*Sensors and Materials*, Vol. 37, No. 11 (2025) 4819–4841
MYU Tokyo



S & M 4216

# Enhancing Artificial Olfactory Reasoning via Integration of Electronic Nose Sensing, Large Language Models, and Knowledge Graphs: A Case Study on Coffee E-Nose

Chung-Hong Lee,[1*] Hsin-Chang Yang,[2] Jun-Teng Sun,[1] and Zhen-Xin Fu[1]

[1]Department of Electrical Engineering, National Kaohsiung University of Science and Technology, 415, Jiangong Rd., Sanmin Dist., Kaohsiung City 807618, Taiwan
[2]Department of Information Management, National University of Kaohsiung, 700, Kaohsiung University Rd., Nanzih District, Kaohsiung 811, Taiwan

(Received June 23, 2025; accepted September 22, 2025)

***Keywords:*** artificial olfactory reasoning, electronic nose, large language model, knowledge graphs

Artificial olfactory systems have been applied in domains such as food quality assessment, environmental monitoring, and medical diagnostics. However, progress in enabling machines to perform high-level reasoning based on odor perception remains limited. To address this gap, we propose a novel hybrid system that integrates electronic nose (E-Nose) sensing with large language models (LLMs) and knowledge graphs, enabling human-like olfactory reasoning through the interaction of sensory and linguistic data. A case study on coffee aroma interpretation demonstrates the system's ability to generate descriptive narratives, infer semantic relationships, and contextualize odor signals meaningfully. To simulate odor perception, we employed a TETCN model—combining a transformer encoder and a temporal convolutional network—to predict aroma types and generate structured labels. These labels guide the retrieval of relevant knowledge from a memory database, which is then processed by the LLM for advanced reasoning. By bridging signal-level perception and abstract cognition, this work presents a significant advancement toward cognitively intelligent olfactory systems.

## 1. Introduction

Artificial olfactory reasoning, as a core branch of perceptual AI, is progressively overcoming the technical limitations of traditional odor analysis. Similar to the human olfactory system, artificial olfactory perception plays an important role in many areas, including food quality assessment, environmental monitoring, and medical diagnosis. However, progress in artificial olfactory reasoning from detected odors has been surprisingly slow. This can be attributed to the complexity of recognizing olfactory information and various technical limitations. Nevertheless, interest in the study of olfactory recognition mechanisms has been steadily increasing. In particular, electronic nose (E-Nose) technology[(1–10)] has recently emerged as an important method for digitizing odor data to address the challenge of objectively capturing and interpreting

*Corresponding author: e-mail: leechung@mail.ee.nkust.edu.tw
https://doi.org/10.18494/SAM5826

ISSN 0914-4935 

complex olfactory information. Despite advances in sensor precision and pattern recognition, most systems remain limited to low-level signal classification and lack the cognitive capacity to reason over complex inputs or infer contextual meaning. Real-world interpretation often demands the integration of background knowledge, domain expertise, and historical patterns—capabilities that traditional machine learning methods struggle to support. To address these limitations, recent advancements in AI offer promising solutions; in particular, the emergence of large language models (LLMs) introduces the ability to interpret odor-related descriptors and generate human-like inferences, whereas knowledge graphs[11,12] provide a structured backbone for linking sensory data to domain-specific concepts. However, a systematic integration of these components for olfactory reasoning remains underexplored in the literature. Therefore, in this study, we propose a hybrid system that combines E-Nose sensing with the reasoning capabilities of LLMs and knowledge graphs to explore deep interactions between olfactory and linguistic data, where human-like reasoning functions can be performed. To verify the feasibility of the proposed system, we conducted a case study on coffee aroma interpretation, demonstrating how the proposed system can enhance the understanding of odors by generating descriptive narratives, inferring underlying relationships, and contextualizing sensory inputs in a semantically meaningful way. To simulate odor perception, we employed a hybrid model, the TETCN algorithm, which combines a transformational encoder (TE) and a temporal convolutional network (TCN), to predict aroma types and generate structured labels. These labels guide the retrieval of relevant knowledge from a memory database, which is then processed by the LLM for advanced reasoning. The main contribution of this work is that our hybrid system approach overcomes the current limitations of artificial olfactory reasoning, bridges the gap between complex odor information and contextual reasoning models, and allows the machine to simulate the cognitive mechanism of human abstract thinking and reasoning about odor, which in turn provides a new path for AI-based olfactory research.

The structure of this paper is as follows. In Sect. 2, we review related work on artificial olfactory systems. In Sect. 3, we outline the system framework and methods. In Sect. 4, we present benchmarking experiments of various LLMs and evaluate system performance, including the quality of generated odor descriptions. In Sect. 5, we conclude the study and discuss future research directions.

## 2. Related Work

The artificial olfactory system developed in this study is inspired by the human olfactory system.[13] During inhalation, volatile molecules reach the interior of the nasal cavity. The olfactory epithelium in the nasal cavity interacts with these odor molecules. The olfactory neurons, which act as receptors, transmit the molecular binding process to the brain via electronic signals. Thus, the essence of odor perception is the conversion of chemical interactions between olfactory receptors and volatile molecules into electronic signals that transmit external information to the brain.[13,14] Information about odors is encoded in the olfactory bulbs in the form of patterns. In other words, olfactory judgment is determined by the pattern formed by different combinations of receptors that recognize the specific molecular characteristics of each

odor molecule. Therefore, to mimic the above olfactory sensing process, machine learning and artificial neural network techniques are used to categorize pattern data from the sensor array (i.e., E-Nose) in the developed artificial olfactory system.[1,2,15] In our system, the E-Nose component is used to perform sensing and odor pattern recognition, whereas the LLM component is used to perform more advanced inference tasks, such as olfactory-based interpretation. To the best of our knowledge, there is no similar research work that integrates E-Nose, LLM, and knowledge graph approaches to explore olfactory reasoning functionality in the development of an artificial olfactory system. This is an interdisciplinary research effort. Several issues and studies related to this research are discussed below. Odor analysis is widely applied across diverse domains and can be implemented using an E-Nose. A conventional E-Nose comprises a multichannel gas-sensor array; each sensor exhibits a distinct sensitivity toward volatile organic compounds (VOCs), allowing the system to characterize gaseous constituents.[3–5] Odor analysis is most frequently deployed in the food sector—for instance, Ren *et al.*[6] identified food types by classifying gaseous components and measuring their concentrations—but it also extends to other areas, such as discriminating alcoholic beverages[7] and identifying coffee cultivars.[8,9] Contemporary E-Nose systems are typically trained with deep-learning algorithms whose stacked hidden layers yield more granular and accurate analyses of odor data. Wang *et al.*,[10] for example, combined a convolutional neural network with a wavelet-scattering network to gauge freshness by exploiting odor differences that arise at various spoilage levels. Despite this progress, E-Nose research has largely remained limited to gas identification; in-depth studies on the relationships among different odor datasets are lacking, leaving current solutions short of meeting the full range of human olfactory needs.

Reasoning ability is highly correlated with human memory;[16–18] consequently, effective reasoning applications must integrate human sensory modalities such as olfaction. LLMs designed for reasoning emulate human cognitive mechanisms, and as new reasoning-oriented models continue to emerge, researchers have begun to explore diverse approaches for reasoning tasks.[19–21] Earlier work on olfactory reasoning in AI has concentrated on several fronts, one of which is a model's capacity to describe and comprehend olfactory information.[22] For example, Shaari *et al.*[23] examined the accuracy of LLMs (e.g., GPT-4o and Google Gemini) when answering odor-related questions, whereas Esteban-Romero *et al.*[24] investigated cross-modal information integration to enhance model understanding in the olfactory domain. Schwarz and Hamburger[25] confirmed the strong link between odors and memory, thereby motivating research on reasoning driven by olfactory cues. Although these studies have achieved incremental advances, they mainly assess a model's ability to handle odor-centric questions. Mahmud *et al.*[26] combined odor signals with machine-learning techniques to localize scent sources within physical spaces, thereby tying olfaction to real-world environments; however, the model's odor cognition still relies heavily on prior knowledge, and research on reasoning across different odors and their interrelations remains limited. To date, no comprehensive framework has succeeded in instilling a human-like olfactory reasoning mechanism within an AI system.

Since E-Nose systems primarily generate low-level digital signals (e.g., changes in resistance and VOC concentrations) but lack the ability to reason about "what the odor is" or "what it means," the use of knowledge graphs allows for the mapping of these signals into semantic

entities and relationships. In addition, knowledge graphs define relationships between entities (such as aroma, compound, origin, and roast level), enabling cross-level logical inference.

## 3. Methods

The developed system uses the E-Nose, which is based on a multisensor array device, to mimic human olfactory function as the basic unit of olfactory sensing. The neural mechanism of olfactory odor recognition starts from the differential interaction between different types of receptors and odor molecules, similar to the interaction between neurotransmitters and receptors in the nervous system. The system framework and configuration of the artificial olfactory system we developed are consistent with the basic principles of the human olfactory system, which are described in detail below. In this study, we used the interpretation and reasoning of coffee aroma as a case study to explain how the integration of E-Nose with LLMs and knowledge graphs can achieve the above functions.

