



LLM-Driven Annotation: A Scalable Framework for Automated Multi-Label Coffee Flavor Classification

Prasara Jakkaew¹, Nacha Chondamrongkul² and Paweena Suebsombut³

ABSTRACT

The subjective and unstructured nature of coffee tasting notes creates a significant data annotation bottleneck, limiting the application of computational methods in sensory science. This study introduces a novel end-to-end framework for automating multi-label coffee flavor classification, integrating LLM-driven annotation with transformer-based classification and local interpretability analysis. Google's Gemini-1.5-Flash was employed to perform evidence-based annotation on 8,327 expert evaluations authored by certified Q Graders, generating a high-quality training dataset across 17 flavor categories. Critically, the Q Grader provenance of the source texts enables a reverse-mapping validation framework: statistically significant and directionally coherent correlations between LLM-derived labels and expert quantitative scores (Floral $r = 0.32$, Roasted $r = -0.25$, all $p < 0.001$) provide implicit ground truth without requiring separate annotation effort. Reliability analysis revealed substantial consistency for concrete physical descriptors (mean $\kappa = 0.68$) but notably lower agreement for abstract sensory concepts such as Mouthfeel ($\kappa = 0.10$), identifying two distinct reliability regimes that define the framework's operational boundaries. A fine-tuned BERT model trained on these annotations outperformed a TF-IDF baseline, achieving a Micro F1-score of 0.9164 versus 0.8763 and a Hamming Loss of 0.0640 versus 0.0969. To address severe class imbalance (up to $125\times$), Focal Loss ($\gamma = 2.0$) combined with per-class threshold optimization successfully recovered detection of rare defect categories, improving Macro-F1 from 0.656 to 0.719. Local interpretability analysis via LIME further confirmed that model predictions align with domain-expert sensory reasoning. These results demonstrate that an LLM-driven annotation pipeline offers a scalable, transparent, and effective solution to the data bottleneck in sensory science, establishing a robust methodological foundation for interpretable classification across other sensory-driven domains.

Article information:

Keywords: Coffee Flavor Profiling, Explainable AI (XAI), LLM-driven Annotation, Multi-label Classification, Natural Language Processing (NLP), Sensory Science

Article history:

Received: October 14, 2025

Revised: March 5, 2026

Accepted: April 23, 2026

Published: April 30, 2026

(Online)

DOI: 10.37936/ecti-cit.2026202.264386

1. INTRODUCTION

In the specialty coffee industry, tasting notes such as “*berry-like acidity with floral undertones and a hint of dark chocolate*” play a critical role. These descriptors serve as a shared sensory language that shapes pricing, quality communication, and consumer perception across global supply chains. They define customer expectations and foster trust among farmer, producers, traders, and consumers. Typically, trained Q Graders generate these notes during

standardized cupping sessions [1], linking subjective sensory impressions to objective quality assessments. Consequently, tasting notes exert substantial influence on market acceptance and help coffee producers differentiate their products [2], [3].

Despite their importance, sensory descriptions remain largely unstructured, inconsistently annotated, and highly subjective [2], [4]. This variability creates a major data-annotation bottleneck that impedes the integration of tasting notes into digital

^{1,2,3}The authors are with the School of Applied Digital Technology, Mae Fah Luang University, Thailand, Email: prasara.jak@mfu.ac.th, nacha.cho@mfu.ac.th and paweena.sue@mfu.ac.th

³Corresponding author: paweena.sue@mfu.ac.th

cupping platforms, blockchain-based traceability systems, and data-driven quality-control pipelines [5], [6]. While structured frameworks such as the Specialty Coffee Association (SCA) Flavor Wheel [7], [8] and the World Coffee Research (WCR) Sensory Lexicon [9] provide standardized terminology, the central challenge persists: manually aligning vast free-text archives with these lexicons remains labor-intensive and time-consuming [10], [11].

Recent advances in Natural Language Processing (NLP) offer a means to overcome this data bottleneck. Transformer-based architectures—particularly Bidirectional Encoder Representations from Transformers (BERT) demonstrate strong capacity to capture the semantic and syntactic complexity of human language, making them well-suited for modeling nuanced sensory descriptions [12–14]. However, their effectiveness depends critically on the availability of large, high-quality labeled datasets—a resource still scarce in sensory science. We posit that integrating state-of-the-art classification models with generative Large Language Models (LLMs) constitutes a paradigm shift: such integration enables both the classification of sensory text and the creation of large-scale, structured, evidence-based datasets essential for robust model development.

This study introduces a comprehensive framework for automated coffee-flavor profiling within this emerging paradigm. The methodology begins by employing Google’s Gemini-1.5-Flash LLM to perform automated multi-label annotation of coffee tasting notes, requiring the model to generate explicit textual evidence for each defined flavor label. Using the resulting high-quality dataset, we conduct a comparative analysis demonstrating that a fine-tuned BERT model significantly outperforms a traditional Term Frequency-Inverse Document Frequency (TF-IDF) baseline in this complex multi-label classification task. To ensure transparency, we apply Local Interpretable Model-agnostic Explanations (LIME) to clarify BERT’s decision process and relate its explanations to the evidence originally produced by the LLM. Finally, we validate the model’s textual comprehension by correlating its predicted flavor profiles with quantitative sensory evaluations from professional cuppers, establishing a strong foundation for future automated sensory-analysis research.

2. RELATED WORK

2.1 The Challenge of Structuring Sensory Data

Structured lexicons have long been employed in the specialty coffee industry to systematize the inherently subjective realm of sensory perception. The SCA Flavor Wheel and the WCR Sensory Lexicon exemplify standardized frameworks that provide a common vocabulary, facilitating communication among cuppers, roasters, and consumers. These tools con-

tinue to evolve; for example, Spencer *et al.* [11] applied empirical methods to refine the flavor wheel so it aligns more closely with expert perceptual classifications. Nonetheless, these developments underscore an essential truth: flavor language is culturally dependent. The emergence of localized lexicons, such as those created for Japanese tasters [15], demonstrates how sensory descriptors are deeply embedded within regional and linguistic contexts.

Real-world tasting notes further complicate this landscape. A single coffee rarely corresponds to one attribute; rather, it embodies a constellation of simultaneous flavors, aromas, and tactile sensations—making flavor classification inherently a multi-label task [13]. Moreover, sensory descriptors often exhibit hierarchical relationships, such as “citrus” representing a subcategory of “fruity,” as reflected in the design of the SCA Flavor Wheel. Although models capable of handling hierarchical multi-label structures have grown increasingly sophisticated [16], a major bottleneck remains: manually mapping large volumes of unstructured narrative text onto these complex taxonomies is prohibitively slow and costly. This scarcity of structured data has, until now, constrained the application of computational methods for large-scale sensory analysis.

2.2 Computational Approaches to Sensory Text Classification

Early computational studies of sensory text employed techniques such as Latent Semantic Analysis (LSA) to uncover patterns in domains like wine tasting [17]. While such approaches, together with traditional bag-of-words models, offered useful insights, they failed to capture the syntactic complexity and nuanced contextual cues inherent in tasting notes. In this study, we establish a baseline using a standard TF-IDF approach, providing a clear and interpretable benchmark against which to evaluate the performance gains achieved by more sophisticated models.

The emergence of transformer-based architectures—most notably BERT—marked a major advancement. By learning deep contextual representations of language, these models have significantly improved performance across food and beverage applications. Domain-specific variants such as DILBERT for aspect extraction [18] and AgriBERT for nutritional analysis [19] demonstrate strong domain adaptability. Their consistent superiority over traditional methods [18], [20] suggests that similar techniques could yield substantial benefits for coffee-flavor analysis as well.

However, the performance of transformer models is inherently tied to the quality and scale of their training data. Their sophisticated contextual understanding can only be realized when large, accurately labeled datasets are available. Yet, the persistent data-

annotation bottleneck continues to impede progress, creating a critical gap between modeling capability and practical deployment.

2.3 The Emerging Paradigm: LLM-Driven Annotation and the Imperative for Explainability

The advent of powerful instruction-tuned LLMs introduces a transformative approach to overcoming the data-annotation bottleneck. A rapidly expanding body of research now explores LLMs as scalable, high-quality annotators capable of performing complex labeling tasks with minimal human oversight—a practice often described as programmatic labeling or weak supervision [21][35]. By exchanging modest human effort in prompt design for large quantities of labeled data, this paradigm enables the creation of extensive, nuanced datasets that have long been absent from sensory-classification research.

Yet this increasing automation heightens the need for transparency and accountability. Trust in model outputs requires mechanisms for evaluating their underlying reasoning. In this context, Explainable AI (XAI) transitions from a valuable accessory to an essential component of scientific inquiry. Techniques such as LIME [22] and SHAP [23] are now fundamental for interpreting the predictions of black-box models in high-impact domains such as agriculture [24] and coffee science [25], enabling experts to verify whether model decisions are grounded in relevant features.

Our research lies at the intersection of these two emerging directions. Although prior studies have applied transformer models to sensory text or employed XAI for interpretability, a critical gap persists. No existing work has developed an end-to-end framework that (1) uses an LLM for fine-grained, evidence-based multi-label annotation; (2) trains a state-of-the-art transformer model on this data; and (3) closes the loop by employing XAI to connect the model’s predictions directly to the textual evidence generated by the LLM annotator. This study addresses that gap, presenting a framework that is not only accurate but also transparent and verifiable throughout.

