking, then salt, fat, tenderness and color

5 **Text analysis**

IRaMuTeQ, descriptor tables, mixed models and MR-CA

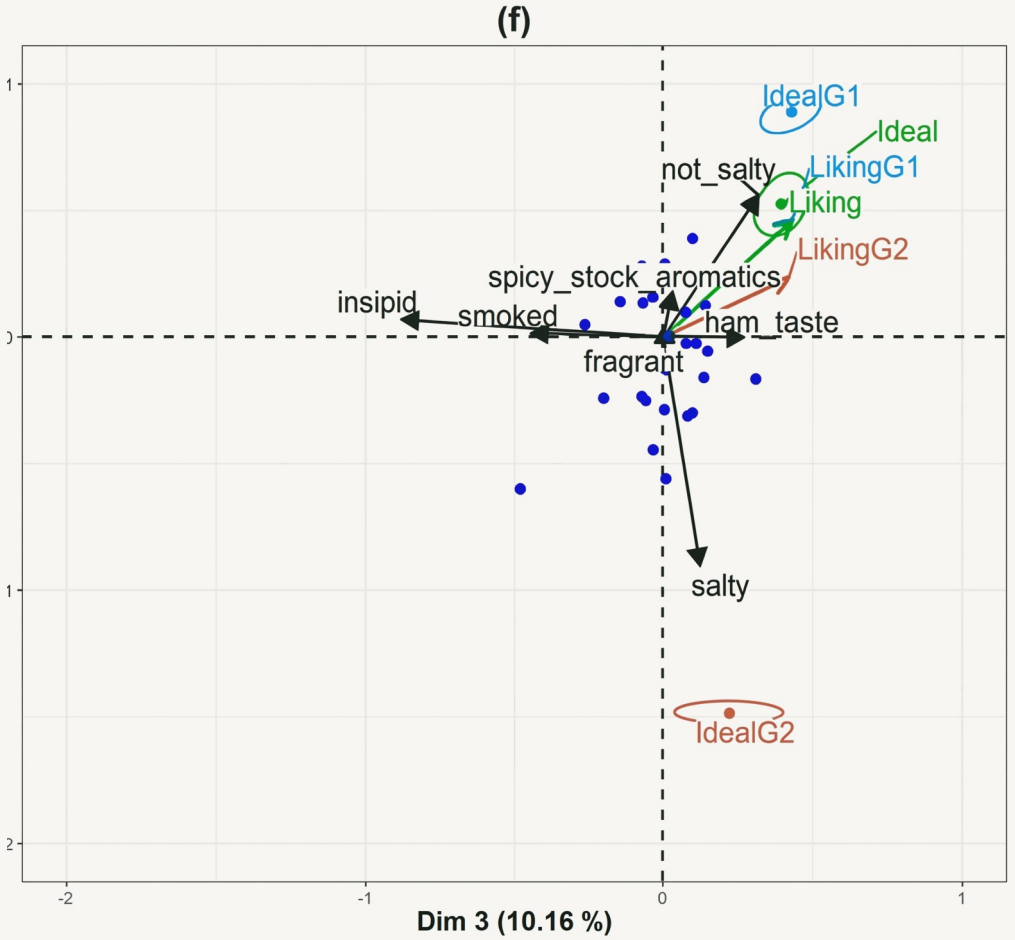
The 2022 setting was home-use.

Regular consumers bought listed hams, tasted them at home, and entered free comments, liking and JAR data online.

The bridge to our work is the raw consumer language.

Mahieu, Visalli & Schlich (2022), Food Quality and Preference, 96, 104389.

Mahieu et al. (2022) provide a reference sensory map.



A known sensory structure.

MR-CA maps descriptor citations into a sensory space. Liking and ideal positions are projected afterward.

30
commercial cooked hams

483
French consumers

2,758
home-use evaluations

Test for the new pipeline

Can three LLM similarity scores predict liking and still point back to the same sensory hierarchy?

Mahieu, Visalli & Schlich (2022), Food Quality and Preference, 96, 104389.

Our workflow: LLM-as-judge, then TabPFN.

1 Input

actual and ideal free-comments

2 Score

visual, texture and flavor similarity

3 Predict

TabPFN estimates 0 to 10 liking

Held out by consumer. No leakage across train and test.