In Fig. 1, the proposed system framework is illustrated. As shown in Fig. 1, coffee aromas are first collected by an E-Nose system; the resulting signals then undergo wavelet denoising to suppress noise and $Z$-score normalization to place all channels on a common scale. Dimensionality reduction techniques then compress the high-dim tensional odor data into low-dimensional vectors while retaining both local and global structural features. These reduced vectors are encoded by a TE to reinforce temporal-context awareness and are subsequently fed into a TCN that classifies tested coffee samples. This study made full use of our previously established coffee-aroma dataset, which was created using an E-Nose based on a multisensor

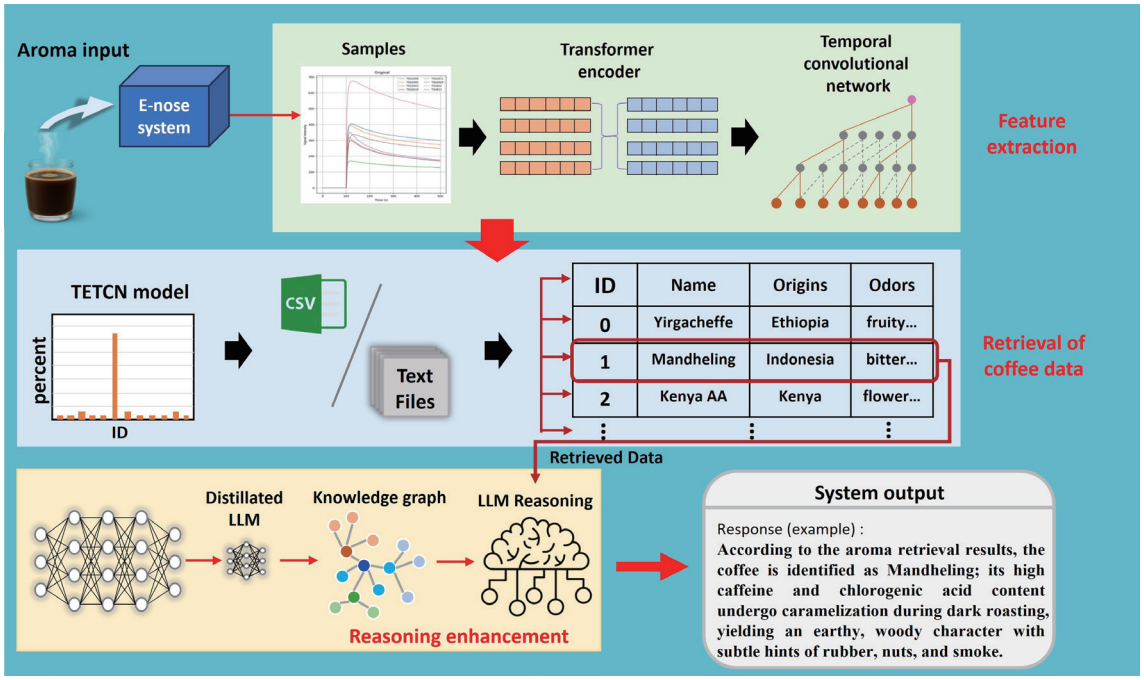



Fig. 1. (Color online) Proposed system framework.

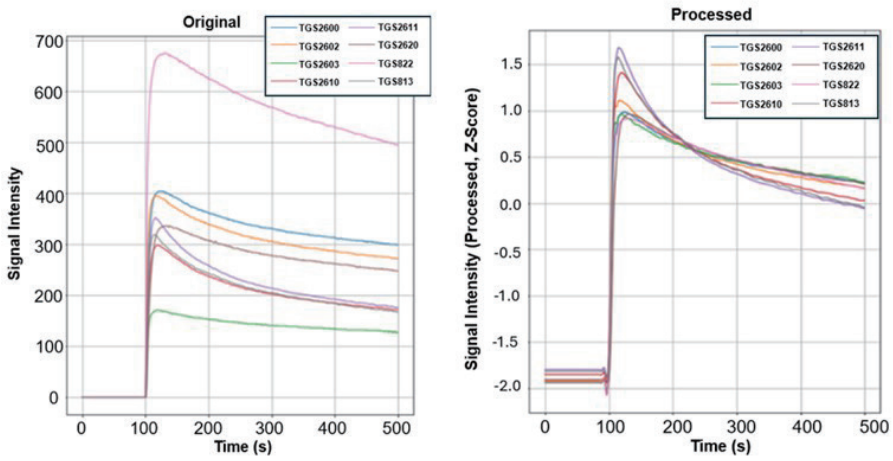
array device and covers the data related to aromas of various specialty coffees from around the world. This dataset proved indispensable when applying the TETCN method to coffee identification. During the preparation phase, digital data and metadata on the key aromatic characteristics of each specialty coffee tested were preserved and edited into this dataset, providing a basic reference for sample identification during training. TCN's output is linked to a Retrieval-Augmented Generation (RAG) database containing fields such as variety name, origin, odor descriptors, and application suggestions, so the system can fetch the pertinent information. Enhanced reasoning is achieved via two complementary components: a knowledge graph built from enthusiasts' reviews—including roast level, processing method, brewing technique, and aroma—and a distilled model that restricts the language model's focus to the coffee domain while curbing computational cost. By fusing these two reasoning resources with RAG retrieval, the system delivers coherent, context-rich responses, thereby creating an end-to-end user interaction pipeline that spans from coffee odor acquisition to LLM output. This reasoning mechanism primarily utilizes the reasoning capabilities of LLM and knowledge graphs representing the relationships between diverse semantic concepts in the field of coffee, enabling the model to think more deeply and connect relevant information. In the following sections, the system modules are described in detail.

### 3.1 E-Nose data collection and preprocessing

To ensure that high-quality odor data are collected and better train the coffee recognition model, ambient air is first recorded to establish a baseline, then signal peaks are captured to reflect dynamic odor characteristics, and finally, measurements are collected at a steady state to minimize variability. The resulting time-series data (high-dimensional responses of multiple sensors to volatile organic compounds) are denoised by wavelet transformations to preserve low-frequency odor characteristics while suppressing high-frequency noise, further smoothed using a moving average window to remove random fluctuations, and finally standardized by $Z$-score (*mean* = 0, *standard deviation* = 1) transformations to remove sensor-specific biases, which accelerates model convergence and compensates for sensor-specific biases and baseline drift (see Fig. 2). The sample in the figure is Honey Hears Geisha coffee produced in Colombia.

In the feature extraction stage, we collect each coffee's aroma in two parts: an initial phase, during which the E-Nose system draws ambient air and, over the next 500 s, records sensor data to capture baseline fluctuations so that later baseline drift can be corrected, and a response phase, in which the coffee aroma is detected. From the system response, we obtain both the sensors' peak and steady-state signals, which serve as the main cues for identifying the coffee aroma. After standard preprocessing, we apply principal component analysis to derive the feature set P, composed of initial phase characteristics and response phase features; taken together, these form the key odor signatures that feed into model training.

In this study, the hardware component of the E-Nose system we developed primarily consists of sensor-array-related circuits, which are composed of specially selected sensors suitable for coffee aroma detection. Table 1 shows the sensors and their target gases used for constructing




(a) (b)

Fig. 2. (Color online) Comparison chart after denoising and smoothing.

Table 1
Sensor and specific target gases.

| Sensor type | Target gas |
|---|---|
| TGS2600 | Methane, carbon monoxide, butane, ethanol, hydrogen |
| TGS2602 | Toluene, ammonia, ethanol, hydrogen |
| TGS2603 | Amine series, sulfurous odors |
| TGS2610 | Ethanol, hydrogen, methane, butane, propane |
| TGS2611 | Ethanol, hydrogen, methane, butane |
| TGS2620 | Methane, carbon monoxide, butane, hydrogen, ethanol |
| TGS813 | Methane, propane, butane |
| TGS822 | Ethanol, organic solvent |

the sensor array. All of the sensors used in the experiment are metal-oxide–semiconductor devices manufactured by Figaro and are designed chiefly for detecting VOCs in coffee aromas.

### 3.1.1 Experimental protocol and feature extraction

A standardized experimental protocol was followed to ensure reproducibility and minimize external variability. For each coffee variety, beans were roasted on two separate days to evaluate cross-batch consistency. During each session, 15 replicate measurements were performed per variety. The sensing experiments were conducted under controlled environmental conditions of 27 ± 2 ℃ and 55 ± 5% relative humidity. Ambient air was first recorded to establish a baseline. Subsequently, odor responses were collected for 1000–1500 s per trial, and the stable 500 s segment covering both the transient peak and the steady state was retained for analysis.

Raw time-series signals were then segmented using a sliding window approach, resulting in approximately 20000 data segments derived from all measurements. This segmentation strategy increased the effective sample size for training while preserving the temporal dynamics of sensor responses.