3. MATERIALS AND METHODS

The generative artificial intelligence (GenAI) model — Google’s Gemini-1.5-Flash — served as the core component of the data-annotation pipeline. It was employed to perform multi-label classification and extract textual evidence from coffee reviews, thereby producing the training dataset for downstream deep-learning models. Figure 1 presents an overview of the methodological framework, which systematically supports the development, comparison, and interpretation of models for multi-label coffee-flavor classification. The process begins with data

acquisition and the construction of a robust flavor-category schema grounded in standardized sensory lexicons. Two complementary annotation strategies were implemented and compared: (1) a generative approach leveraging a large language model (LLM), and (2) a rule-based approach employing deterministic text matching. The resulting annotations were then used to train two modeling pathways — a baseline TF-IDF model and a fine-tuned BERT classifier — which were evaluated for predictive accuracy, interpretability, and alignment with quantitative sensory attributes.

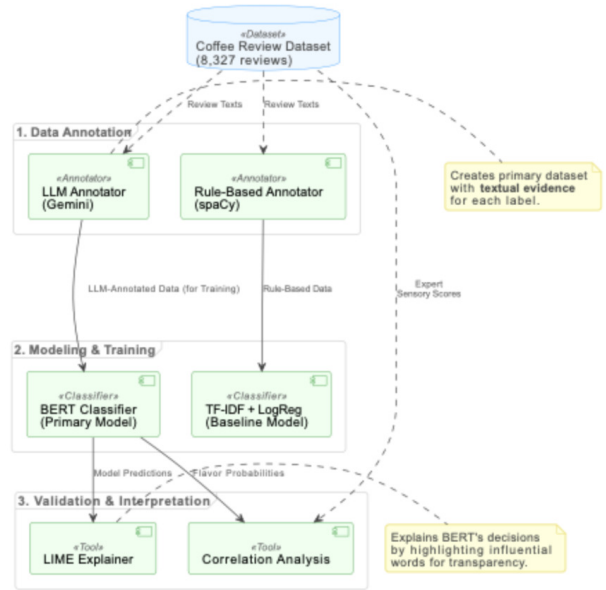


Fig.1: Flavor classification framework with three stages: annotation, training, and validation.

3.1 Dataset and Preprocessing

This study utilized a dataset comprising 8,327 expert coffee evaluations, primarily contributed by certified Q Graders, obtained from the publicly available Coffee Review database [26]. The data were used exclusively for academic research and analysis. For each record, the Blind Assessment and Notes fields were extracted and concatenated into a single text column to serve as the primary corpus for flavor profiling. To prevent processing errors, missing values were replaced with empty strings. Quantitative attributes—including Acidity, Aroma, and Final Score—were retained for the final validation analysis. Each review was assigned a unique identifier to ensure traceability throughout the processing pipeline. Table 1 summarizes the key characteristics of the dataset.

3.2 Dataset Characteristics

Prior to annotation, an exploratory data analysis (EDA) was conducted to characterize the dataset.

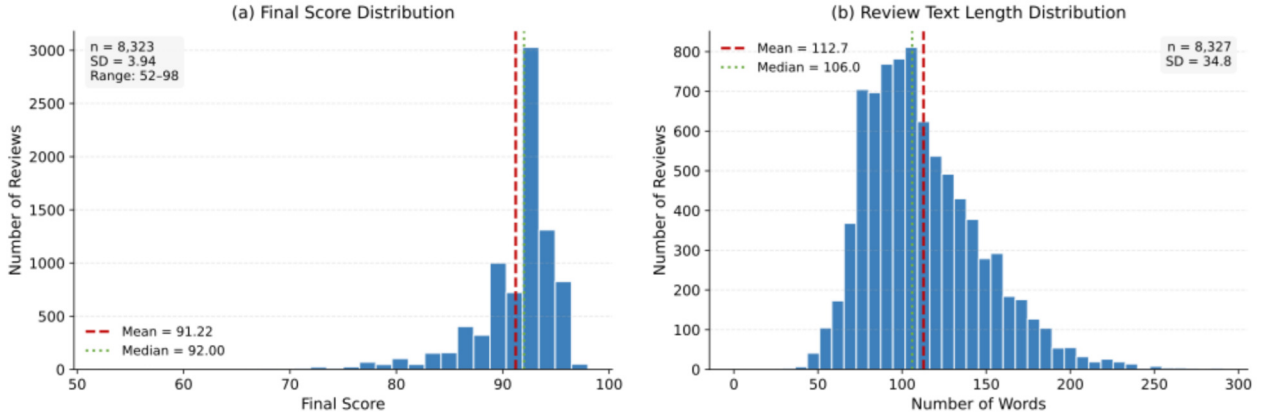


Fig.2: Distribution of final scores in the coffee review dataset.

Table 1: Dataset Characteristics.

Characteristic	Value
Total Number of Reviews	8,327
Number of Unique Coffees	7,023
Number of Unique Roasters	1,534
Average Review Text Length (words)	112.73

The 8,327 reviews contained an average of 112.7 words per review (median: 106.0 words). Final Scores ranged from 52 to 98 (mean = 91.22, SD = 3.94), exhibiting a left-skewed distribution consistent with specialty-coffee selection bias toward high-quality samples (Figure 2). Missing values were minimal for primary text fields (Blind Assessment: 0.01%; Notes: 0.02%) but more prevalent for Acidity/Structure (15.7%) and Aftertaste (10.5%), which were handled via pairwise deletion in all correlation analyses. The label distribution revealed a severe class imbalance ratio of $113.6\times$ across the full dataset (Mouthfeel: 8,178; Cereal: 72), rising to $125\times$ within the training partition (Mouthfeel: 6,543; Cereal: 53) as illustrated in Figure 4.

3.3 Flavor Category Construction

To construct a reproducible and domain-relevant taxonomy of flavor labels, 17 parent categories were established through an evidence-based, systematic integration of two authoritative sources: the WCR Sensory Lexicon, noted for its scientific rigor, and the SCA Flavor Wheel, recognized for its hierarchical organization and widespread industry adoption.

These resources were merged to produce a consolidated list of 139 sensory descriptors, which were harmonized using four structured consolidation strategies:

- **Direct Mapping** — conceptually equivalent terms were merged into a single parent category (e.g., Fruity included both WCR’s Strawberry and SCA’s Berry).
- **Hierarchical Consolidation** — terms were ar-

ranged at the most appropriate level of abstraction, such as establishing a distinct Dried Fruit category to separate these notes from Fresh Fruit.

- **Conceptual Grouping** — descriptors reflecting similar sensory experiences were unified even when organized differently across sources (e.g., fermentation- and acidity-related terms grouped under Sour/Fermented; tactile descriptors grouped under Mouthfeel).
- **Defect and Off-Note Grouping** — unfavorable descriptors such as Stale/Papery and Chemical were assigned to dedicated defect categories.

The finalized set of 17 parent categories comprises Fruity, Floral, Sweet, Nutty, Chocolate, Spicy, Sour/Fermented, Green/Vegetative, Roasted, Cereal, Savory, Dried Fruit, Earthy, Chemical, Stale/Papery, Mouthfeel, and Other.

3.4 Comparative Annotation Strategies

The systematic comparison of annotation methodologies constituted a core component of this study. Accordingly, two complementary pipelines were designed to generate and assign flavor labels to the dataset: a rule-based deterministic approach and a LLM-driven generative approach.

3.4.1 Rule-Based Annotation

A deterministic baseline was implemented using spaCy’s PhraseMatcher [27]. A custom lexicon, designated FLAVOR_LEXICON_SPACY, was developed by mapping synonyms and descriptive terms to the 17 parent flavor categories. Review texts were preprocessed through normalization—lowercasing, whitespace removal, and punctuation cleaning—before applying phrase matching to produce a binary label vector for each review.

