Bridging Perceptual Gaps in Food NLP: A Structured Approach Using Sensory Anchors

Kana Maruyama
Sony AI
kana.maruyama@sony.com

Angel Hsing-Chi Hwang
University of Southern California
angel.hwang@usc.edu

Tarek R. Besold
Sony AI
tarek.besold@sony.com

Abstract

Understanding how humans perceive and describe food is essential for NLP applications such as semantic search, recommendation, and structured food communication. However, textual similarity often fails to reflect perceptual similarity, which is shaped by sensory experience, wine knowledge, and individual context. To address this, we introduce Sensory Anchors—structured reference points that align textual and perceptual representations. Using Red Wine as a case study, we collect free-form descriptions, metaphor-style responses, and perceptual similarity rankings from participants with varying levels of wine knowledge. These rankings reflect holistic perceptual judgments, with wine knowledge emerging as a key factor. Participants with higher wine knowledge produced more consistent rankings and moderately aligned descriptions, while those with lower knowledge showed greater variability. These findings suggest that structured descriptions based on higher wine knowledge may not generalize across users, underscoring the importance of modeling perceptual diversity. We also find that metaphor-style prompts enhance alignment between language and perception, particularly for less knowledgeable participants. Sensory Anchors thus provide a flexible foundation for capturing perceptual variability in food language, supporting the development of more inclusive and interpretable NLP systems.

1 Introduction

Understanding how humans perceive and describe food is essential for developing NLP-driven applications in food analysis. These include structured food descriptions, personalized recommendations, pairing systems, and models that integrate human sensory perception. While traditional NLP approaches often rely on textual similarity (Agirre et al., 2012; Reimers and Gurevych, 2019), human food perception is influenced by a combination of sensory experience, domain knowledge, cultural context, and personal preference (Majid, 2021).

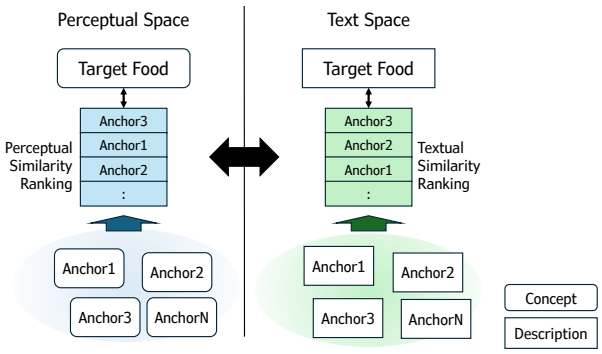

Figure 1: Overview of the Sensory Anchors Framework.

This raises a critical question: Can textual similarity alone adequately reflect perceptual similarity in the human experience of food?

Prior work has explored knowledge-driven representations of food perception, such as expert-defined flavor wheels and structured lexicons (Barbe et al., 2021), as well as consumer-generated taxonomies (Rodríguez-Mendoza et al., 2024). However, these frameworks typically assume a fixed vocabulary and are not easily adaptable to users with different knowledge levels or interpretive styles. Moreover, most perceptual modeling has focused on specific products (e.g., branded wines), whereas category-level modeling (e.g., Red Wine) remains underexplored. Yet, modeling perception at the category level is essential for building generalizable systems that align with human conceptual organization (Rosch et al., 1976) and support perception-aware NLP.

Perceptual framing differs substantially by knowledge level. High-knowledge individuals tend to describe food using structured sensory categories (e.g., Black Fruits, Red Fruits, Oak), while low-knowledge individuals often rely on more impressionistic, less differentiated expressions (Parr et al., 2011). Interestingly, prior research suggests

Proceedings of the Fourth Workshop on NLP for Positive Impact (NLP4PI), pages 86–98

that despite these differences in language, their underlying sensory perceptions may be quite similar (Parr et al., 2011). This implies that the divergence in descriptions arises not from differences in raw sensory sensitivity, but from differences in prior exposure and conceptual organization. As individuals gain more experience, their representations of similarity become increasingly refined—not necessarily because perception itself changes, but because experience reshapes how similarity is conceptualized (McAuley and Leskovec, 2013).