Preprocessing steps included baseline correction, wavelet-based denoising, moving-average smoothing, and *Z*-score normalization. From the preprocessed signals, both transient and steady-state features were extracted. The transient features included rise time, decay time, peak derivative values, and the area under the transient curve, reflecting the adsorption–desorption dynamics of the sensors. The steady-state features captured plateau responses at equilibrium.

Together, these features provided a comprehensive representation of odor dynamics. This enriched time-series data input was then processed by the TETCN model, enabling the system to leverage both the static and dynamic aspects of the sensor signals for robust coffee discrimination.

### 3.2 Identification of sensor data

To utilize the odor data collected by the E-Nose to predict the type of coffee sample and generate appropriate labels for subsequent retrieval, a TETCN model integrating a TE and a TCN was trained in this study[5], as shown in Fig. 3. Each measurement lasted 1000–1500 s, and 500 s of data covering both peak and stable signals was retained. The raw time-series signals were further segmented into multiple overlapping windows, resulting in approximately 20000 data segments derived from all coffee measurements. These segments served as the effective training samples for the TETCN model. The input is a time series cut into different time windows as queries, and the key is the index of its time point. The output is an ID representing the predicted coffee type, which is used as the basis for subsequent retrieval.

The TE specializes in sequence processing, allowing the model to capture the contextual information inherent in E-Nose measurements, while the TCN can accept variable-length inputs and model long-range temporal dependencies, properties that are well suited for time-series

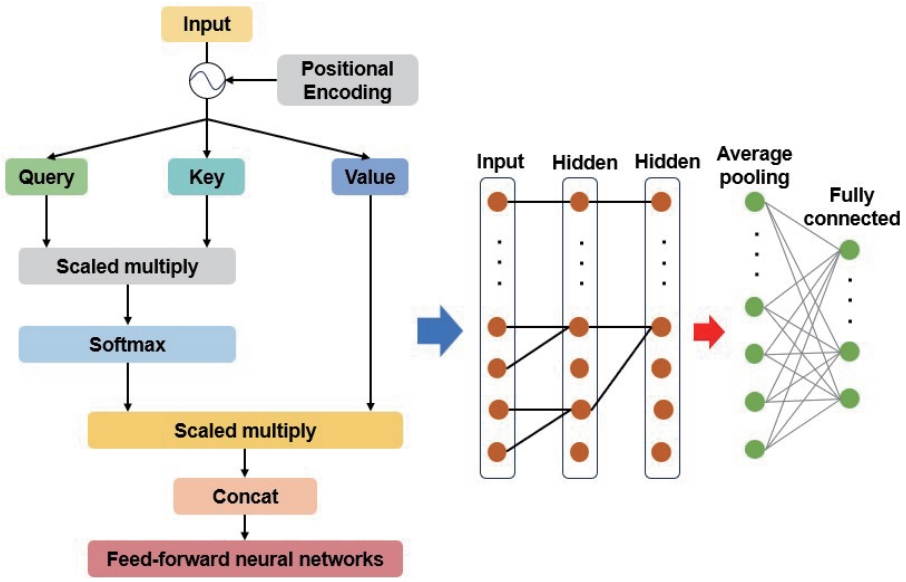



Fig. 3. (Color online) Mechanism of TE and TCN.[5]

data. The odor signal is weighted by the TE and then fed into the TCN, which efficiently extracts salient features from the odor data.

For model evaluation, the dataset was first split into training and test data subsets at a ratio of 8:2. To ensure robust assessment, a stratified 5-fold cross-validation was applied within the training dataset, preventing data leakage and maintaining the balanced representation of coffee varieties across folds. Final performance was reported on the held-out 20% test dataset, and additional chance-level baselines together with Cohen's $\kappa$ statistics were calculated to confirm reliability.

The TE leverages a self-attention mechanism to capture inter-temporal relationships within the sequence—an aspect that is especially critical for tracking odor fluctuations across different time points in sensor data. Equation (1) specifies how the corresponding attention weights are computed. Q, K, and V denote Query, Key, and Value, respectively.[5]

$$Attention(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

Here, $d_k$ acts as a scaling factor that stabilizes the gradients. Once the data have been processed by the TE, the context-enriched time-series data inputs are fed into the TCN for training. The TCN applies causal convolutions together with dilated convolutions to regulate the spacing of the inputs (see Fig. 4).

To mitigate the vanishing-gradient issue that can arise during training, the TCN adopts residual connections, as illustrated in Eq. (2). In this formulation, *TCN_Block*($x$) denotes the output produced after several layers of dilated convolution and nonlinear transformations.

$$Output = Activation\left(x + TCN\_Block\left(x\right)\right) \tag{2}$$

The TETCN algorithm (see Algorithm 1) outlines the workflow for applying the model to coffee-aroma recognition. First, the E-Nose-sensed time-series data are normalized and

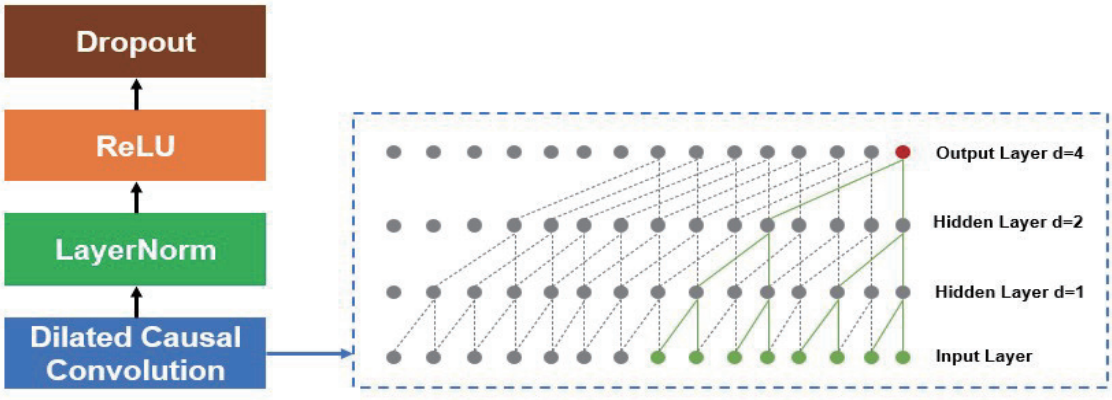



Fig. 4. (Color online) Convolution layer and block of TCN.[5]

Algorithm 1
TETCN-based odor recognition.

```
Input: TimeSeries // from sensor output
Output: OdorLabel // predicted class
Normalize (TimeSeries)
Segment ← SlidingWindow(TimeSeries, windowSize, stride)
PosEncoded ← AddPositionalEncoding(Segment)
if useTransformer then
    for each layer in TransformerLayers do
        PosEncoded ← MultiHeadAttention(PosEncoded)
        PosEncoded ← FeedForward(PosEncoded)
        PosEncoded ← LayerNorm(PosEncoded)
    end for
else
    PosEncoded ← PosEncoded // skip Transformer block
end if
TCNinput ← PosEncoded
for each layer in TCNLayers do
    TCNinput ← DilatedCausalConv(TCNinput)
    TCNinput ← ReLU(TCNinput)
    TCNinput ← Dropout(TCNinput)
end for
Output ← GlobalAveragePooling(TCNinput)
Logits ← FullyConnected(Output)
OdorLabel ← Softmax(Logits)
return OdorLabel
```

segmented into consecutive slices with a sliding window. A TE module then extracts global dependency patterns: after positional encoding, the sequence passes through multiple layers of self-attention and feed-forward networks, preserving rich contextual information. The resulting representation enters a TCN, where multilayer dilated and causal convolutions capture local features. Global average pooling followed by a fully connected layer produces recognition logits, which are converted to aroma-class probabilities via a softmax function. By jointly modeling global and local characteristics, this architecture markedly improves the accuracy of odor recognition.

After normalizing the time series, we apply a sliding window to slice the sequence into overlapping segments, thereby enlarging the training dataset and preserving local dynamic features. Position encoding is then added so that the model can capture relationships among different time points. During the Multi-Head Attention stage, correlations across temporal segments are detected, while the subsequent Feed-Forward and LayerNorm layers integrate the data through linear transformations and facilitate convergence. The sequence then enters a TCN: dilated causal convolutions maintain the causal direction of the series, expand the receptive field, and enable the model to learn long-term dependencies with relatively few layers. In the decision layer, Global Average Pooling aggregates the sequence over the time dimension, compressing variable-length inputs into a fixed-length vector that passes through a fully connected layer to produce logits, which are finally converted into a probability distribution by a softmax to yield the predicted label.

The experimental environment and hyperparameter settings are summarized in Table 2. Because hyperparameters greatly affect model performance, choosing suitable values is crucial. In this study, Bayesian optimization was used to determine key hyperparameters such as key dimension, the number of attention heads, and dropout rate, whereas the remaining hyperparameters were obtained through manual fine-tuning during experimentation.

### 3.3 Knowledge distillation

To enhance the accuracy and professionalism of olfactory description of coffee flavors while reducing the computational resource requirements, in this study, we incorporated the technique of knowledge distillation as shown in Fig. 5.