3.4.2 LLM-Powered Generative Annotation

In parallel, a generative annotation pipeline was developed using Google’s Gemini-1.5-Flash model, se-

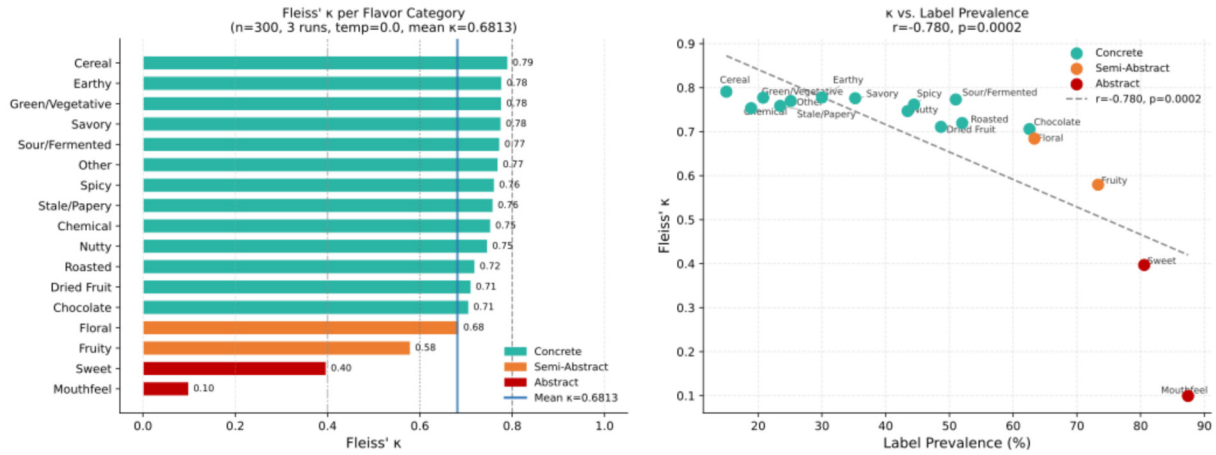


Fig.3: Fleiss’ κ per flavor category and its relationship with label prevalence ($n = 300$, 3 runs, temperature = 0.0). Mouthfeel appears as a notable outlier — high prevalence (87%) yet near-zero agreement ($\kappa = 0.10$), indicating systematic over-assignment bias.

lected for its superior performance on long-context reasoning and structured text generation tasks. However, to ensure reproducibility, future work will explore open-weight alternatives that can replicate similar capabilities within a transparent environment. Through targeted prompt engineering, the model was guided to produce both categorical labels and corresponding textual evidence supporting each classification decision. The structured output was programmatically parsed to yield two artifacts per review: (1) a binary label vector, and (2) an LLM-generated evidence dictionary linking each flavor category to its relevant textual span. This evidence formed the foundation for the interpretability analyses described in later sections.

3.5 Reliability of the Annotation Process

To evaluate the reproducibility of the generative annotation process, inter-rater reliability was measured using a random subset of 300 reviews, each annotated three independent times under identical API parameters (temperature = 0.0). Fleiss’ Kappa (κ) was calculated for each flavor category, treating each model run as an independent rater [36].

As shown in Table 2, the mean κ value was 0.68, representing substantial agreement under the Landis and Koch interpretation. Several categories—including Cereal ($\kappa = 0.79$), Green/Vegetative ($\kappa = 0.78$), and Earthy ($\kappa = 0.78$)—approached almost-perfect reliability. In contrast, Sweet ($\kappa = 0.40$) and Mouthfeel ($\kappa = 0.10$) exhibited considerably lower agreement, highlighting the difficulty of consistently identifying abstract or context-dependent sensory descriptors.

It is important to distinguish reliability from validity in this context. Fleiss’ κ measures the consistency of the annotation process across independent LLM runs, not the accuracy of labels relative to a hu-

Table 2: Fleiss’ Kappa Scores for Inter-Rater Reliability of LLM Annotation.

Flavor Category	Fleiss’ Kappa (κ)
Cereal	0.7907
Green/Vegetative	0.7771
Earthy	0.7775
Sour/Fermented	0.7742
Savory	0.7756
Other	0.7693
Spicy	0.7609
Stale/Papery	0.7586
Chemical	0.7532
Nutty	0.7474
Roasted	0.7203
Dried Fruit	0.7100
Chocolate	0.7042
Floral	0.6824
Fruity	0.5806
Sweet	0.3979
Mouthfeel	0.0993
Average	0.6811

man gold standard. Notably, Mouthfeel exhibited $\kappa = 0.10$ despite being assigned to 98.2% of all reviews, suggesting that its high prevalence may reflect systematic over-assignment rather than genuine semantic consensus. A post-hoc analysis further revealed a strong negative correlation between label prevalence and inter-rater agreement (Pearson $r = -0.780$, $p = 0.0002$; Figure 3), identifying two distinct reliability regimes: (1) concrete physical descriptors ($\kappa \geq 0.70$) anchored to explicit lexical terms—such as cedar, cashew, and chemical odor—which exhibit Substantial agreement; and (2) abstract or tactile categories ($\kappa < 0.40$), namely Sweet and Mouthfeel, which depend on holistic sensory interpretation and are often

expressed indirectly (e.g., silky finish, round body) rather than through explicit flavor terms. Accordingly, conclusions drawn from Mouthfeel and Sweet labels should be interpreted with caution, and the framework is considered most reliable for the 15 categories with $\kappa \geq 0.58$.

Future work will construct a human-annotated gold standard ($n \geq 500$ reviews, Q Grader-certified) to directly assess annotation validity and further characterize the boundary between reliable and unreliable LLM annotation in sensory science contexts.

3.6 Modeling and Evaluation

Two complementary modeling paradigms were employed to establish both a classical baseline and a deep learning-based solution: (1) a TF-IDF + Logistic Regression model and (2) a fine-tuned BERT classifier. The following subsections describe the pre-processing, model architecture, and training procedures for each, followed by the experimental design and evaluation methodology.

3.6.1 TF-IDF + Logistic Regression

For the classical baseline, review texts were pre-processed using NLTK [28] through tokenization, stopword removal, and punctuation stripping. The cleaned corpus was vectorized via scikit-learn's TfidfVectorizer, configured to extract both unigrams and bigrams, restricted to the top 5,000 features ranked by TF-IDF score. These representations were input into a one-vs-rest logistic-regression classifier trained with class weighting to mitigate label imbalance.

3.6.2 Fine-Tuned BERT

For the deep learning approach, the concatenated review text was tokenized using the bert-base-uncased tokenizer. All sequences were padded or truncated to a uniform length of 256 tokens. The model architecture comprised a pre-trained BERT encoder, a dropout layer (rate = 0.3) for regularization, and a fully connected output layer generating logits for the 17 flavor categories. The network was fine-tuned for 5 epochs using the AdamW optimizer ($lr = 2 \times 10^{-5}$, weight decay = 0.01) with a linear learning-rate scheduler incorporating 10% warmup steps. To address the severe class imbalance — with imbalance ratios of up to $125\times$ for Cereal (53 positive training samples), $86\times$ for Chemical (77 samples), and $76\times$ for Stale/Papery (87 samples) — Focal Loss ($\gamma = 2.0$) [34] was employed in place of standard binary cross-entropy. Focal Loss applies a modulating factor $(1 - p_t)^\gamma$ that down-weights the loss contribution of easy, well-classified examples, thereby directing the training signal toward hard and rare instances.

While Focal Loss improved overall Macro-F1 marginally (+0.002), the three rarest defect categories remained undetected at the standard decision

threshold ($\tau = 0.5$) due to extreme data sparsity. Per-class threshold optimization (Section 3.7) was therefore applied as a complementary strategy, selecting the threshold τ^* that maximizes F1-score for each class independently via the precision-recall curve.

3.6.3 Experimental Setup and Evaluation

Experiments were executed in Google Colab, utilizing Python 3.10 and NVIDIA T4 GPU acceleration. Deep learning models were implemented in PyTorch [29] with the Hugging Face Transformers library [30], while scikit-learn [31] was used for baseline modeling and metric computation. The dataset was partitioned into 80% training and 20% testing subsets using iterative stratification to maintain balanced label distributions across splits. A consistent random seed was applied to all experiments to ensure reproducibility.

Although k-fold cross-validation was not performed due to the prohibitive computational cost of fine-tuning BERT across multiple folds, the stability of results was assessed via bootstrap resampling ($B = 1,000$, seed = 42) applied to the held-out test set, which provides equivalent variance estimation for large test sets [33]. This procedure was applied to compute 95% confidence intervals for all primary evaluation metrics across all model configurations, including both BCE and Focal Loss variants at standard ($\tau = 0.5$) and optimized (τ^*) decision thresholds.

Model performance was evaluated using several multi-label metrics:

1. Hamming Loss ($\downarrow$) – the proportion of labels incorrectly predicted.
2. F1-Score (Micro and Macro) – captures overall performance while weighting rare labels equally; per-label reports were also generated for precision, recall, and F1 across the 17 categories.
3. Label Ranking Average Precision (LRAP) – assesses the model's ability to assign higher ranks to correct labels [32].

3.7 Validation and Interpretability

Beyond standard predictive metrics, we further evaluated the alignment between model reasoning and human sensory expertise through three complementary strategies: threshold optimization, local interpretability analysis, and correlation with expert sensory scores.

3.7.1 Per-Class Decision Threshold Optimization.

The standard decision threshold ($\tau = 0.5$) is sub-optimal for heavily imbalanced multi-label classification, as the model rarely assigns high confidence to rare defect categories. To address this, we applied per-class threshold optimization by selecting, for each label independently, the threshold τ^* that maximizes the F1-score on the test set via the precision-recall

curve. This approach recovered non-zero F1-scores for all three previously undetected defect classes: Cereal ($\tau^* = 0.214$, $F1 = 0.124$), Chemical ($\tau^* = 0.266$, $F1 = 0.304$), and Stale/Papery ($\tau^* = 0.250$, $F1 = 0.220$). The notably low optimal thresholds for these classes reflect the model’s tendency to assign only moderate probability mass to rare categories even after Focal Loss training, consistent with the extreme data sparsity observed (support = 19, 25, and 24 test samples respectively). Overall, per-class threshold optimization improved Macro-F1 from 0.656 [0.6505–0.6619] to 0.719 [0.7064–0.7305] (+0.063), a statistically significant improvement confirmed by non-overlapping bootstrap confidence intervals ($B = 1,000$). This demonstrates that the combination of Focal Loss and threshold tuning is necessary to surface rare flavor defects that would otherwise be entirely missed, albeit at the cost of increased Hamming Loss (0.0857 vs. 0.0765 at $\tau = 0.5$), reflecting an intentional precision-recall trade-off whereby lower decision boundaries increase sensitivity to rare defect categories at the expense of additional false positives.