This variability poses a challenge for NLP systems, which must bridge divergent perceptual structures across users (Hamilton et al., 2023; Croijmans and Majid, 2016). Models based solely on textual similarity may fail to capture meaningful sensory similarity, particularly when they overlook how knowledge shapes both perception and language use (Iatropoulos et al., 2018; Speed and Majid, 2020).

Although prior work in computational gastronomy and sensory science has advanced models of flavor networks (Ahn et al., 2011), ingredient pairings (Maruyama and Spranger, 2022), and multisensory integration (Prescott, 2015), few efforts have addressed perceptual modeling at the category level or across diverse knowledge levels. While large-scale food NLP datasets have been enabled by crowdsourced annotations (Callison-Burch, 2009; Snow et al., 2008), the subjectivity of perception—especially among heterogeneous users—remains a core obstacle. We respond to this challenge by proposing a flexible framework that can capture diverse sources of perceptual variation—including domain knowledge, sensory experience, and cultural background—without attempting to reduce perceptual judgments to any one factor.

As an instantiation of this framework, we introduce Sensory Anchors—structured reference points designed to align perceptual similarity judgments and textual expressions across user groups. While this study focuses on participants' knowledge level as the analytic lens, the framework itself is general and can accommodate other sources of perceptual variability, such as sensory experience or cultural background, by substituting the grouping axis and comparative analysis accordingly.

Using Red Wine as a case study, we collect both free-form descriptions and perceptual similarity rankings from participants with varying levels of wine knowledge. For clarity, we refer to participants with higher or lower wine knowledge scores as "high-knowledge" and "low-knowledge" participants, respectively. We treat the similarity rankings as holistic judgments, potentially shaped by a range of factors including direct sensory experience, conceptual associations, and prior exposure. Rather than disentangling these factors, we focus our analysis on how domain knowledge influences the relationship between perception and language.

To support participants in articulating nuanced perceptual similarities, we incorporate metaphor-style prompts that encourage them to frame their judgments using familiar conceptual language.

Our study makes the following contributions:

- We propose a novel framework for analyzing perceptual similarity by systematically comparing textual and perceptual rankings across knowledge levels.
- We show that high-knowledge participants produce more consistent perceptual rankings and moderately aligned descriptions for prototypical Red Wine attributes such as Black Fruits, Red Fruits, and Oak.
- We demonstrate that structured descriptions from high-knowledge participants do not generalize to low-knowledge perception, underscoring the need to model conceptual and perceptual diversity.
- We find that metaphor-style prompts improve alignment between language and perception, especially for low-knowledge participants, highlighting the value of linguistic scaffolding.
- We extend the Sensory Anchors framework to category-level modeling, enabling more robust and knowledge-aware NLP applications.

By bridging the gap between textual and perceptual similarity, this study offers insights that may inform the design of perception-aware NLP systems to support inclusive, interpretable, and user-aligned food communication. Such systems could enhance tasks such as search and retrieval, knowledge-sensitive recommendations, and structured description generation. More broadly, our framework may contribute to applications in dietary education, accessibility, and culturally-aware food design—supporting socially relevant goals aligned with the potential of NLP technologies.

2 Related Work

2.1 Diversity in Food Perception

Food perception is inherently diverse, influenced by factors such as cultural background, prior experience, and individual differences. Studies have examined how multisensory interactions shape food preferences and descriptions (Spence, 2015; Prescott, 2015), how cultural variations contribute to differences in perception (Jeong and Lee, 2021; Ahn et al., 2011), and how linguistic patterns shape food descriptions across cultures (Speed and Majid, 2020).

While these studies highlight the variability in food perception, they often rely on experimental or qualitative methods, lacking systematic computational modeling approaches. Recent efforts have begun to address this gap through computational methods, including cross-lingual analyses of culinary perception (Leng et al., 2019) and machine learning-based modeling of taste perception (Aliya et al., 2024; Androutsos et al., 2024). However, these approaches often depend on predefined taxonomies, which may not fully capture the nuances of food perception across different cultures and individual preferences.