Related LLM scoring example

Maier et al. (2025) use LLMs to elicit Likert ratings from text. arXiv:2510.08338.

TabPFN v2: fast tabular regression

Hollmann et al. (2025). TabPFN v2: accurate predictions on small tabular data. Nature.

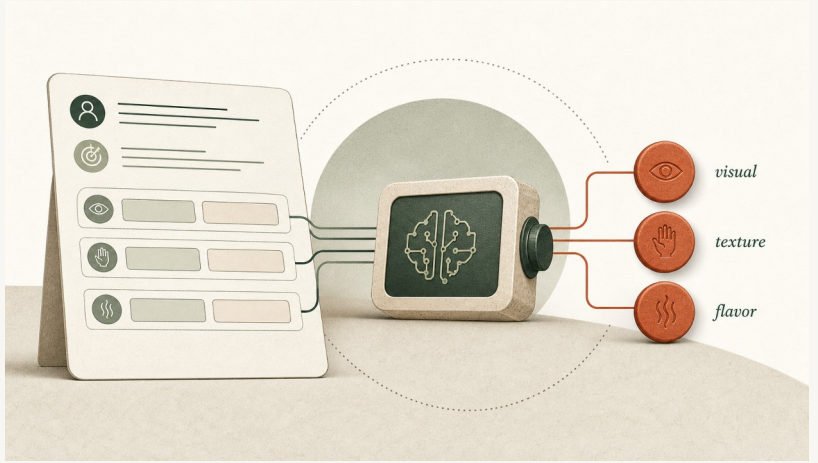
Scoring was a fixed prompt per evaluation.

Same role, same six text fields, same schema. The liking score stayed out.

Prompt template

1	**Role + task**	"You are a sensory + consumer scientist." Compare ACTUAL ham to IDEAL; score alignment.
2	**Six fields**	IDEAL: IdealVisual, IdealTexture, IdealFlavor ACTUAL: DescriptionVisual, DescriptionTexture, DescriptionFlavor
3	**Rules**	French unchanged; sensory ham descriptions only. Ignore non-sensory comments; no liking score; blanks = "(empty)".
4	**Scale + return**	Inject 4-, 6-, or 8-point labels. Return JSON only: {"visual":X, "texture":X, "flavor":X}

6-point example — 1 Extremely Different ... 6 Extremely Similar



Run contract

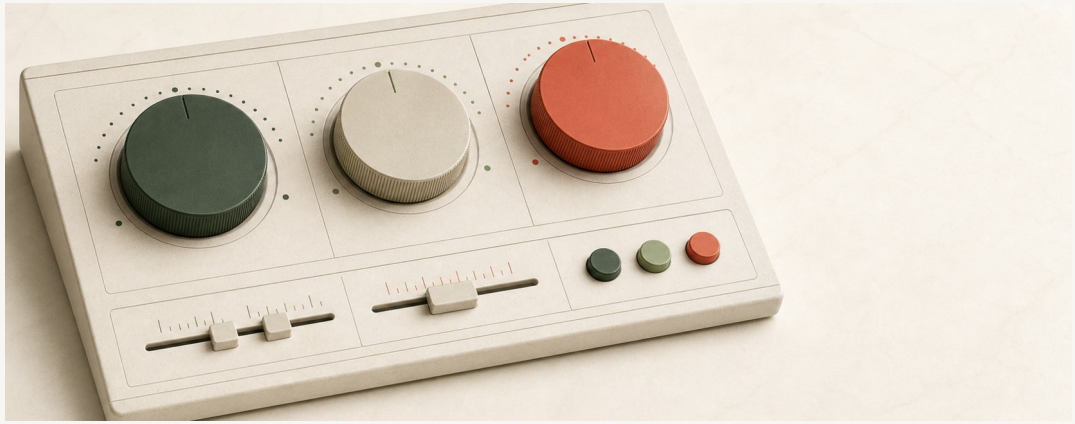
Row unit	one consumer-product evaluation
API	responseMimeType: application/json
Parse	JSON first, regex fallback
Failure	3 retries → midpoint + error flag

Example output — **H041 / J09 | visual 4 texture 5 flavor 6**

Only visual, texture, and flavor scores enter TabPFN; liking is used later as the target.