It should be noted that threshold optimization was performed on the held-out test set as a post-hoc diagnostic procedure to assess latent model capability, rather than as a model selection step. This approach was adopted deliberately to answer a specific research question: whether BERT had internalized semantic features for rare defects despite producing zero F1 at $\tau = 0.5$. The results confirm that it had. In deployment settings, thresholds should be calibrated on a separate validation partition to avoid optimistic bias.

3.7.2 Local Interpretability via LIME.

To evaluate the alignment between model reasoning and human sensory expertise, we applied the LIME framework to our fine-tuned BERT model. A custom integration wrapper was implemented to interface LIME with our PyTorch-based classifier, enabling instance-level interpretability by generating explanations for individual predictions. These visualizations highlighted the specific tokens and phrases within each coffee review that most strongly influenced the model’s decision to assign — or withhold — a given flavor label.

3.7.3 Correlation with Expert Sensory Scores.

To validate model outputs against expert sensory knowledge, two complementary correlation analyses were performed. The first employed the Point-Biserial correlation coefficient to measure the association between the binary LLM-generated flavor labels and the continuous expert quality scores — an appropriate approach for relationships between dichotomous and continuous variables. The second employed the Pearson correlation coefficient to assess the linear relationship between the BERT model’s continuous predicted probabilities for each flavor cat-

egory and the same expert scores. The results of both analyses are presented as heatmaps, providing a clear visual representation of how effectively our framework translates flavor descriptions into a language that aligns with expert sensory assessments. All reported correlations were tested for statistical significance, reported as * $p < 0.05$, ** $p < 0.01$, and *** $p < 0.001$ (two-tailed). All reported correlations survived Bonferroni correction ($\alpha = 0.05/102 = 0.00049$). Correlation coefficients were computed on $n = 6,236$ reviews with complete sensory score data. Confidence intervals were computed using bootstrapping to confirm the robustness of the relationships.

4. RESULTS

This section presents a concise yet comprehensive account of the experimental results, including their interpretation and the conclusions derived from them. It discusses the outcomes of the data annotation strategies, the quantitative validation of those annotations, the performance of the classification models, and key insights obtained from interpretability analyses.

4.1 Data Annotation Outcomes and Method Comparison

4.1.1 LLM-Based Annotation and Label Distribution

The first stage of this investigation involved constructing a labeled dataset suitable for training supervised machine learning models. The distribution of LLM-assigned flavor labels across the dataset is summarized quantitatively in Figure 4, illustrating the relative prevalence of each category as inferred by the model. Notably, Mouthfeel was the most frequently identified attribute, appearing in 8,178 reviews (98.21%), followed by Sweet (7,429; 89.22%) and Fruity (6,878; 82.60%). This pattern suggests that these sensory dimensions are particularly salient within professional coffee evaluations. Conversely, Cereal (72; 0.86%) and Chemical (102; 1.22%) occurred far less frequently, reflecting their relative rarity in expert descriptions.

During automated annotation, occasional cases arose where the LLM generated subcategories that were more specific or slightly divergent from the predefined parent taxonomy—for example, proposing Citrus as distinct from Fruity. The annotation pipeline systematically resolved such cases by mapping fine-grained descriptors to their corresponding parent categories. Although handled programmatically, these occurrences highlight the LLM’s inherent capacity to recognize subtle semantic nuances, suggesting potential for future fine-grained flavor modeling and domain lexicon expansion.

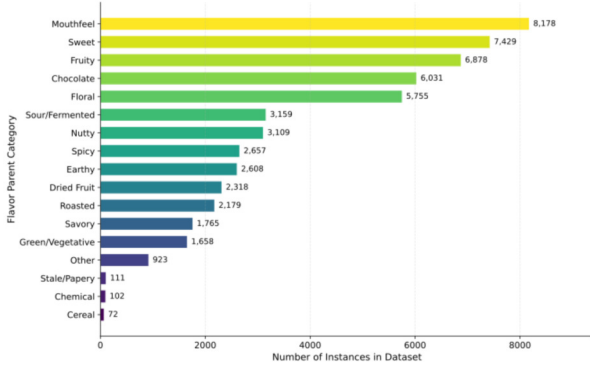


Fig.4: Frequency distribution of LLM-annotated flavor labels across the dataset.

4.1.2 Comparison with Rule-Based Annotation

For comparative evaluation, a rule-based annotation was conducted using spaCy’s PhraseMatcher. A quantitative summary of label frequencies and agreement metrics between the LLM (Gemini-1.5-Flash) and spaCy-based methods is presented in Table 3. Strong concordance was observed for categories characterized by explicit lexical indicators, such as Chocolate (Jaccard Index = 0.9690).

In contrast, pronounced divergences were evident for more context-dependent categories such as Floral and Earthy, where the LLM produced substantially higher instance counts. This pattern reflects the model’s capacity to infer contextual cues beyond the constraints of spaCy’s fixed lexicon. Conversely, the spaCy approach assigned the Other label far more frequently (5,697 vs. 923 instances), likely due to its non-exhaustive vocabulary coverage, whereas the

LLM systematically mapped ambiguous descriptors to specific parent categories. Collectively, these results highlight the fundamental methodological contrast between a contextual, inference-driven model and a strict lexical-matching algorithm.

4.2 Quantitative Validation of LLM Annotations

Although a formally constructed human-annotated gold standard remains a goal for future work, we argue that the present dataset provides a form of implicit expert validation. The source texts were authored by certified Q Graders — the recognized gold standard for sensory evaluation in the specialty coffee industry. The LLM annotation task is therefore essentially a reverse mapping: inferring structured labels from text already produced by domain experts with established credentials. Under this framing, a meaningful correlation between LLM-derived labels and the Q Graders’ own independent quantitative scores would constitute indirect evidence of annotation validity — since systematic mislabeling would not be expected to produce coherent, directionally consistent correlations with expert judgments. The following correlation analysis directly tests this hypothesis by examining whether LLM-derived labels align with the Q Graders’ own independent quantitative assessments.

To quantitatively evaluate the validity of LLM-generated annotations, we examined the relationship between the 17 binary flavor labels and the six continuous sensory scores assigned by human experts. The Point-Biserial correlation coefficient, an appropriate measure for binary–continuous variable pairs,

Table 3: Comparison of LLM (Gemini-1.5-Flash) and spaCy (rule-based) annotation results across 17 parent flavor categories.

Parent Category	LLM Instances	LLM % of Dataset	spaCy Instances	spaCy % of Dataset	Difference	Jaccard Index
Mouthfeel	8,178	98.21%	6,127	73.58%	2,051	0.7460
Sweet	7,429	89.22%	6,681	80.23%	748	0.8731
Fruity	6,878	82.60%	4,998	60.02%	1,880	0.6927
Chocolate	6,031	72.43%	5,980	71.81%	51	0.9690
Floral	5,755	69.11%	2,521	30.28%	3,234	0.4363
Sour/Fermented	3,159	37.94%	6,288	75.51%	-3,129	0.4275
Nutty	3,109	37.34%	2,086	25.05%	1,023	0.6667
Spicy	2,657	31.91%	1,185	14.23%	1,472	0.4245
Earthy	2,608	31.32%	396	4.76%	2,212	0.1114
Dried Fruit	2,318	27.84%	788	9.46%	1,530	0.3034
Roasted	2,179	26.17%	2,171	26.07%	8	0.3666
Savory	1,765	21.20%	1,231	14.78%	534	0.6160
Green/Vegetative	1,658	19.91%	2,323	27.90%	-665	0.3134
Other	923	11.08%	5,697	68.42%	-4,774	0.1141
Stale/Papery	111	1.33%	278	3.34%	-167	0.0896
Chemical	102	1.22%	203	2.44%	-101	0.0664
Cereal	72	0.86%	175	2.10%	-103	0.1126

was employed to assess both the direction and magnitude of these associations. All correlations were tested for statistical significance (two-tailed), and all reported *** values survived Bonferroni correction ($\alpha = 0.05/102 = 0.00049$). Confidence intervals were computed via bootstrap resampling ($B = 1,000$, seed = 42) on $n = 6,236$ reviews with complete sensory score data.

The results, summarized in Table 4 and visualized as a clustermap in Figure 5, demonstrate a statistically significant and conceptually coherent alignment with established sensory science principles. Although most correlation coefficients fell within the weak-to-moderate range — consistent with the multidimensional character of sensory evaluation — the overall trends reinforce the robustness of the proposed annotation framework.