By compressing the model and transferring the knowledge from the large "teacher" model to the smaller "student" model, the student model can maintain high performance with significantly reduced computational requirements. In this study, the loss function for knowledge distillation

Table 2
Hyperparameter settings of TETCN.

| Hyperparameter | Value |
|---|---|
| Key dimension | 18 |
| Number of attention heads | 2 |
| Learning rate | 0.001 |
| Dropout rate | 0.5 |
| Filter size | 2 |

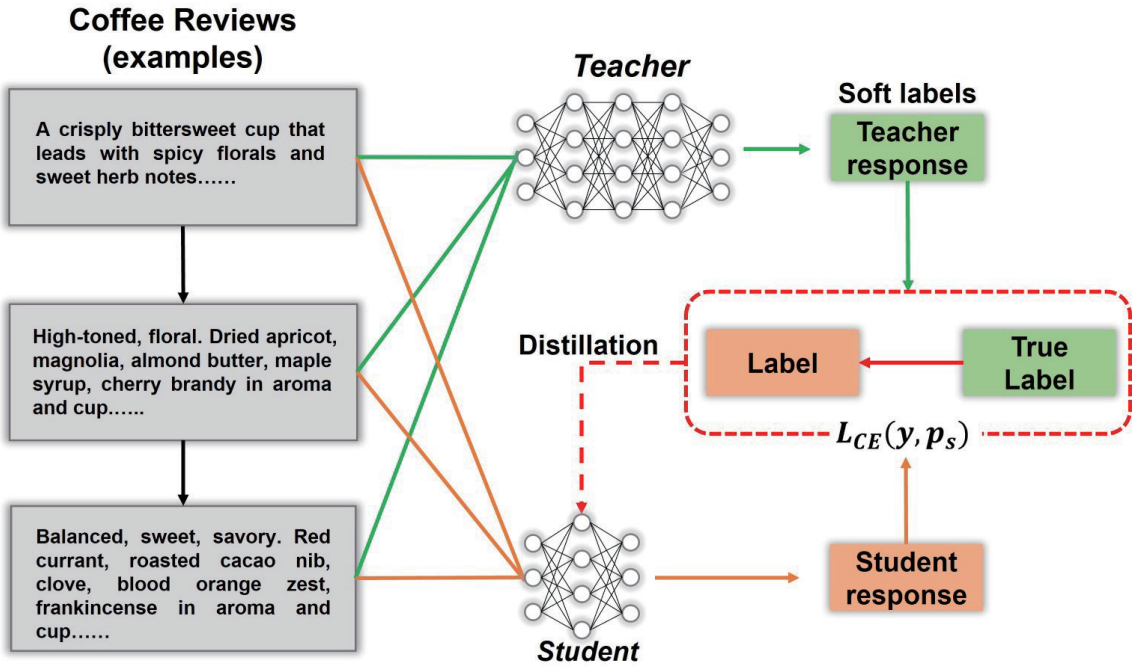



Fig. 5. (Color online) Knowledge distillation process: Coffee reviews as an example of training texts.

consists of two components. The first is the cross-entropy loss between the student model's output and the ground truth labels, which measures the student model's predictive accuracy on actual data. The second is the Kullback–Leibler (KL) divergence loss between the student model's output and the soft labels produced by the teacher model after temperature scaling, which quantifies the extent to which the student model mimics the knowledge of the teacher model.

$$L = \alpha \cdot L_{CE}\left(y, p_s\right) + \left(1-\alpha\right) \cdot T^2 \cdot L_{KL}\left(q_t, q_s\right) \tag{3}$$

As shown in Eq. (3), the loss function incorporates a weighting factor $\alpha$, which controls the balance between the cross-entropy loss (based on ground truth labels) and the distillation loss (based on the teacher model's soft labels). $L_{KL}$ represents the Kullback–Leibler (KL) divergence, which measures how closely the student model's predictions align with the teacher model's soft labels. This loss function is designed to achieve the goal of compressing the model size while maintaining high prediction accuracy.

### 3.4 Knowledge graph

In addition, in performing more advanced reasoning functions, we integrated knowledge graphs into the system to enhance the reasoning capability of the developed system, that is, to allow the model to help us better understand the relationship between different coffee varieties and the similarity between their aroma characteristics, as shown in Fig. 6. The knowledge graph systematically represents the relationships among coffee varieties, aromas, and roast levels, thus

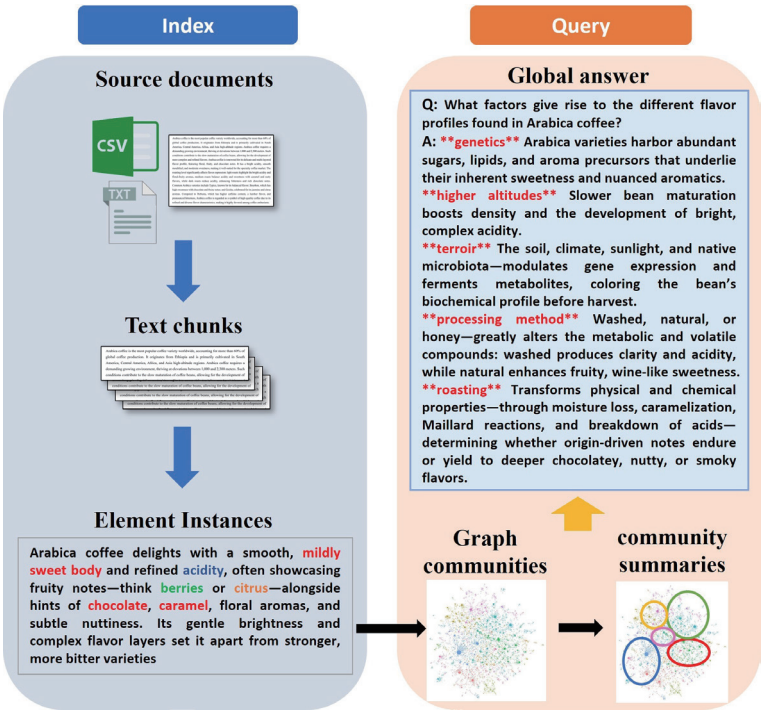



Fig. 6. (Color online) Example of a process for building a knowledge graph using LLMs.

enhancing the inference capability of the model. Although inference-based LLMs have strong general knowledge processing capabilities, they are limited in their ability to infer subtle relationships among coffee aromas. To address this limitation, knowledge maps provide a structured, predefined mapping of coffee-related information that effectively complements LLMs. The knowledge graph constructed for this case study consists of four main components: coffee varieties, origins, roasting levels, and corresponding aroma profiles.

The entire process is divided into two primary stages: indexing and querying. In the indexing stage, the dataset is segmented on the basis of its content to generate element instances, relations, and statement descriptions. These extracted elements are then transformed into a graph-based structure. During the querying stage, the knowledge graph generated in the indexing phase is utilized to perform community-based segmentation, where the graph is partitioned into several regions on the basis of the similarity between nodes. From these partitions, community summaries are constructed. The summaries are ranked according to the significance of each node and incorporated into the language model's context in a prioritized manner.

In Fig. 7, an example of part of a knowledge graph that reveals a subset of its underlying relationships is illustrated. The knowledge graph attaches relevance and interpretability to each node: most nodes inherit new relationships from their respective sources and generate