Which knobs matter in an LLM scoring run?

The dataset, task and JSON schema stayed fixed. We varied the model, scale and temperature.



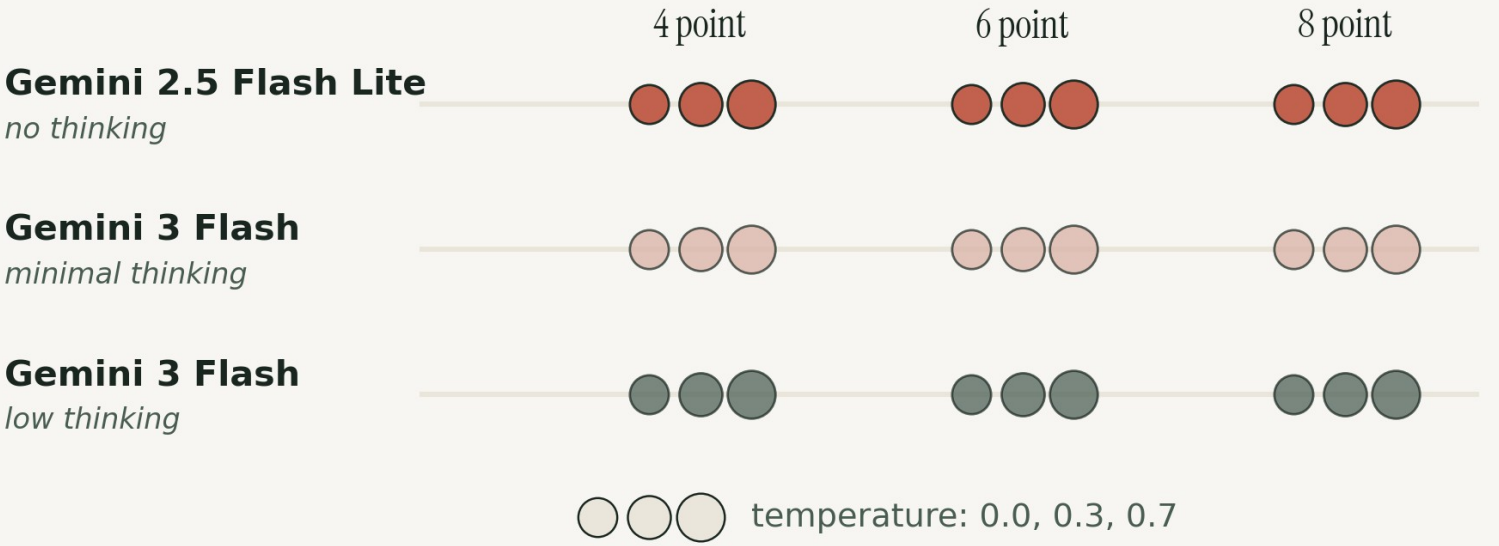
Temperature	Sampling randomness. Tested: 0.0, 0.3, 0.7.
topP	Fixed. Keeps tokens until probability reaches p.
topK	Fixed. Limits sampling to the K most likely tokens.
Candidate count	Fixed. Number of independent answers requested.

Varied in this study	Model family and thinking level. Similarity scale: 4, 6 or 8. Temperature: 0.0, 0.3 or 0.7.
Checked downstream	TabPFN, XGBoost, Gradient Boosting, Ridge and Random Forest after LLM scoring.

27 configurations, one held-out test.

27 scoring passes

Rows are model settings. Columns are similarity scales. Dots are temperatures.