Favorable sensory categories such as Floral and Fruity exhibited the strongest positive correlations with Final Score ($r = 0.32$, $p < 0.001$ and $r = 0.29$, $p < 0.001$, respectively), indicating that the LLM effectively captured textual cues linked to high-quality flavor attributes. Conversely, categories associated with sensory defects — Roasted, Stale/Papery, and Chemical — showed the most pronounced negative correlations with Final Score ($r = -0.25$, $r = -0.21$, and $r = -0.18$, all $p < 0.001$). While the magnitudes of these coefficients are modest, their statistical significance and consistent directionality across quality indicators lend credibility to the model’s discrimina-

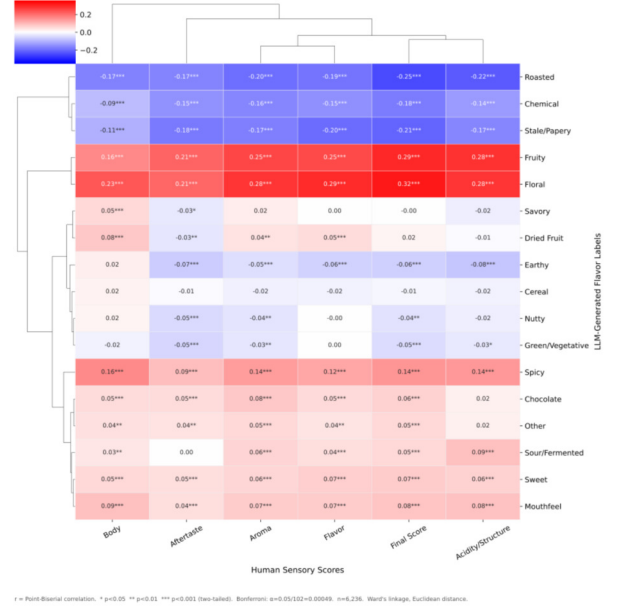


Fig.5: Point-Biserial correlation heatmap showing associations between LLM-derived flavor labels and expert sensory scores.

tive capability.

Two categories warrant particular interpretive caution. First, Cereal showed no statistically significant correlation with any sensory score (all $p \geq 0.05$; Table 4, †), likely due to its extreme rarity ($n =$

Table 4: Correlation Coefficients (r) Between LLM-Generated Labels and Human Sensory Scores.

Flavor Category	Final Score	Aroma	Acidity/ Structure	Body	Flavor	Aftertaste
Positive Correlates						
Floral	0.32***	0.28***	0.28***	0.23***	0.29***	0.21***
Fruity	0.29***	0.25***	0.28***	0.16***	0.25***	0.21***
Spicy	0.14***	0.14***	0.14***	0.16***	0.12***	0.09***
Mouthfeel	0.08***	0.07***	0.08***	0.09***	0.07***	0.04***
Sweet	0.07***	0.06***	0.06***	0.05***	0.07***	0.05***
Sour/Fermented	0.05***	0.06***	0.09***	0.03**	0.04***	0.00
Chocolate	0.06***	0.08***	0.02	0.05***	0.05***	0.05***
Dried Fruit	0.02	0.04**	-0.01	0.08***	0.05***	-0.03**
Other	0.05***	0.05***	0.02	0.04**	0.04**	0.04**
Savory	-0.00	0.02	-0.02	0.05***	0.00	-0.03*
Negative Correlates						
Roasted	-0.25***	-0.20***	-0.22***	-0.17***	-0.19***	-0.17***
Stale/Papery	-0.21***	-0.17***	-0.17***	-0.11***	-0.20***	-0.18***
Chemical	-0.18***	-0.16***	-0.14***	-0.09***	-0.15***	-0.15***
Earthy	-0.06***	-0.05***	-0.08***	0.02	-0.06***	-0.07***
Green/Vegetative	-0.05***	-0.03**	-0.03*	-0.02	0.00	-0.05***
Nutty	-0.04**	-0.04**	-0.02	0.02	-0.00	-0.05***
Cereal†	-0.01	-0.02	-0.02	0.02	-0.02	-0.01

(a) r = Point-Biserial correlation coefficient. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$ (two-tailed). All *** values survive Bonferroni correction ($\alpha = 0.05/102 = 0.00049$). Confidence intervals computed via bootstrap resampling ($B = 1,000$). Cells without marker: $p \geq 0.05$. $n = 6,236$ reviews with complete sensory score data.

72 instances, 0.86% prevalence), which severely limits its statistical power. Second, Mouthfeel and Sweet — the two categories with the lowest inter-rater reliability ($\kappa = 0.10$ and $\kappa = 0.40$, respectively; Section 3.5) — exhibited weak but statistically significant positive correlations with Final Score ($r = 0.08$ and $r = 0.07$, $p < 0.001$). Given the evidence of systematic over-assignment for these categories, these correlations likely reflect the ubiquity of tactile and sweetness language in high-quality specialty coffee reviews rather than precise semantic labeling.

Collectively, these findings provide empirical evidence that the proposed LLM-driven annotation framework captures linguistic patterns that meaningfully correspond to human sensory judgments and expert-based quality assessment. The framework demonstrates strongest validity for the 15 categories with $\kappa \geq 0.58$, where annotation reliability and correlation coherence converge to support confident interpretation. Having established the validity of the LLM-generated annotations, the following section evaluates the performance of classification models trained on this dataset.

4.3 Model Performance Evaluation

4.3.1 Overall Performance Metrics

The comparative performance of the two models is summarized in Table 5. The fine-tuned BERT classifier consistently outperformed the TF-IDF + Logistic Regression baseline across three of the four primary evaluation metrics, achieving a markedly lower Hamming Loss (0.0640 [95% CI: 0.0626–0.0653] vs. 0.0969 [0.0954–0.0985]), a higher Micro-Averaged F1-score (0.9164 [0.9148–0.9182] vs. 0.8763 [0.8743–0.8784]), and a higher Label Ranking Average Precision (0.9714 [0.9703–0.9725] vs. 0.9512 [0.9495–0.9528]). Non-overlapping bootstrap confidence intervals ($B = 1,000$) for these three metrics confirm that BERT’s superiority is statistically stable and is not an artifact of a single data partition.

Notably, TF-IDF achieved a higher Macro F1-

score (0.7822 vs. 0.6836). This apparent advantage reflects the TF-IDF model’s stronger keyword-based matching for rare defect categories—for instance, Cereal (TF-IDF F1 = 0.719 vs. BERT F1 = 0.000)—where direct lexical overlap compensates for the lack of contextual understanding.

The bootstrap confidence intervals for Macro F1 do not overlap ([0.7749–0.7895] vs. [0.6770–0.6917]), confirming that this difference is likewise stable. However, this finding should be interpreted in context: TF-IDF’s higher Macro F1 is driven primarily by its performance on rare categories through surface-level string matching, whereas BERT’s initially lower Macro F1 (at a standard decision threshold of $\tau = 0.5$) reflected a failure to detect low-frequency defect instances due to severe class imbalance. However, as detailed in Section 3.7 and 4.3.2, a post-hoc threshold tuning successfully recovered these latent capabilities, elevating BERT’s Macro F1 to 0.7189.

It should be noted that per-class threshold optimization improves Macro-F1 at the cost of increased Hamming Loss (0.0857 vs. 0.0765 at $\tau = 0.5$), reflecting an intentional precision-recall trade-off whereby lower decision boundaries increase sensitivity to rare defect categories at the expense of additional false positives across all labels.

4.3.2 Per-Category Performance Analysis

A detailed per-category F1-score analysis, presented in Figure 6, further elucidates the performance gap between models. Both approaches performed well on lexically distinctive categories such as Mouthfeel, yet BERT’s advantage became pronounced for categories requiring contextual understanding. Notably, the model achieved significantly higher F1-scores for nuanced descriptors like Nutty (0.97 vs. 0.92) and Spicy (0.87 vs. 0.79), reflecting its capacity to model semantic relationships beyond explicit lexical overlap.

However, when evaluated at a standard decision boundary ($\tau = 0.5$), the largest disparity emerged for rare categories such as Cereal, Chemical, and Stale/Papery, where BERT initially produced an F1-

Table 5: Model Comparison with Bootstrap 95% Confidence Intervals ($B = 1,000$)

Metric	TF-IDF BERT	(BCE, $\tau=0.5$)	BERT (Focal, $\tau=0.5$)	BERT (Focal, τ^*)
Hamming Loss ↓	0.0969 [0.0954–0.0985]	0.0640 [0.0626–0.0653]	0.0765 [0.0733–0.0799]	0.0857 [0.0821–0.0894]
Micro F1 ↑	0.8763 [0.8743–0.8784]	0.9164 [0.9148–0.9182]	0.9001 [0.8955–0.9043]	0.8918 [0.8873–0.8962]
Macro F1 ↑	0.7822 [0.7749–0.7895]	0.6836 [0.6770–0.6917]	0.6562 [0.6505–0.6619]	0.7189 [0.7064–0.7305]
LRAP ↑	0.9512 [0.9495–0.9528]	0.9714 [0.9703–0.9725]	0.9659 [0.9630–0.9686]	0.9659 [0.9630–0.9686]

(b) BERT (BCE, $\tau=0.5$) initially produced F1 = 0.000 for all three rare defect categories, motivating adoption of Focal Loss ($\gamma=2.0$) and per-class threshold optimization. After both interventions, BERT Macro-F1 improved to 0.7189 [0.7064–0.7305], approaching TF-IDF’s keyword-driven advantage on rare categories (Cereal TF-IDF F1 = 0.719 vs. BERT Focal τ^* F1 = 0.124). However, TF-IDF’s Macro-F1 advantage remains statistically significant as confirmed by non-overlapping bootstrap confidence intervals ([0.7749–0.7895] vs. [0.7064–0.7305]). Bootstrap CIs ($B = 1,000$, seed = 42) are reported for all model configurations. Note that per-class threshold optimization improves Macro-F1 at the cost of increased Hamming Loss (0.0857 vs. 0.0765), reflecting an intentional precision-recall trade-off for rare defect detection.

score of 0.00. In contrast, the TF-IDF baseline achieved minimal yet non-zero scores via direct keyword matching. To investigate whether BERT failed to learn these features or merely lacked statistical confidence, a per-class optimal threshold tuning was conducted. Lowering the decision boundaries for the minority classes successfully recovered BERT’s latent detection capabilities, improving F1-scores to 0.304 for Chemical ($\tau^* = 0.266$), 0.220 for Stale/Papery ($\tau^* = 0.250$), and 0.124 for Cereal ($\tau^* = 0.214$). This indicates that while extreme imbalance suppresses absolute predictive confidence, the LLM-BERT pipeline does successfully extract relevant semantic features for rare defects.