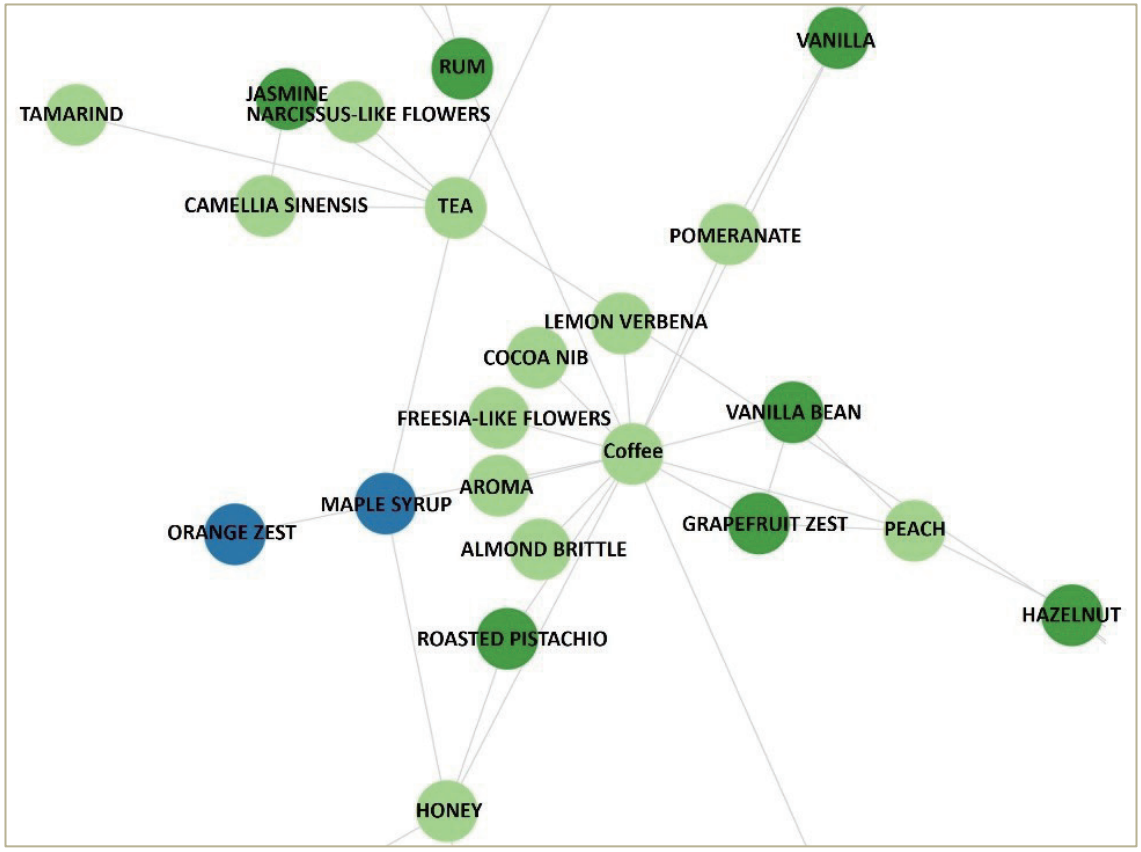



Fig. 7. (Color online) Sample knowledge graph of coffee, roast, and aromas.

corresponding descriptions to ensure that the paired links remain interpretable. As shown in Fig. 7, coffee acts as the source for many other nodes whose connections are highly coherent—largely additives or odor notes found in coffee beverages. Different colors indicate that, when the graph is constructed, the nodes originate from different source paragraphs yet remain related in some way. The more similar the colors, the higher the degree of relatedness. For example, grapefruit zest is a flavoring agent sometimes infused into coffee drinks, while cocoa nibs are often added to various coffees to enrich flavor. Maple syrup serves not only as a sweetener in coffee but also as an additive in certain teas, and tea shares some acidity characteristics with coffee. By explicitly linking such ingredients and flavor attributes, the model can reason more effectively about coffee aroma and its related information, thus enhancing the reasoning ability of the system.

## 4. Experimental Results

As stated previously, to ensure that the selected LLMs remain focused on coffee-flavor-related tasks, we employed a knowledge distillation strategy into the system development. We first collected a large corpus of coffee-related data, such as coffee reviews. After data cleaning and annotation to ensure quality and relevance, the refined dataset was used to fine-tune selected LLMs, enabling them to generate more accurate and nuanced descriptions of coffee flavors. During distillation with temperature $T = 20$, each model undergoes 100 training iterations on a coffee flavor dataset supervised by the teacher model; during training, the student model learns to increase its coffee knowledge by mimicking the teacher model's responses to each sensory description. This distillation pattern allows the student model to learn only about the coffee domain, thereby reducing the excess computational overhead while maintaining its expertise in coffee-oriented queries. The results showed that the student model retains more than 80% of the original performance of the teacher model in this configuration. Afterwards, we conducted benchmarking experiments using the trained student models.

### 4.1 Benchmarking the performance of various LLMs for this study

To select the most suitable LLM for ensuring experimental stability and high performance—and because our study must reason about odors under various scenarios while drawing on extensive historical data, which demands robust long-context handling—we assessed several models across five evaluation axes: multitask competence, multiturn dialogue, logical reasoning, situational reasoning, and long-text comprehension. Specifically, MMLU-pro gauges multitask ability through 12k challenging questions spanning diverse disciplines; MT-Bench tests multiround conversational skill with 80 context-dependent prompts from multiple fields; BBH offers 23 tasks explicitly designed to push the logical-reasoning limits of current LLMs; HellaSwag probes contextual understanding via ~70000 multiple-choice questions, each pairing a passage with four candidate endings; and MuSR evaluates situational reasoning and long-context grasp across three domains, each containing several hundred items. In all experiments, the temperature was set to 0 so that each model deterministically selected the highest-probability

token, guaranteeing consistent outputs for identical inputs. Evaluations were run on the FastChat platform, and every reasoning task employed a chain-of-thought prompting strategy.

As shown in Table 3, we evaluated various LLMs across multiple datasets and tasks to compare their contextual understanding and reasoning capabilities. The results revealed that Grok-3 exhibits the most outstanding overall performance in logical and contextual reasoning tasks, while also achieving high scores in the remaining tasks. In particular, Grok-3's training corpus incorporates social-media reviews, giving the model a richer aroma-related vocabulary that strengthens its ability to reason about coffee questions and to express aromatic descriptions in text. As a result, Grok-3 was selected as the best model in this application-specific study to maintain stable system performance and ensure a high degree of robustness.

### 4.2 Evaluation on knowledge-graph-enhanced LLMs

In this study, we adopted knowledge graph technology as the core mechanism for enhancing an LLM's reasoning capability. By building a graph-structured space within the odor database and performing graph-based searches, the knowledge graph furnishes explicit explanations of the relationships among diverse aromas, enabling the model to reason across a broader spectrum of odor sources—an approach we term "*Knowledge-graph-enhanced LLM*" technique. In this case study, to quantify how the coffee-flavor descriptions of the system differ from those of human experts, we validated its outputs with two standard text metrics: BLEU and ROUGE. BLEU is a metric for evaluating machine translation that compares the similarities of two text segments and assesses both fluency and lexical accuracy. As shown in Eq. (5), BLEU is computed by measuring n-gram precision to gauge the similarity between a generated text and its reference description, while incorporating a brevity penalty (BP) to penalize overly short outputs, as defined in Eq. (6).

$$BLEU = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \cdot \log p_n\right) \tag{5}$$

$$BP = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \le r \end{cases} \tag{6}$$

Table 3
Performance evaluation of different LLMs within the proposed system.

| Model | MMLU-PRO | MT-Bench | BBH | HellaSwag | MuSR |
|---|---|---|---|---|---|
| Deepseek-r1-14b | 84.82 | 78.41 | 40.69 | 18.34 | 28.71 |
| Grok 3 | 79.91 | 82.84 | 68.74 | 84.62 | **73.33** |
| o1-preview | **89.33** | 67.92 | 67.92 | 75.74 | 69.81 |
| OpenAI o3 | 73.43 | 80.61 | **73.21** | **87.79** | 68.89 |
| Claude 3.7 sonnet | 84.71 | **86.93** | 65.82 | 84.86 | 71.14 |
| Llama-3.3-70b | 65.92 | 56.56 | 56.56 | 10.51 | 15.57 |
| Gemma-2-27b | 56.54 | 49.27 | 49.27 | 16.67 | 9.11 |
| Qwen-2.5-72b | 64.38 | 54.62 | 54.62 | 20.69 | 19.64 |
| Phi-4-14b | 70.40 | 55.24 | 55.24 | 11.63 | 10.13 |

When calculating *n*-gram precision, we set $n = 4$ to assess the fluency of the generated text and ran 1000 iterations for each model to compute its BLEU score. Figure 8 shows the comparison of the BLEU scores achieved by the various LLMs, revealing that reasoning augmentation yields a marked performance gain across all models. The gray and green bars respectively indicate the performance scores of the baseline and knowledge-graph-enhanced LLMs in describing coffee flavors. Grok-3 records the highest score overall, while among open-source models, Llama 3.3 exhibits the most stable performance and enjoys a clear advantage over its counterparts.

To measure the effectiveness of the reasoning component—specifically, the model's ability to retrieve and summarize information from the database and to compare its coverage and completeness with expert reviews—we employed the ROUGE evaluation metric. ROUGE is computed here using the longest common subsequence approach, as shown in Eqs. (7)–(9). In these equations, $P_L$ and $R_L$ represent precision and recall, respectively, whereas the F-score is derived from these two measures.

$$R_L = \frac{LCS(X,Y)}{|Y|} \tag{7}$$

$$P_L = \frac{LCS(X,Y)}{|X|} \tag{8}$$

$$F_L = \frac{(1+\beta^2)\cdot P_L \cdot R_L}{P_L + \beta^2 R_L} \tag{9}$$

Using the ROUGE-L metric to assess performance on the flavor-description task offers two principal advantages: (i) it requires no fixed n-gram length, allowing richer semantic