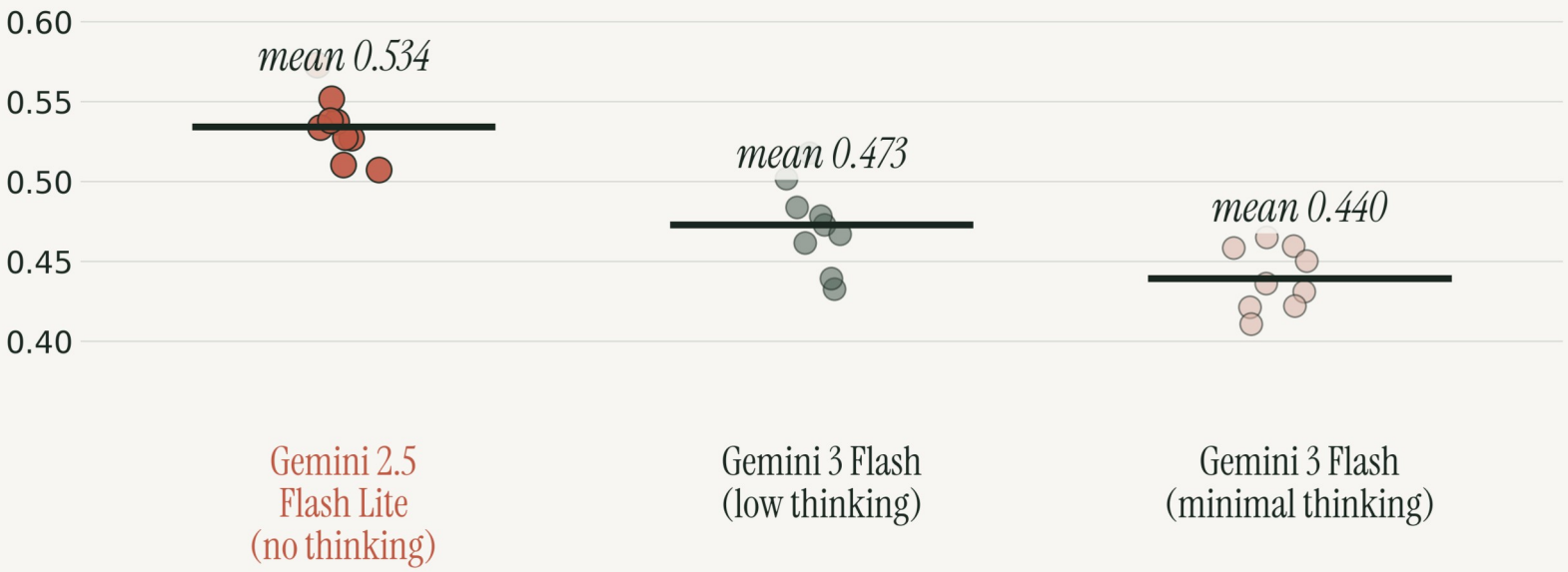


Full data	No leakage	Uncertainty	ML check
2,758 evaluations	Group split by consumer	200 bootstrap resamples	5 downstream learners

Flash Lite wins this benchmark.

R-squared across the full 3 × 3 × 3 grid

Nine points per model family; Flash Lite stays above both Gemini 3 thinking modes.



Best config

Flash Lite

6-point scale · t = 0.7

R-squared 0.573

MAE 1.22

All 9 Flash Lite configs beat all 18 Gemini 3 Flash configs on MAE.

Overthinking can make a simple liking case look complicated.

Same row, same prompt, same 1 to 6 scale: Flash Lite and Gemini 3 low read the texture signal differently.

Actual comment, consumer H041, liking = 8.90

"The ham breaks apart, but it has a good texture."
"Very good, not too salty."

Ideal comment

"A slightly thick slice that doesn't fall apart in the mouth, without being too dry."
"A light smoky taste, not too salty."

Model scores

visual / texture / flavor

Flash Lite	*no thinking*	4/5/6
Gemini 3 Flash	*low thinking*	4/2/4

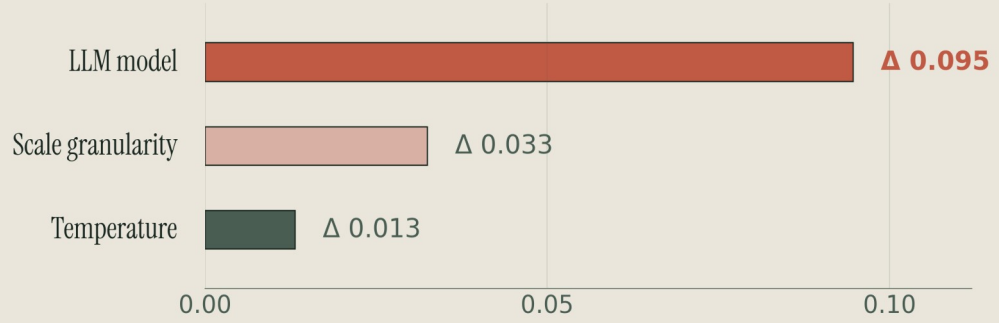
My read: Gemini 3 punishes the literal mismatch.
Flash Lite keeps the high-liking signal.