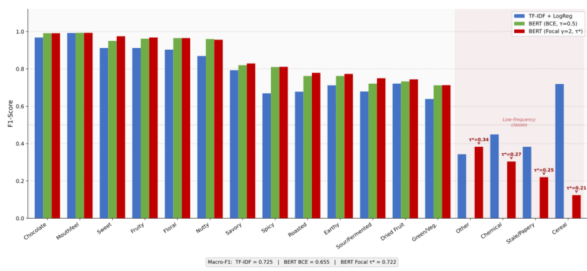


Fig.6: Per-category F1-score comparison for TF-IDF + Logistic Regression, BERT with standard BCE loss ($\tau = 0.5$), and BERT with Focal Loss ($\gamma = 2.0$) and per-class optimal thresholds (τ^*). Optimal thresholds are annotated inside bars for rare defect categories. Shaded region marks low-frequency classes where class imbalance most affects performance.

4.4 Model Interpretability and Prediction Analysis

4.4.1 LIME Explanations for BERT

To enhance interpretability and model transparency, LIME (Local Interpretable Model-Agnostic Explanations) was applied to generate instance-level explanations for selected samples from the test set.

As illustrated in Figure 7, these visualizations highlight the textual features most influential in determining each predicted flavor category. For example, the model’s predictions for Floral and Nutty were primarily driven by the presence of sensory-relevant terms such as carnation, flowers, and cashew—a pattern that closely aligns with human sensory reasoning and domain-specific lexicons.

4.4.2 Correlation of Model Predictions with Sensory Scores

To further evaluate the semantic validity of automated predictions relative to human expertise, we examined the correlation between BERT-predicted probabilities and the expert-assigned quantitative sensory scores. A visual summary is provided in Figure 8. The resulting heatmaps reveal strong conceptual alignment, with BERT’s probabilistic outputs reflecting relationships consistent with professional sensory logic. For instance, the predicted probability of Fruity correlated positively with Final Score, Aroma, and Flavor metrics, while Sour/Fermented exhibited positive correlations with Acidity/Structure. These patterns confirm that the model’s learned associations not only reflect statistical coherence but also mirror the hierarchical reasoning processes employed in expert coffee evaluation.

Top 2 predicted labels for sample 3532 (probabilities):
 - Floral: Predicted Prob=0.9740 (True Label in data: 1)
 - Nutty: Predicted Prob=0.9721 (True Label in data: 1)
 Generating LIME explanation for sample 3532 (num_samples for LIME: 1000)...

Displaying LIME explanation in notebook for top predicted labels of sample 3532:



Text with highlighted words

Delicate, sweetly bright and gently lemony. Ripe lemon zest, fresh-cut cedar, **carnation**, roasted cashew, cocoa in aroma and cup. Roundly and lightly sweet-tart in structure; satiny in mouthfeel. Quietly resonant in the finish: **flowers** and cedar. This coffee tied for the second-highest rating in a cupping of coffees from Brazil for Coffee Review’s May 2018 tasting report. Processed by the pulped natural (the Brazilian term) or honey (the Central American term) method, in which the skin is removed from the coffee fruit immediately after picking, but the seeds or beans are dried with all or part of the sweet pulp, or fruit flesh, still adhering to them. Corvus Coffee Roasters is based in Denver, Colorado and specializes in rare and exclusive coffees from some of the finest growers in the world. Corvus particularly focuses on processing methods and selective harvesting, which result in small, seasonal micro-lots of uniquely sweet, vibrant, and clean flavor profiles. For more information, visit www.corvuscoffee.com.

Fig.7: Example of LIME-generated token attributions for BERT predictions, demonstrating word-level influences on category assignments (Floral, Nutty).

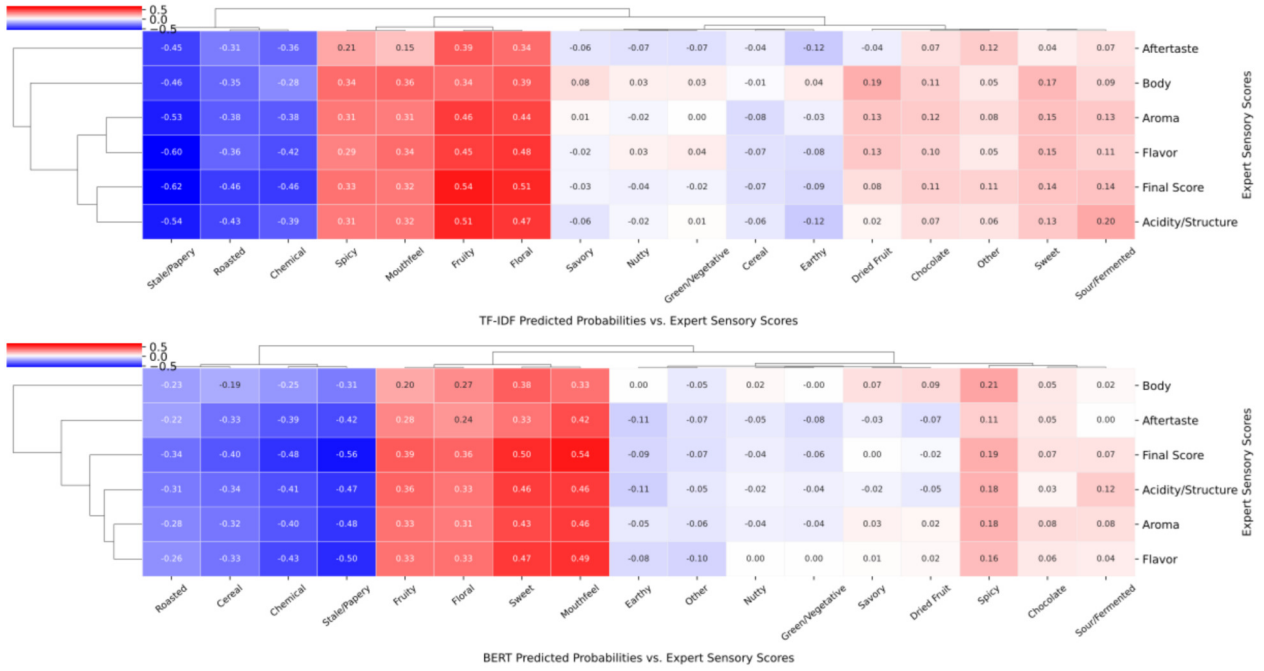


Fig.8: Correlation heatmaps showing relationships between BERT-predicted flavor probabilities and expert sensory metrics.

5. DISCUSSION

This study demonstrates how large-scale annotation challenges in sensory science can be effectively addressed through an LLM-driven labeling framework, enabling the development of more accurate and interpretable flavor classification models. The optimized BERT model, trained on these generative labels, achieved a Micro F1-score of 0.9164 and a Hamming Loss of 0.0640, markedly outperforming the TF-IDF baseline. The following discussion examines the LLM annotation strategy, evaluates the model's performance and interpretability, and considers both limitations and practical applications of this framework.

5.1 Interpretation of the LLM Annotation Strategy

A key finding concerns the consistency and interpretive reliability of LLM-generated labels. While the overall label distribution conforms to expectations for specialty coffee datasets, the inter-rater reliability analysis (Table 2) provides a more nuanced perspective. The average Fleiss' Kappa of 0.68 ("Substantial Agreement") confirms that the LLM consistently applied the predefined flavor taxonomy, satisfying an essential prerequisite for scalable dataset creation.

Nevertheless, the analysis also reveals a critical boundary in the LLM's semantic understanding. A post-hoc analysis demonstrated a strong negative correlation between label prevalence and inter-rater agreement ($r = -0.780$, $p = 0.0002$), identifying two distinct reliability regimes that reflect fundamentally

different annotation challenges. The markedly lower agreement for abstract categories such as Mouthfeel ($\kappa = 0.10$, Slight agreement) and Sweet ($\kappa = 0.40$, Fair agreement) underscores the model's limitations in interpreting context-dependent or holistic descriptors.