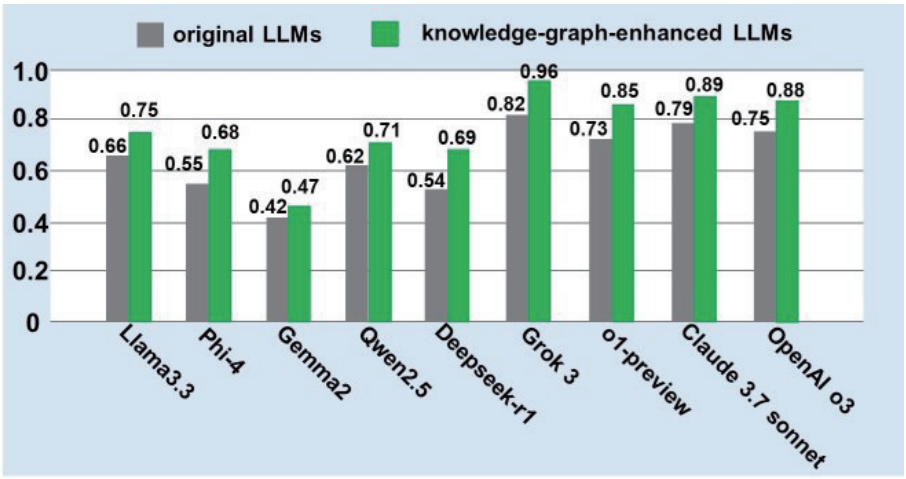



Fig. 8. (Color online) Comparison of BLEU scores achieved by baseline and knowledge-graph-enhanced LLMs.

relationships to be captured in longer texts, and (ii) its tolerance for non-contiguous word order gives the model greater flexibility in phrasing. This method therefore lets us verify divergences between the coffee aroma descriptions of the model and those of human experts. The ROUGE-L evaluation results for the model before and after reasoning enhancement are presented in Table 4. The best performing model, Grok-3, improves by more than 10%, indicating that it covers the greatest amount of key information and achieves the highest similarity to expert flavor descriptions.

The experimental results demonstrated that integrating knowledge graph techniques effectively and accurately enhances the reasoning capabilities of LLMs in the olfactory domain. When benchmarked against human experts' coffee flavor descriptions, reasoning through knowledge graph entities and relations yields clear gains in the textual fluency, precision, and coverage of key information.

### 4.3 Overall performance of E-Nose and LLM-integrated system

To assess the overall performance of the system that combines an E-Nose with an LLM, we adopted a classification consistency protocol that examines the correlation and agreement between the model's coffee predictions and its descriptive outputs. We began by constructing a test set composed of samples bearing ground-truth flavor labels. Each sample underwent the full processing pipeline, after which the olfactory-reasoning system generated a flavor description; from that description, the coffee category was reverse-inferred. These inferred categories were then compared with the E-Nose classifications over multiple trials on the labeled aroma samples. Finally, Cohen's $\kappa$ coefficient was calculated to quantify the system's consistency in coffee flavor categorization, thereby verifying that the reasoning module reliably captured both the label information and the E-Nose results when formulating its responses.

First, we assessed the system's practical improvement. Using Kilimanjaro coffee as an example, Fig. 9 shows that the knowledge-graph-enhanced LLM supplies fuller and clearer information than the baseline LLM when answering flavor-related questions. This comparison reveals that the knowledge-graph-enhanced model can draw on more factors such as the region's terrain, soil, and climate, and reason about how these elements give rise to Kilimanjaro's distinctive taste.

Table 4
ROUGE-L evaluation results of LLMs.

| Model | Original LLM | | | Knowledge-graph-enhanced LLM | | |
|---|---|---|---|---|---|---|
| | Precision | Recall | F-score | Precision | Recall | F-score |
| Deepseek-rl | 0.5337 | 0.5196 | 0.5266 | 0.6457 | 0.6328 | 0.6392 |
| o1-preview | 0.7048 | 0.6932 | 0.6990 | 0.7563 | 0.7492 | 0.7529 |
| OpenAI o3 | 0.7561 | 0.7353 | 0.7456 | 0.8229 | 0.8081 | 0.8157 |
| Grok-3 | **0.8122** | **0.7825** | **0.7971** | **0.9065** | **0.8933** | **0.8998** |
| Claude-3.7-sonnet | 0.7897 | 0.7802 | 0.7849 | 0.8382 | 0.8464 | 0.8419 |
| Llama 3.3 | 0.6339 | 0.6121 | 0.6228 | 0.6977 | 0.7056 | 0.7014 |
| Phi-4 | 0.5528 | 0.5203 | 0.5361 | 0.6561 | 0.6312 | 0.6431 |
| Gemma 2 | 0.4824 | 0.4774 | 0.4799 | 0.5334 | 0.5117 | 0.5227 |
| Qwen 2.5 | 0.4108 | 0.4926 | 0.4480 | 0.5106 | 0.5496 | 0.5294 |

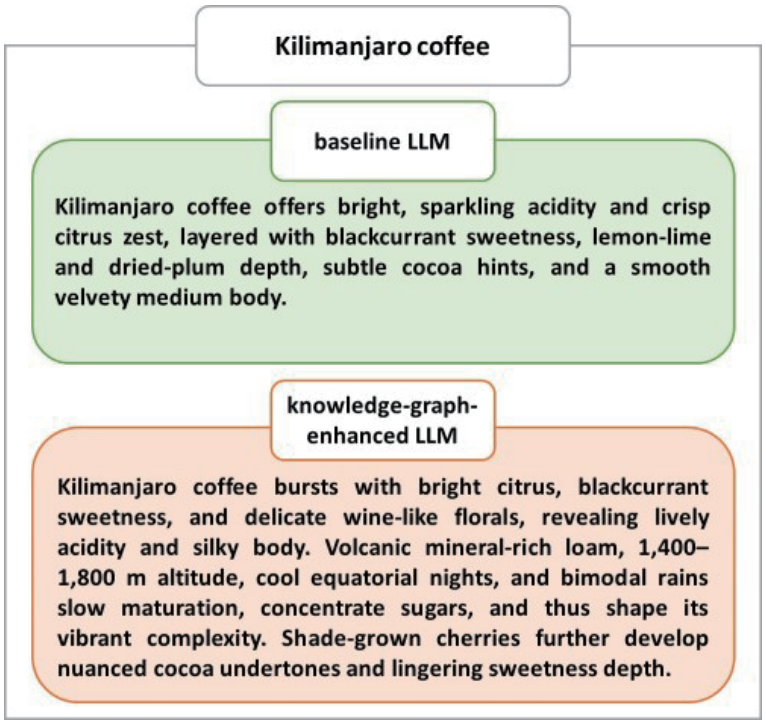



Fig. 9. (Color online) Qualitative comparison example of text generated by baseline LLM (green) and knowledge-graph-enhanced LLM (orange).

After the E-nose system classifies the sample, it outputs an ID such as Mandheling. This ID is then used to query a RAG database, which returns the corresponding coffee category together with information such as origin and flavor descriptors (e.g., citrus notes and jasmine aroma). A model that has been distilled and enhanced with a knowledge graph uses the retrieved information to produce a detailed flavor description for that coffee. In the reverse-inferred stage, the model's task is inverted: with the RAG component disabled, the system must infer the coffee category solely from the flavor description generated in the previous step. This reverse reasoning serves to verify that, when responding to coffee-related questions, the model relies on the E-nose recognition and retrieval outputs rather than on its own pre-training alone; if it can accurately deduce the coffee category from the flavor description, it demonstrates that the entire system operates coherently and consistently.

Table 5 shows the classification accuracies achieved for various coffee varieties using the TETCN algorithm approach. Ten distinct coffees—representing diverse origins and processing methods that introduce subtle flavor differences—served as the experimental samples. The results indicated that the training procedure performed well: for most varieties, the classification accuracy exceeded 90 percent.

To validate the system's classification consistency, we also evaluated the reverse-reasoning capability of the LLM-based, reasoning-enhanced module. In this test, odor samples were first classified by the E-Nose; the system then received the same samples without labels and had to infer each coffee category solely from the flavor descriptions it generated. The goal was to

Table 5
Evaluation results of E-Nose system for odor prediction of sampled coffee beans.

| Coffee Name | Country | Type | Processing Method | Accuracy | | |
|---|---|---|---|---|---|---|
| | | | | CNN | LSTM | TE + TCN |
| Mandheling | Indonesia | Arabica | Wet | 0.91 | 0.90 | 0.93 |
| Kilimanjaro | Tanzania | Kent/Bourbon | Wet | 0.90 | 0.91 | 0.92 |
| Guji Adola | Ethiopia | Heirloom | Wet | 0.86 | 0.88 | 0.90 |
| | | Heirloom | Dry | 0.90 | 0.85 | 0.91 |
| Yirgacheffe | Ethiopia | Arabica | Wet | 0.87 | 0.88 | 0.94 |
| | | Arabica | Dry | 0.86 | 0.88 | 0.96 |
| Kenya AA | Kenya | Wet | Wet | 0.95 | 0.92 | 0.93 |
| Sigri Estate | Paoua New Guinea | Typica | Dry | 0.91 | 0.93 | 0.91 |
| Mocha Matari | Yaman | Mocha Java | Dry | 0.90 | 0.91 | 0.94 |
| Hartmann Estate | Panama | Pacamara | Wet | 0.94 | 0.89 | 0.88 |
| Finca El General | Guatemala | Maragogype | Dry | 0.91 | 0.87 | 0.89 |
| Brazil Santos | Guatemala | Bourbon | Wet | 0.88 | 0.92 | 0.92 |

determine whether the model's descriptions followed the classifications and retrieved labels. Figure 10 shows the coffee classification performance with E-Nose and LLM experiments. Because the LLM has not been trained on large volumes of labeled data, its accuracy is naturally lower than that of the E-Nose system; nevertheless, reasoning backward from flavor descriptions still achieved a high level of identification, indicating that the overall system performance in this study is highly consistent. Figure 11 illustrates the actual sample counts classified by both the E-Nose system and the LLM. With ten samples provided for each coffee variety, the results showed that the two classifiers generally produce consistent categorizations. Misclassifications occur only in a few cases where certain varieties share very similar flavor profiles, leading the LLM to make incorrect predictions on the basis of its flavor descriptions.