Same pattern elsewhere: liking 10.00, Flash Lite 3/4/5 vs Gemini 3 2/2/3; liking 7.62, Flash Lite 5/6/5 vs Gemini 3 4/3/4.

Why it's not noise.

Drivers of variance

Range in mean R-squared across factor levels in the 27-run grid.

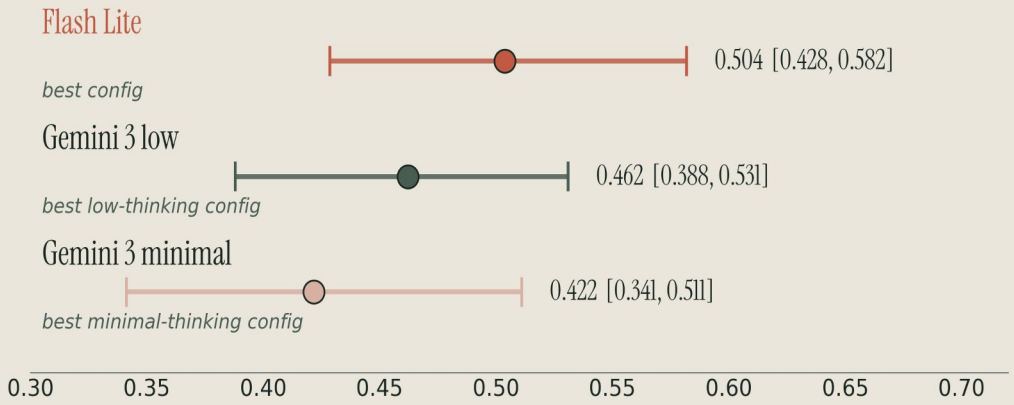


ANOVA readout

LLM model	11 %	F = 18.42, p < 0.0001
Scale granularity	3 %	F = 4.12
Temperature	2 %	F = 0.18, n.s.

The performance gap is robust

Mean R-squared with 95 % CI, 200 resamples.