2.2 Structured Representations of Food Perception

Traditional approaches to food description rely on structured sensory lexicons, expert-defined taxonomies, and flavor wheels that provide standardized vocabularies for characterizing sensory experiences (Rodríguez-Mendoza et al., 2024; Su et al., 2022; Lawless and Heymann, 2010). While widely used in professional sensory evaluation, these frameworks often fail to capture the variability and subjectivity found in consumer-generated descriptions.

Expert-oriented frameworks typically use technical terms and fixed categories, whereas consumers tend to describe sensory experiences in more intuitive, emotionally grounded language. This mismatch creates a gap between professional and everyday food descriptions (Croijmans and Majid, 2016; Croijmans et al., 2020).

To address these limitations, recent work has proposed data-driven methods for modeling sensory perception, including the integration of computational approaches into flavor perception analysis (Hamilton, 2022), computational modeling of flavor compounds (Ahn et al., 2011), comparisons between expert and consumer language (Hamilton et al., 2023), and integration of chemical and linguistic data (Prescott, 2015). Further, multimodal embeddings and domain-specific large language models have shown promise for representing food knowledge in structured NLP systems (Rodríguez-Mendoza et al., 2024; Huang et al., 2024).

Despite these advances, modeling fine-grained sensory distinctions remains a challenge. For example, recent work using large language models (LLMs) as virtual tasters has shown that these models tend to produce generic or overly positive descriptions, failing to capture subtle perceptual differences (Torrico, 2025). Similarly, deep learning models trained on whisky reviews—authored by a mix of professional and semi-professional tasters—perform well in identifying descriptors, but the underlying corpora may not reflect the variability found in general consumer language (Miller et al., 2021).

2.3 Crowdsourcing and Annotation for Food NLP

Crowdsourcing has played a central role in food NLP, enabling the large-scale collection of sensory descriptions, ingredient categorizations, and recipe annotations (Min et al., 2019), and has been further expanded through computational gastronomy approaches that leverage user-generated content for modeling food perception (Trattner and Elsweiler, 2017).

However, food perception poses unique challenges due to its subjective nature. A growing body of work shows that individuals with higher domain knowledge produce more structured and precise sensory descriptions than those with less knowledge (Croijmans and Majid, 2016; Parr et al., 2011). Similar patterns have been observed in wine and coffee, where expertise correlates with more consistent and abstract flavor language. These findings highlight the need for modeling strategies that account for differences in knowledge level and descriptive style.

In sensory science, structured reference points such as sensory lexicons and calibrated reference samples are used to enhance consistency and reproducibility in evaluations (Lawless and Heymann, 2010). These techniques provide structured methodologies that can help improve the quality and consistency of data collection in subjective domains like food perception. In NLP, structured annotation formats such as Best-Worst

Scaling have proven effective in improving inter-annotator agreement in sentiment analysis and may be adapted to food-related tasks (Kiritchenko and Mohammad, 2017).

2.4 Positioning Sensory Anchors

Building on the limitations identified above, we propose Sensory Anchors as a structured yet adaptable framework for modeling perceptual similarity in food NLP. Unlike existing approaches that rely on fixed taxonomies or unstructured textual data, Sensory Anchors offer a mechanism for systematically comparing perceptual judgments and textual descriptions across users with different levels of domain knowledge.

The framework centers on category-level reference points (e.g., Black Fruits, Red Fruits, Oak) selected from established sensory taxonomies such as those used in professional tasting protocols. These Anchors serve as consistent points of comparison, enabling perceptual and linguistic responses to be aligned even when participants use diverse descriptive strategies or vocabulary.

By linking perceived similarity and language through interpretable reference categories, Sensory Anchors support the analysis of how sensory concepts are represented across knowledge levels. This suggests their potential usefulness in applications such as food recommendation, search, and structured description generation, where sensitivity to variation in user background and expression is essential.