Equation (10) illustrates how Cohen's $\kappa$ is calculated. $p_o$ is the observed proportion of agreement between the two raters, while $p_e$ is the expected proportion of agreement by chance—that is, the probability that both raters either classify the sample correctly or misclassify it into the same coffee category. Because a small sample size can bias these probability estimates and make $\kappa$ lower than expected, we tested 100 aroma samples from different coffees and tracked how $\kappa$ changed. The experimental findings indicated that a small sample size affects classification consistency. Thus, we examined how various sample sizes affect Cohen's $\kappa$, as illustrated in Fig. 12. When the number of samples falls below ten, $\kappa$ drops considerably. In addition, because chance agreement ($p_e$) depresses the $\kappa$ statistic, this index can also reveal bias in a dataset: when bias is present, the probability of chance agreement rises and $\kappa$ falls sharply. In our experiment, however, most $\kappa$ values remained above 0.7, underscoring the dataset's stability and reliability. Even across two roasting batches, the classification accuracy remained above 90% and Cohen's $\kappa$ values exceeded 0.7, confirming robustness across batches. Even so, for our reasoning-enhanced system, overall $\kappa$ values remained above 0.6, signifying strong agreement between coffee classification and flavor description performance. This outcome confirms that the E-Nose system and the reasoning model integrate and link their outputs effectively.

$$K = \frac{p_o - p_e}{1 - p_e} \tag{10}$$

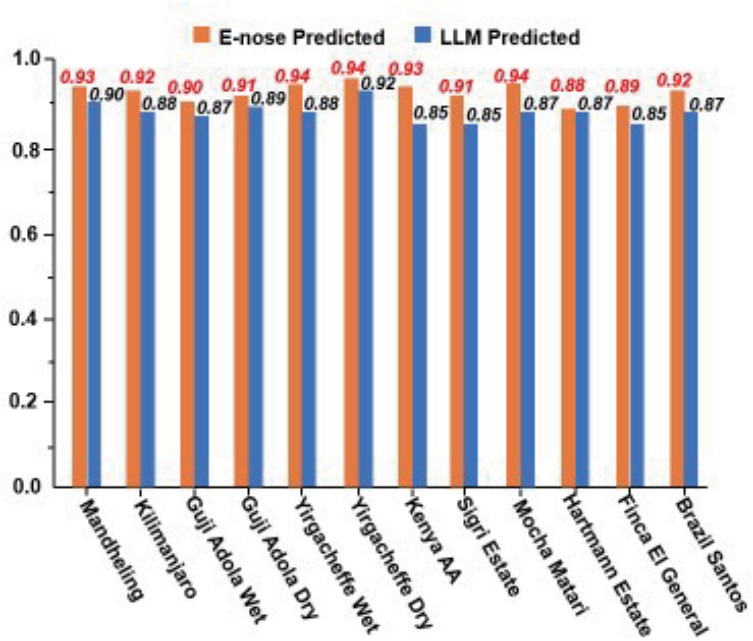



Fig. 10. (Color online) Coffee classification performance with E-Nose and LLM experiments.

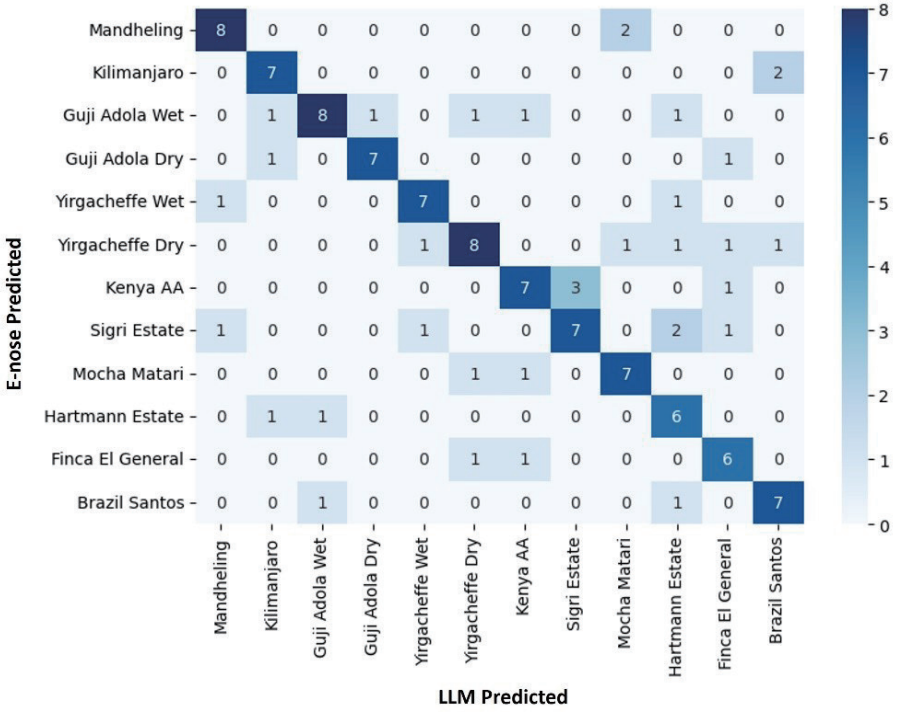



Fig. 11. (Color online) Visualization of classification consistency for coffee samples.

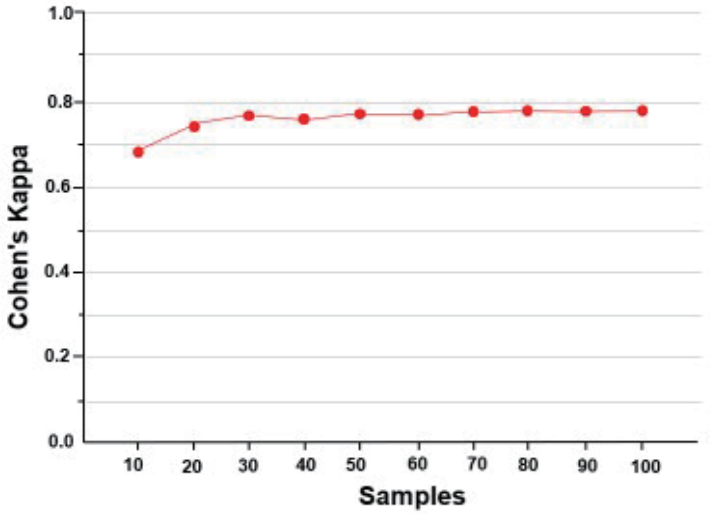



Fig. 12. (Color online) Variation of Cohen's $\kappa$ with different sample sizes.

### 4.4 Discussion

In this study, we proposed a hybrid artificial olfactory system that utilizes an E-Nose for sensing odor data and combines LLMs with knowledge graph models to overcome the bottleneck of traditional odor recognition, which can only classify, but seldom explains, interprets, and reasons; a few findings from our experimental results are discussed below.

(1) Owing to the rapid advancement of AI language models, the LLM selected through our benchmarking experiments may not consistently represent the best available option. For instance, the best performing model in our experiment (i.e., Grok-3) cannot always maintain the highest score. Therefore, the system will continuously add newer evaluations and updated models to maintain the best performance. We have evaluated the ability of each model to interpret coffee aroma after enhancing reasoning capability and compared the results with those of human experts, as shown in Table 4 and Fig. 8. On the BLEU scale, our system reached almost the same level as a human expert, showing that its ability to reason about aroma descriptions is close to that of an expert.

(2) The knowledge distillation technique used in this study can make the LLMs more focused, which effectively improves the accuracy and expertise of the olfactory description of coffee flavor while reducing the computational resource requirements. In our experiments, we found that the LLMs generated after rounds of conversations and reasoning retained the knowledge of the coffee domain while producing a more streamlined model that serves as the core question-and-answer engine of the system. It would be very helpful to apply this approach to the future development of next-generation E-Nose systems based on smaller Edge-AI devices.

(3) The vector retrieval results are mapped onto a knowledge graph whose multidimensional nodes represent variety, origin, processing, roasting methods, flavor attributes, and more;

reasoning is then performed along graph paths, enabling the LLM to generate evidence-backed descriptions. Under the ROUGE-L evaluation metric, this architecture outperforms conventional retrieval schemes by 6–10% and significantly reduces hallucinations.