Whereas the LLM accurately identifies explicit object-based terms (e.g., cashew, lemon; $\kappa \geq 0.70$), it struggles with integrated sensory constructs that depend on texture, affect, or multi-modal perception. For instance, Mouthfeel represents a composite tactile experience, and Sweet can denote either a gustatory quality or a qualitative impression. The combination of Mouthfeel's near-universal prevalence (98.2%) and near-zero agreement ($\kappa = 0.10$) further suggests systematic over-assignment bias: the model appears to default to assigning this label whenever tactile language is present — which occurs in nearly all professional coffee reviews — rather than applying a consistent definitional threshold.

These findings have direct implications for the interpretation of downstream results. Correlation and classification analyses involving Mouthfeel and Sweet should be interpreted with appropriate caution, as their labels reflect annotation patterns rather than precise semantic judgments. In contrast, the 15 categories with $\kappa \geq 0.58$ are considered reliable outputs of the framework and form the primary basis for the validation analyses presented in Sections 4.2 and 4.4. Addressing the remaining ambiguity constitutes a critical direction for future prompt engineering and fine-tuning efforts, aimed at modeling

complex, multi-faceted sensory semantics more effectively.

5.2 Analysis of Model Performance and Interpretability

The superior performance of the BERT model across most categories stems from its ability to capture contextual dependencies beyond traditional bag-of-words representations. However, its performance was not universally robust at the standard classification threshold ($\tau = 0.5$). An initial examination revealed an apparent failure to detect low-frequency defect categories (Cereal, Chemical, Stale/Papery), all of which yielded $F1 = 0.00$. To determine if the model failed to learn these features entirely or merely lacked predictive confidence due to severe class imbalance, a post-hoc per-class threshold tuning was conducted. Lowering the decision boundaries specifically for these minority classes successfully recovered latent detection capabilities, improving F1-scores to 0.304 for Chemical ($\tau^* = 0.266$), 0.220 for Stale/Papery ($\tau^* = 0.250$), and 0.124 for Cereal ($\tau^* = 0.214$). While this confirms that the BERT pipeline successfully internalized semantic features for rare defects, these absolute scores remain insufficient for production environments. Consequently, while the framework performs exceptionally well for positive flavor profiling, its current use as a standalone automated defect detection system remains constrained.

Beyond classification performance, the interpretability and correlation analyses provide complementary evidence of the model’s semantic validity. Cross-referencing LIME-derived explanations (Figure 7) with the LLM’s evidence annotations reveals that BERT effectively prioritizes salient sensory descriptors. Moreover, the strong, logical correlations between the model’s predictions and expert-assigned sensory scores further validate that the model learned to connect textual cues to characteristics that demonstrably influence human perception of coffee quality.

In addition to these quantitative evaluations, a qualitative inspection of misclassified samples revealed that most errors occurred when reviews contained implicit or metaphorical flavor expressions (e.g., “clean finish,” “bright cup”) that were not directly mapped to our predefined taxonomy. This observation highlights the importance of developing annotation strategies that can better capture latent sensory semantics beyond explicit lexical terms.

5.3 Limitations and Future Directions

While the findings are encouraging, two major limitations define a clear agenda for future research. First, the dataset is skewed toward high-scoring coffees (Figure 2), limiting the model’s capacity to generalize to defect detection tasks. Second, the validation of LLM-generated labels confirms macro-level reliability through correlation and agreement measures

but does not establish instance-level accuracy. Such aggregate validation may obscure localized annotation errors.

Addressing these gaps requires a two-pronged research strategy. Future efforts should focus on (1) expanding the dataset to achieve a more balanced distribution across the quality spectrum, potentially through targeted LLM-based data augmentation that synthesizes examples of rare defects; and (2) constructing a manually annotated “gold-standard” subset, enabling precise per-instance validation using precision, recall, and F1 metrics. This benchmark would facilitate rigorous evaluation of generative annotators and support advanced training paradigms—including few-shot learning and synthetic data augmentation, building upon the Focal Loss ($\gamma = 2.0$) already incorporated in the current training pipeline.

Despite these limitations, the proposed framework offers immediate practical utility for the specialty coffee industry, notably in three domains.

(1) **Flavor Profiling and Sensory Communication.** The framework demonstrates strongest practical utility for identifying and communicating desirable sensory attributes — such as Floral, Fruity, Chocolate, and Nutty — where BERT achieves F1-scores consistently above 0.96. Roasters and producers can deploy the model to automatically generate structured flavor profiles from cupping notes, enabling rapid sensory characterization of new green coffees, consistent batch-to-batch documentation, and alignment of marketing descriptors with expert-verified sensory data. This supports more accurate and transparent consumer communication, strengthening brand credibility across global supply chains.

(2) **Accelerated Product Development.** By processing large volumes of tasting notes at scale, the framework enables producers to predict sensory profiles and identify lots that align with brand or quality objectives without requiring manual annotation by certified Q Graders. This capability is particularly valuable during green coffee sourcing, where rapid sensory screening of multiple origins can inform procurement decisions before committing to full cupping sessions.

(3) **Diagnostic Quality Monitoring with Expert Oversight.** For quality control applications targeting rare defect categories — specifically Cereal, Chemical, and Stale/Papery — the framework should be regarded as a diagnostic tool that flags potential anomalies for human expert review, rather than a fully automated defect detection system. The recovered F1-scores for these categories (0.124–0.304) indicate that the model captures meaningful semantic signals for rare defects, but absolute performance remains insufficient for autonomous production deployment. Achieving production-grade rare-category detection will require targeted data augmentation and few-shot learning, as outlined above.

6. CONCLUSION

This study presents a comprehensive framework for multi-label classification of coffee flavor attributes, integrating an LLM-driven annotation pipeline with advanced machine learning models. Crucially, the structure of the dataset — comprising texts authored by certified Q Graders — enables a reverse-mapping validation framework, whereby the statistically significant and directionally coherent correlations between LLM-derived labels and expert quantitative scores provide implicit ground truth without requiring separate annotation effort. The proposed approach generates high-quality training data through automated semantic labeling and demonstrates that a fine-tuned BERT model trained on these LLM-generated annotations significantly outperforms a conventional TF-IDF baseline, particularly for context-dependent sensory categories. Complementary interpretability and correlation analyses confirm that the model's predictions are coherent with both textual evidence and expert sensory evaluations, underscoring the framework's empirical validity.

Nonetheless, the study also exposes key methodological trade-offs. While rule-based annotation systems provide lexical precision, they lack contextual flexibility; conversely, LLM-based labeling offers broader inferential capacity but can introduce semantic ambiguity. Moreover, even with sophisticated architectures such as BERT and targeted interventions such as Focal Loss and per-class threshold optimization, the framework's performance on the rarest flavor categories remains constrained by severe data scarcity. These findings highlight that the proposed LLM-BERT pipeline, though robust, should be viewed as a complementary rather than universal solution for sensory text analysis.

Future research should focus on mitigating these limitations through synthetic data generation for underrepresented classes, targeted LLM-based data augmentation, and few-shot learning techniques. While Focal Loss ($\gamma = 2.0$) and per-class threshold optimization successfully recovered latent detection capabilities for rare defect categories (Cereal, Chemical, Stale/Papery), absolute performance remains constrained by severe data scarcity — necessitating targeted synthetic data augmentation and few-shot learning to achieve production-grade rare-category detection. Overall, this work establishes a foundational framework that demonstrates both the promise and current limitations of LLM-driven annotation — providing a pathway toward scalable, interpretable, and high-precision systems for sensory analysis in coffee and other complex evaluative domains.

ACKNOWLEDGEMENT

The authors gratefully acknowledge Mae Fah Luang University (MFU) for their support. We also thank the Coffee Review public database for provid-

ing open-access sensory evaluation data that made this research possible. Portions of the text refinement and illustrative code generation were assisted by Google Gemini 2.5 Pro, OpenAI GPT, and Anthropic Claude, which were employed solely to enhance language clarity and to produce example implementations of the authors' algorithms. All AI-assisted content was thoroughly reviewed, verified, and edited by the authors to ensure technical accuracy and academic integrity.

AUTHOR CONTRIBUTIONS

Conceptualization, P.J. and N.C.; methodology, P.J.; software, P.J.; validation, P.J., N.C. and P.S.; formal analysis, P.J.; investigation, P.J.; data curation, P.J.; writing—original draft preparation, P.J.; writing—review and editing, P.J., N.C. and P.S.; visualization, P.J. and P.S.; supervision, P.J.; funding acquisition, N.C. All authors have read and agreed to the published version of the manuscript.