3 Data Collection and Annotation

This section describes the data collection methodology, the selection of Sensory Anchors, and the annotation process.

3.1 Data Collection Methodology

We conducted a pilot study to investigate how individuals describe and evaluate food perception, recruiting 34 participants through Amazon Mechanical Turk (MTurk). To ensure response quality, we required participants to have a HIT approval rate of $\geq$98% and at least 1,500 approved HITs. All participants were based in the United States. Prior to participation, all participants were presented with a consent form outlining the nature and purpose of the study and the intended use of their responses. Only those who provided informed consent were allowed to proceed.

Each participant completed two main tasks: (1) a food description task and (2) a perceptual similarity ranking task.

In the first task, participants provided open-ended descriptions of the sole target food item (Red Wine) and seven Sensory Anchors, focusing on sensory attributes such as taste, aroma, and texture. Participants were instructed to base their descriptions on their general impressions of each item, for example by recalling the last time they consumed red wine, rather than referring to a specific brand or product label. This approach was designed to elicit intuitive, memory-based representations grounded in personal experience, while avoiding brand-driven or overly idiosyncratic descriptions. In addition to free-text descriptions, participants responded to a series of metaphor-style prompts designed to elicit intuitive associations with specific sensory dimensions. For each food item, they were asked to complete sentences such as "The sweetness of the red wine is like ____." across a set of predefined attributes including basic tastes (e.g., *sweetness*, *bitterness*, *sourness*), texture (e.g., *smoothness*), and intensity-related qualities (e.g., *potency*, *acidity*). If a sensory dimension was not relevant to a given food item, participants were allowed to skip that prompt. A complete list of prompts is provided in the Appendix B.

In the second task, participants ranked the seven Sensory Anchors based on their perceptual similarity in taste and flavor to the target food.

To account for individual differences in domain knowledge, participants completed the Wine Knowledge Assessment Test. We adapted 24 questions from the knowledge test employed in Qi et al. (2024), which was originally developed in Velikova et al. (2015).

Participants were categorized into high-knowledge (18 participants) and low-knowledge (16 participants) groups based on their scores, using the median score (23) as the threshold. Given the relatively small sample size (N = 34), we employed a median split to create a simple and approximately balanced grouping. The distribution was concentrated around a score of 23, with a few participants scoring lower, resulting in a slight asymmetry toward the lower end (see Appendix Figure 2).

Overall, we collected 272 food descriptions (34 participants × 8 food items: Red Wine and 7 Sensory Anchors) and 34 perceptual similarity rankings for Red Wine, forming the dataset for subse-

quent analysis.

3.2 Sensory Anchor Selection

To provide structured stimuli for perceptual comparison, we selected seven Sensory Anchors from established wine flavor categories defined in the WSET tasting framework (WSET, 2020). Each category represents a class of food descriptors commonly used in wine education (e.g., Red Fruits, Citrus Fruits, Oak).

For each participant, one representative food item (e.g., *strawberry*, *orange*, *coffee*) was randomly selected from each sensory category to serve as the anchor. This ensured variation at the item level while maintaining consistent coverage across the seven categories. The selected categories capture key aromatic and taste dimensions relevant to wine perception and are listed in Appendix Table 7, along with their corresponding food items.

3.3 Annotation of Sensory Terms and Description Quality

We manually annotated all descriptions to identify sensory-related terms across seven perceptual categories: Acidity, Aroma, Aftertaste, Flavor, Taste, Weight, and Texture (see Appendix Table 5). Wine-specific attributes (e.g., "Body") were mapped to general categories (e.g., Weight) to ensure compatibility with our cross-domain sensory framework.

Each description was also rated for overall descriptive quality and categorized into one of three levels:

- High: Multiple concrete sensory terms; specific and informative enough to meaningfully distinguish the target item.
- Mid: Generally relevant but lacking detail or precision.
- Low: Vague, generic, or minimally sensory.