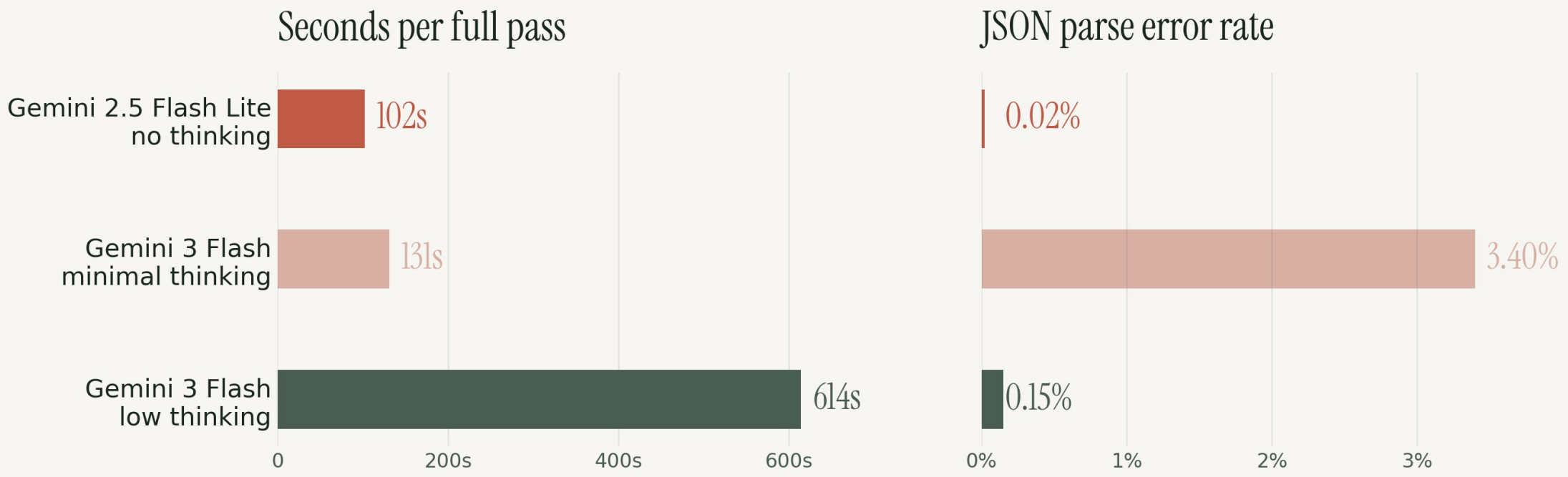


Probability of outperforming

99.3 %
Flash Lite > Gemini 3 minimal

91.2 %
Flash Lite > Gemini 3 low

Reasoning costs more and breaks format more.



The reasoning model either takes longer, breaks format more often, or both.

Why TabPFN lets us run the full loop.

Accuracy was nearly tied.

Mean R-squared across the LLM configurations

TabPFN v2	0.476
XGBoost	0.474
Gradient Boosting	0.473
Ridge	0.469
Random Forest	0.444

TabPFN won 74 % of configs.

The accuracy case is modest. The workflow case is strong.

The choice was operational.

1 **No tuning**

No separate hyperparameter search after each LLM run.

2 **One forward pass**

A fast tabular model kept the comparison loop short.

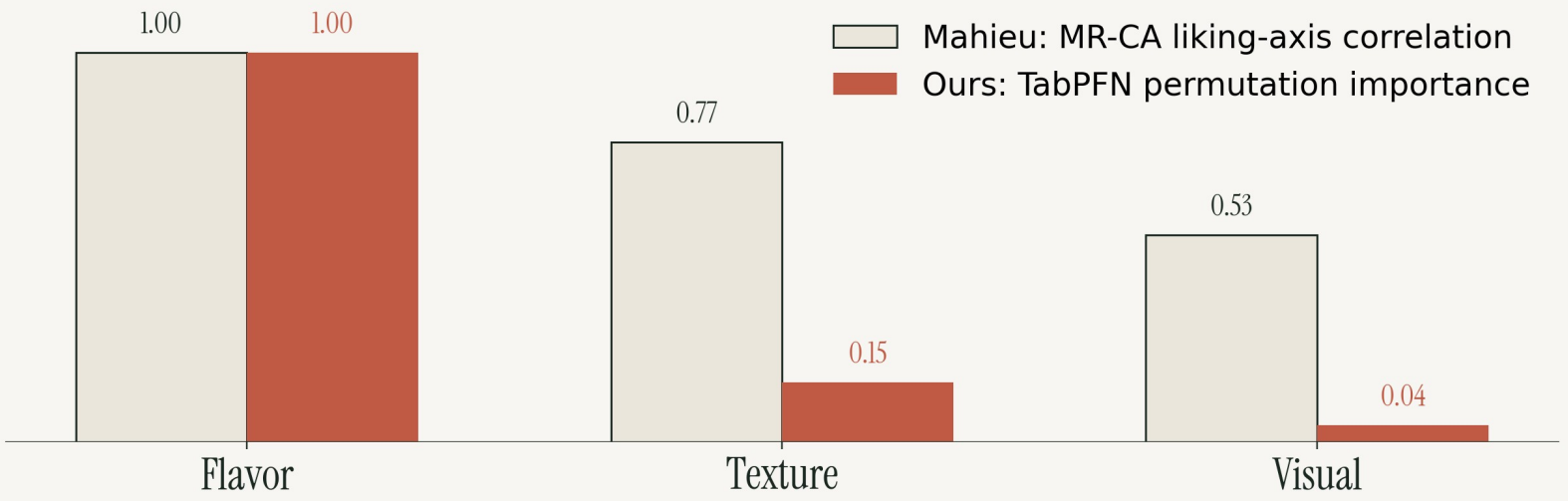
3 **Bootstrap-friendly**

The 200-resample uncertainty check stayed practical.

The model recovers the sensory hierarchy.

Different statistics, same sensory ordering

Each series is normalized within method so flavor = 1.00.



Flavor first. Texture second. Visual last.

Sensory sanity check

Flavor stays first.

Cream: Mahieu et al. (2022)
average absolute weighted correlation
of mean liking with MR-CA axes.

Terracotta: this study
TabPFN permutation importance
of the three LLM similarity scores.