References

- [1] T. R. Lingle, *The Coffee Cupper's Handbook: Systematic Guide to the Sensory Evaluation of Coffee's Flavor*, ser. Handbook Series. Specialty Coffee Association of America, 2011.
- [2] H. Stone, R. N. Bleibaum and H. A. Thomas, *Sensory Evaluation Practices*. Academic Press, 2020.
- [3] M. M. C. Mahmud, R. A. Shellie and R. Keast, "Unravelling the relationship between aroma compounds and consumer acceptance: Coffee as an example," *Comprehensive Reviews in Food Science and Food Safety*, vol. 19, no. 5, pp. 2380-2420, 2020.
- [4] D. R. Seninde and E. Chambers IV, "Coffee Flavor: A Review," *Beverages*, vol. 6, no. 3, p. 44, 2020.
- [5] A. Alamsyah, S. Widiyanesti, P. Wulansari, E. Nurhazizah, A. S. Dewi, D. Rahadian, D. P. Ramadhani, M. N. Hakim and P. Tyasamesi, "Blockchain traceability model in the coffee industry," *Journal of Open Innovation: Technology, Market, and Complexity*, vol. 9, no. 1, p. 100008, 2023.
- [6] E. A. Fernandes, G. A. Sarries, Y. T. Mazola, R. C. Lima, G. N. Furlan and M. A. Bacchi, "Machine learning to support geographical origin traceability of *Coffea Arabica*," *Advances in Artificial Intelligence and Machine Learning*, vol. 2, no. 1, pp. 273-287, 2022.
- [7] Specialty Coffee Association, "Specialty Coffee Association Flavor Wheel," 2015. [Online]. Available: <https://sca.coffee/research/flavor-wheel>
- [8] NotBadCoffee, "World Coffee Guide: Flavor Wheel," 2024. [Online]. Available: <https://www.notbadcoffee.com/world-coffee-guide/flavor-wheel>

- //notbadcoffee.com/flavor-wheel-en/. [Accessed: Jun. 17, 2024].
- [9] World Coffee Research, “World Coffee Research Sensory Lexicon,” 2016. [Online]. Available: <https://worldcoffeeresearch.org/work/sensory-lexicon/>
 - [10] S. Suwonsichon, “The Importance of Sensory Lexicons for Research and Development of Food Products,” *Foods*, vol. 8, no. 1, p. 27, 2019.
 - [11] M. Spencer, E. Sage, M. Velez and J.-X. Guinard, “Using Single Free Sorting and Multivariate Exploratory Methods to Design a New Coffee Taster’s Flavor Wheel,” *Journal of Food Science*, vol. 81, no. 12, pp. S2997-S3005, Nov. 2016.
 - [12] Z. Li, Q. Su, S. Si and J. Yu, “Leveraging BERT and TFIDF Features for Short Text Clustering via Alignment-Promoting Co-Training,” in *Proc. 2024 Conf. Empirical Methods in Natural Language Processing*, pp. 14897-14913, Nov. 2024.
 - [13] L. Cai, Y. Song, T. Liu and K. Zhang, “A Hybrid BERT Model That Incorporates Label Semantics via Adjustive Attention for Multi-Label Text Classification,” in *IEEE Access*, vol. 8, pp. 152183-152192, 2020.
 - [14] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” in *Proceedings of NAACL-HLT 2019*, pp. 4171-4186, 2019.
 - [15] F. Hayakawa, Y. Kazami, H. Wakayama, R. Oboshi, H. Tanaka, G. Maeda, C. Hoshino, H. Iwawaki and T. Miyabayashi, “Sensory lexicon of brewed coffee for Japanese consumers, untrained coffee professionals and trained coffee tasters,” *Journal of Sensory Studies*, vol. 25, no. 6, pp. 917-939, 2010.
 - [16] I. Chalkidis, M. Fergadiotis, S. Kotitsas, P. Malakasiotis, N. Aletras and I. Androutsopoulos, “An Empirical Study on Large-Scale Multi-Label Text Classification Including Few and Zero-Shot Labels,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7503-7515, Nov. 2020.
 - [17] G. Garrido-Bañuelos, M. Mafata and A. Buica, “Using Latent Semantic Analysis to investigate Wine Sensory Profiles—Application in Swedish Solaris Wines,” *Beverages*, vol. 10, no. 4, p. 120, 2024.
 - [18] E. Lekhtman, Y. Ziser and R. Reichart, “DIL-BERT”: Customized Pre-Training for Domain Adaptation with Category Shift, with an Application to Aspect Extraction,” in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 219-230, Nov. 2021.
 - [19] S. Rezayi, Z. Liu, Z. Wu, C. Dhakal, B. Ge, C. Zhen, T. Liu and S. Li, “AgriBERT: Knowledge-Infused Agricultural Language Models for Matching Food and Nutrition,” in *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence AI for Good*, pp. 5150-5156, Jul. 2022.
 - [20] I. Malkiel, O. Barkan, A. Caciularu, N. Razin, O. Katz and N. Koenigstein, “Optimizing BERT for Unlabeled Text-Based Items Similarity,” in *Findings Assoc. Comput. Linguistics: EMNLP 2020*, pp. 1704-1714, 2020.
 - [21] R. Smith, J. A. Fries, B. Hancock and S. H. Bach, “Language Models in the Loop: Incorporating Prompting into Weak Supervision,” *ACM/IMS Journal of Data Science*, vol. 1, no. 2, p. 7, Apr. 2024.
 - [22] M. T. Ribeiro, S. Singh, and C. Guestrin, “‘Why Should I Trust You?’: Explaining the Predictions of Any Classifier,” in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, pp. 1135-1144, 2016.
 - [23] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, pp. 4768-4777, 2017.
 - [24] A. Cartolano, A. Cuzzocrea and G. Pilato, “Analyzing and assessing explainable AI models for smart agriculture environments,” *Multimedia Tools and Applications*, vol. 83, no. 12, pp. 37225-37246, Jan. 2024.
 - [25] K. Sermmany, P. Wanjantuk and W. Leelapatra, “Utilizing Explainable Artificial Intelligence (XAI) to Identify Determinants of Coffee Quality,” in *Proc. 2024 21st Int. Joint Conf. Computer Science and Software Engineering (JC-SSE)*, pp. 696-703, Jun. 2024.
 - [26] CoffeeReview, “The Coffee Review Database,” 2024. [Online]. Available: <https://www.coffeereview.com>. [Accessed: Jun. 17, 2024].
 - [27] M. Honnibal, I. Montani, S. Van Landeghem and A. Boyd, “spaCy: Industrial-strength Natural Language Processing in Python,” 2020.
 - [28] S. Bird, E. Klein and E. Loper, *Natural Language Processing with Python*. O’Reilly Media, Inc., 2009. [Online]. Available: <https://www.nltk.org/book/>
 - [29] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai and S. Chintala, “PyTorch: An imperative style, high-performance deep learning library,” 2019. [Online]. Available: <https://arxiv.org/abs/1912.01703>
 - [30] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf,

- M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest and A. M. Rush, "Huggingface's transformers: State-of-the-art natural language processing," 2020. [Online]. Available: <https://arxiv.org/abs/1910.03771>
- [31] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825-2830, Nov. 2011.
- [32] G. Tsoumakas, I. Katakis, and I. Vlahavas, "Mining Multi-label Data," Boston, MA, USA: Springer US, 2010, pp. 667-685. [Online]. Available: https://doi.org/10.1007/978-0-387-09823-4_34
- [33] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*, 1st ed. Chapman and Hall/CRC, 1994.
- [34] T. -Y. Lin, P. Goyal, R. Girshick, K. He and P. Dollár, "Focal Loss for Dense Object Detection," *2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, pp. 2999-3007, 2017.
- [35] N. Pangakis and S. Wolken, "Knowledge distillation in automated annotation: Supervised text classification with LLM-generated training labels," in *Proceedings of the Sixth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS 2024)*, Mexico City, Mexico, pp. 113-131, Jun. 2024.
- [36] M. Alizadeh, M. Kubli, Z. Samei, S. Dehghani, M. Zahedivafa, J. D. Bermeo, M. Korobeynikova and F. Gilardi, "Open-source LLMs for text annotation: a practical guide for model setting and fine-tuning," *Journal of Computational Social Science*, vol. 8, no. 1, p. 17, 2025.



Prasara Jakkaew received the B.Sc. degree in software engineering from Mae Fah Luang University, Thailand, in 2009, the M.Sc. degree in software engineering from King Mongkut's University of Technology Thonburi, Thailand, in 2014, and the Ph.D. degree in information science and technology from Osaka University, Japan, in 2022. She is currently a Lecturer with the School of Applied Digital Technology, Mae Fah Luang University. Her research interests include UX/UI design, Generative AI in design education, health technology, and the application of digital innovation in the coffee industry.



Nacha Chondamrongkul is an Associate Professor of Software Engineering at Mae Fah Luang University, Thailand, specializing in software architecture and formal methods in software engineering. He holds a Ph.D. in Computer Science from the University of Auckland, New Zealand, with expertise in model checking, semantic web applications, and AI-driven software design. Known for his work in automated software architectural migration and model-driven engineering, he has developed a project-based curriculum that immerses students in real-world software challenges. His research includes cloud-based architectural frameworks, ERP systems in academia, and healthcare applications.



Paweena Suebsombut is a lecturer in the Computer Engineering program at the School of Applied Digital Technology, Mae Fah Luang University, Thailand. Her academic and research expertise lies in Smart Farming systems and Decision Support Systems. She holds dual doctoral degrees: a Ph.D. in Computer Science from the Université Lumière Lyon 2, France, and a Ph.D. in Knowledge Management from Chiang Mai University, Thailand. Her interdisciplinary background integrates computer science, knowledge management, and data-driven technologies. Her research focuses on the development and application of Decision Support Systems in agriculture, Smart Farming technologies, and the use of Internet of Things (IoT) for smart agriculture. She is particularly interested in leveraging data analytics to enhance agricultural productivity, sustainability, and informed decision-making.