To assess annotation reliability, 72 responses (26%) were independently labeled by two trained coders. Inter-rater agreement was moderate ($\kappa = 0.430$; Landis and Koch (1977)), consistent with prior work on free-form sensory descriptions. Disagreements occurred mainly in borderline cases—especially between Mid and High or Low and Mid—reflecting subjective differences in assessing specificity, relevance, and informativeness. For example, annotators sometimes differed on whether vague but technically accurate sensory terms merited mid- or low-quality labels. These cases were resolved through discussion, leading to a shared understanding and refinement of the annotation guidelines.

Following this calibration, the remaining responses were annotated by a single trained coder using the finalized guidelines.

Among all 272 responses, 20.6% were rated as high-quality, 66.9% as mid-quality, and 12.5% as low-quality. These annotations formed the basis for the analysis in Section 4.1, which examined the relationship between knowledge level and descriptive clarity.

4 Experimental Analysis

This section investigates how domain knowledge affects both perceptual similarity judgments and sensory descriptions, using Sensory Anchors as structured reference points. We analyze (1) the quality and content of free-form descriptions, (2) the structure of perceptual similarity rankings and their alignment with textual data, and (3) the implications of these patterns for perception-aware NLP.

4.1 Data Quality and Sensory Word Usage

This analysis focuses on participants' free-form descriptions, which allow for meaningful variation in lexical and structural features.

To assess how domain knowledge affects the quality of sensory descriptions, we compared several textual features between high- and low-knowledge participants. These included word count, lexical diversity (MSTTR), normalized Shannon entropy, and coverage of predefined sensory categories (see Appendix A for details of metrics, and Appendix Table 5 for predefined sensory categories). To further assess descriptive clarity, we examined the distribution of quality labels across groups and conducted a chi-square test of independence.

Table 1 summarizes the comparison of free-form descriptions. High-knowledge participants produced longer descriptions ($p = 0.001$), with greater lexical variety (entropy: $p < 0.001$) and broader sensory category coverage ($p < 0.001$). Lexical diversity did not differ significantly (MSTTR). While these results indicate that domain knowledge is associated with greater structural and topical variation in sensory language, they do not directly assess semantic accuracy or domain-

specific relevance. We acknowledge that metrics such as word count and entropy capture surface-level variation and do not reflect the semantic accuracy or specificity of the descriptions. To address this, we complement the structural analysis with human-annotated quality labels, as discussed below.

Appendix Table 6 shows the distribution of description quality. A chi-square test revealed a significant association between knowledge level and quality ($\chi^2 = 31.303$, $p < 0.001$), indicating that knowledge level is systematically related to descriptive clarity. While mid-quality descriptions were common across groups, low-knowledge participants were more likely to produce low-quality responses. In contrast, high-knowledge participants more often provided specific and structured descriptions that better support perceptual modeling.

These findings suggest that domain knowledge influences not only what is described, but also how clearly and specifically sensory attributes are expressed. This pattern is evident in both structural metrics and human annotation.

4.2 Knowledge-Level Variation in Perceptual and Textual Similarity

We examine how perceptual similarity judgments vary by knowledge level, and how well free-form and metaphor-style responses align with these judgments. Perceptual similarity serves as the ground truth. We assess (1) structural and variability differences in rankings between high- and low-knowledge participants, and (2) alignment between perception and text across input types.

4.2.1 Structure and Variability of Perceptual Similarity Rankings

Participants ranked seven Sensory Anchors by their perceived similarity to Red Wine. According to wine education frameworks (e.g., WSET), Black Fruits and Red Fruits are typical descriptors of Red Wine, while Green Fruits, Citrus, Stone, and Tropical Fruits are more common in White Wine. Oak appears in both.

Table 2 and Table 3 summarize the rankings across knowledge groups. Mode ranks reveal group-level tendencies: for example, both groups most frequently ranked Black Fruits as most similar (mode = 1). Red Fruits also ranked highly in both groups, with a slightly lower mean rank among high-knowledge participants. Oak had a mid-range mode in the high-knowledge group but ranked lower on average in the low-knowledge group.