(4) In real-world deployment, the proposed system supports a high degree of personalization. Because olfactory perception is inherently subjective and users may experience the same odor quite differently, the model draws on selected portions of each user's interaction history as contextual evidence. By continually learning from these exchanges, it infers individual scent preferences: after every instance in which the user provides feedback on an odor description, the system updates the preference profile and re-weights subsequent odor descriptors accordingly, yielding responses tailored to that user's unique olfactory taste.

(5) Despite recent progress, AI-based olfactory reasoning remains technically constrained. Although this study simulates the interplay between odor and memory using a knowledge graph, the system still lacks a holistic, memory-oriented representation of smell, which limits its flexibility. Future research should aim to align computational models more closely with the associative and dynamic nature of human olfactory memory.

## 5. Conclusion

For the evolution of AI, if perception is "seeing," "hearing," and "smelling," then cognition is "understanding" and "thinking." Nowadays, perception and cognition are promoting the comprehensive upgrade of AI in a synergistic way. From the olfactory perception of food to the autonomous optimization of industrial robots, behind every AI application scene is the deep integration of perception and cognition. In this study, we proposed a novel artificial olfactory system architecture that combines E-Nose technology and LLM to simulate human odor perception and reasoning mechanisms, thus addressing the long-standing limitation that previous studies can only identify odors without elucidating their potential correlations. To verify the feasibility of the proposed system, we conducted a case study on coffee aroma interpretation. Therefore, discrete olfactory data were structured for similarity retrieval, and the descriptive capabilities of knowledge graphs and LLM were utilized to generate detailed coffee flavor profiles. Experimental results show that our hybrid system approach overcomes the current limitations of artificial olfactory reasoning and allows the machine to simulate the cognitive mechanism of human abstract thinking and reasoning about odor, which in turn provides a new path for AI-based olfactory research.

For future work, the knowledge-graph-based reasoning enhancement approach demonstrated in this study can be extended to other food-related domains for quality grading and inspection through olfactory analysis. Furthermore, the technique can be similarly deployed in industrial and other odor-critical settings for the real-time detection of hazardous gas leaks and air quality monitoring, where the LLM can convert sensor data into easily interpretable linguistic descriptions. In addition, the novel architecture of combining the E-Nose system with a multimodal LLM provides a fundamental reference for subsequent E-Nose research and facilitates applications in a wider range of fields.

## References


1 T. Liu, L. Guo, M. Wang, C. Su, D. Wang, H. Dong, and W. Wu: Intell. Comput. **2** (2023) 0012. https://doi.org/10.34133/icomputing.0012
2 S. Baruah and D. H. Mazumder: IEEE Trans. Instrum. Meas. **74** (2025) 1. https://doi.org/10.1109/TIM.2025.3547517
3 P. Jia, X. Li, M. Xu, and L. Zhang: Int. J. Bio-Inspired Comput. **23** (2024) 16. https://doi.org/10.1504/IJBIC.2024.136224
4 M. Jiang, N. Li, M. Li, Z. Wang, Y. Tian, K. Peng, and Q. Li: Sensors **24** (2024) 4126. https://doi.org/10.3390/s24134126
5 F. Wu, R. Ma, Y. Li, F. Li, S. Duan, and X. Peng: Sens. Actuators, B **405** (2024) 135272. https://doi.org/10.1016/j.snb.2024.135272
6 X. Ren, Y. Wang, Y. Huang, M. Mustafa, D. Sun, F. Xue, and F. Wu: IEEE Sens. J. **23** (2023) 6027. https://doi.org/10.1109/JSEN.2023.3241842
7 J.-T. Sun and C.-H. Lee: Sens. Mater. **37** (2025) 23. https://doi.org/10.18494/SAM5375
8 S. D. Astuti, I. R. Wicaksono, S. Soelistiono, P. A. D. Permatasari, A. K. Yaqubi, Y. Susilo, and A. Syahrom: Sens. Bio-Sens. Res. **43** (2024) 100632. https://doi.org/10.1016/j.sbsr.2024.100632
9 D. Erwanto, R. F. Rizal, D. E. Yuliana, M. Munir, Y. Trisnoaji, C. Harsito, and S. D. Prasetyo: J. Intell. Syst. Control **3** (2024) 186. https://doi.org/10.56578/jisc030305
10 M. Wang, Y. Chen, D. Chen, X. Tian, W. Zhao, and Y. Shi: Meas. Sci. Technol. **35** (2024) 056004. https://doi.org/10.1088/1361-6501/ad29e4
11 J. Vizcarra, S. Haruta, and M. Kurokawa: Proc. 2024 IEEE 18th Int. Conf. Semant. Comput. (IEEE, 2024) 231–232. https://doi.org/10.1109/ICSC59802.2024.00043
12 C. Peng, F. Xia, M. Naseriparsa, and F. Osborne: Artif. Intell. Rev. **56** (2023) 13071. https://doi.org/10.1007/s10462-023-10465-9
13 A. Menini, L. Lagostena, and A. Boccaccio: Physiology **19** (2004) 101. https://doi.org/10.1152/nips.1507.2003
14 A. Menini: Curr. Opin. Neurobiol. **9** (1999) 419. https://doi.org/10.1016/S0959-4388(99)80063-4
15 R. Zhong, Z. Ji, S. Wang, and H. Chen: Trends Food Sci. Technol. **135** (2024) 104700. https://doi.org/10.1016/j.tifs.2024.104700
16 A. W. Woolley and P. Gupta: Perspect. Psychol. Sci. **19** (2024) 344. https://doi.org/10.1177/17456916231191534
17 D. Vedejová and V. Čavojová: Think. Reason. **28** (2022) 1. https://doi.org/10.1080/13546783.2021.1891967
18 F. Zhang, Z. Zhang, F. Zhuang, Y. Zhao, D. Wang, and H. Zheng: IEEE Trans. Knowl. Data Eng. **36** (2024) 7115. https://doi.org/10.1109/TKDE.2024.3390683
19 K. Kumar, T. Ashraf, O. Thawakar, R. M. Anwer, H. Cholakkal, M. Shah, and S. Khan: arXiv:2502.21321 (2025). https://doi.org/10.48550/arXiv.2502.21321
20 T. Xie, Z. Gao, Q. Ren, H. Luo, Y. Hong, B. Dai, and C. Luo: arXiv:2502.14768 (2025). https://doi.org/10.48550/arXiv.2502.14768
21 T. Han, Z. Wang, C. Fang, S. Zhao, S. Ma, and Z. Chen: arXiv:2412.18547 (2024). https://doi.org/10.48550/arXiv.2412.18547
22 A. Sundararajan and K. Ravirajan: arXiv:2502.07796 (2025). https://doi.org/10.48550/arXiv.2502.07796 (accessed August 2025)
23 A. L. Shaari, A. M. Saad, D. Patil, J. Yanik, W. D. Hsueh, J. A. Eloy, and A. Filimonov: Laryngoscope Investig. Otolaryngol. **10** (2025) e70130. https://doi.org/10.1002/lio2.70130
24 S. Esteban-Romero, I. Martín-Fernández, M. Gil-Martín, and F. Fernández-Martínez: SSRN Working Paper 4912100 (2025). https://ssrn.com/abstract=4912100 (accessed August 2025)
25 M. Schwarz and K. Hamburger: Cogn. Process. **25** (2024) 37. https://doi.org/10.1007/s10339-023-01169-7
26 K. R. Mahmud, L. Wang, S. Hassan, and Z. Zhang: Robot. Auton. Syst. **186** (2025) 104915. https://doi.org/10.1016/j.robot.2025.104915

## About the Authors

**Chung-Hong Lee** (Senior Member, IEEE) received his M.Sc. degree in information technology for manufacturing from the University of Warwick, U.K., in 1994, and his Ph.D. degree in computer science from the University of Manchester, U.K., in 1997. He is currently a professor in the Department of Electrical Engineering, National Kaohsiung University of Science and Technology, Taiwan. His current research interests include artificial intelligence and data mining. (leechung@mail.ee.nkust.edu.tw)

**Hsin-Chang Yang** received his M.Sc. and Ph.D. degrees in computer science and information engineering from National Taiwan University, Taiwan, in 1990 and 1996, respectively. He is currently a professor in the Department of Information Management at the National University of Kaohsiung, Taiwan. His current research interests include neural networks, pattern recognition, information retrieval, and social network analysis. (yanghc@nuk.edu.tw)

**Jun-Teng Sun** received his B.Sc. degree in electrical engineering from Southern Taiwan University of Science and Technology, Taiwan, in 2023. He is currently pursuing his M.Sc. degree in electrical engineering at National Kaohsiung University of Science and Technology, Taiwan. His research focuses on electronic nose systems and AI applications. (f112154173@nkust.edu.tw)

**Zhen-Xin Fu** received his B.Sc. degree in electrical engineering from National Kaohsiung University of Science and Technology, Taiwan, in 2023. He is currently pursuing his M.Sc. degree at National Kaohsiung University of Science and Technology. His research is dedicated to AI applications. (f112154170@nkust.edu.tw)