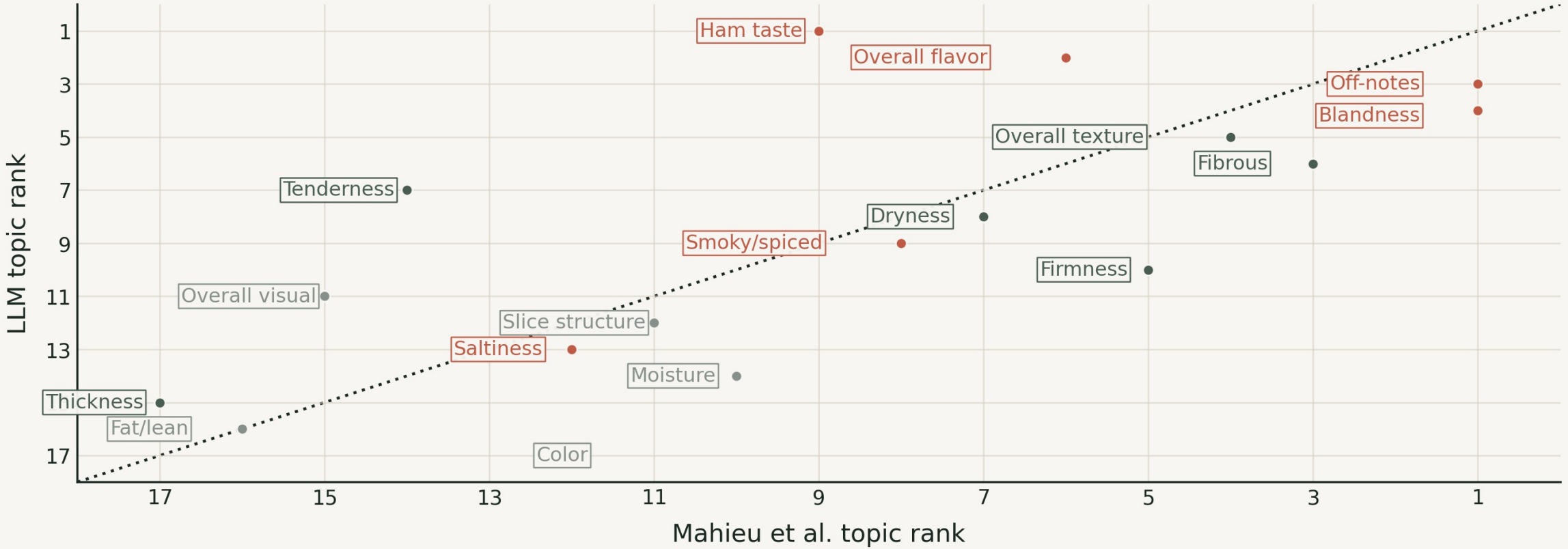
The claim is rank order, not equal units.

The topic-level check preserves named sensory priorities.

The broad modality result also holds when we map 17 LLM topic-alignment scores back to Mahieu et al. descriptor families.

17 mapped topic families: lower rank means stronger liking driver

Spearman rho 0.715 | Kendall tau 0.519 | 9 of top 10 overlap



Models move. The benchmark stays useful.

This is a benchmark, not a permanent ranking.

Gemini 3.5 Flash launched on 19 May 2026.
Model releases move the baseline. Rerun the grid.

1 **Rerun after major releases.**
Check whether the System 1/System 2 gap holds.

2 **Extend the sensory corpus.**
Cooked ham first, then fragrance, texture and aesthetics.

3 **Test smaller local models.**
Gemma 4 E2B/E4B and T5Gemma 2 are candidates once tuned.

Benchmark loop

1 New model release
2 Rerun scoring grid
3 Compare bootstrap CIs
4 Update recommendation

What still needs checking

Still monitor: one product domain, intra-Gemini comparison, and JSON/schema failures that can drop or distort LLM scores.

Three takeaways.

1 The sensometrician keeps checking the model.

Releases move the baseline. Rerun the benchmark when the model family changes.

2 Match LLM cognition to task cognition.

System 1 task. Don’t pay for System 2.

3 The iron triangle bends.

Speed, quality, and low cost are reachable together when model choice fits the work.

John Ennis

aigora.ai