To further explore distributional differences and interpretation consistency, we analyzed three categories—Red Fruits, Oak, and Green Fruits—selected to reflect different degrees of association with Red Wine. As shown in Appendix Figure 3, Red Fruits was generally perceived as similar across groups, but high-knowledge participants showed occasional divergence, suggesting participants may interpret specific items (e.g., "cranberry" vs. "strawberry") differently within the same category. Oak showed stronger contrasts: high-knowledge participants often rated it moderately, while low-knowledge participants more frequently ranked it as dissimilar. Green Fruits revealed the clearest consistency gap, with high-knowledge participants forming a clear peak and low-knowledge participants exhibiting broader spread.

These findings indicate that domain knowledge shapes not only category-level associations but also how consistently participants apply them. Mode ranks identify dominant perceptual intuitions, while remaining variability underscores item-specific interpretation.

4.2.2 Alignment Between Textual and Perceptual Similarity

To evaluate whether participants' textual responses reflect their perceptual judgments, we computed Spearman's rank correlations between textual and perceptual similarity scores across the seven Sensory Anchors. Perceptual similarity scores were defined as the inverse of the mean rank (1 / Mean Rank), such that anchors perceived as more similar to Red Wine received higher scores. This transformation ensured that both similarity metrics were directionally aligned for correlation analysis.

Textual similarity scores were computed using TF-IDF cosine similarity under two conditions: (1) free-form descriptions, and (2) free-form descriptions combined with metaphor-style responses. These two input types enabled a direct comparison between unconstrained language and language scaffolded by structured prompts. Only participant-generated text was included in the computation of metaphor-style responses; prompt templates were excluded.

We used TF-IDF instead of contextual embeddings to avoid introducing external knowledge

Metric	High-Knowledge Mean	Low-Knowledge Mean	Mann-Whitney U	p-value
Word Count	23.306	21.211	11322.0	$p = 0.001$
MSTTR	0.888	0.901	8336.0	$p = 0.174$
Normalized Shannon Entropy	0.748	0.721	11222.5	$p < 0.001$
Sensory Category Coverage	0.444	0.371	8843.5	$p < 0.001$

Table 1: Comparison of Text Characteristics and Sensory Category Coverage

Rank	Sensory Anchor	Mean Rank	Mode Rank
1	Black Fruits	2.056	1
2	Red Fruits	2.778	3
3	Green Fruits	3.944	4
4	Oak	4.056	4
5	Citrus Fruits	4.778	6
6	Stone Fruits	4.833	6
7	Tropical Fruits	5.556	7

Table 2: Perceptual Similarity Rankings for the High-Knowledge Group: Mean and Mode

Rank	Sensory Anchor	Mean Rank	Mode Rank
1	Black Fruits	2.188	1
2	Red Fruits	3.062	2
3	Green Fruits	3.750	5
4	Citrus Fruits	4.312	7
5	Oak	4.688	6
6	Stone Fruits	4.938	4
7	Tropical Fruits	5.062	6

Table 3: Perceptual Similarity Rankings for the Low-Knowledge Group: Mean and Mode

from pretrained models, ensuring that similarities reflect only participant-generated text.

Table 4 presents the correlation results. In the free-form condition, high-knowledge participants showed moderate alignment between textual and perceptual similarity scores. Low-knowledge participants exhibited weaker and more variable alignment. We also tested whether high-knowledge descriptions could explain the perceptual judgments of low-knowledge participants—a common assumption in prior work. These low correlations suggest that descriptions grounded in domain knowledge may not effectively generalize to users with less expertise or different perceptual frameworks.

The inclusion of metaphor-style responses led to stronger correlations in both groups. Although the differences did not reach conventional thresholds for statistical significance ($p < 0.05$), the trend suggests that structured prompts helped participants—particularly those in the low-knowledge group—produce descriptions whose textual similarity more closely reflected their own perceptual rankings.

Taken together, these findings indicate that domain knowledge facilitates more consistent correspondence between linguistic and perceptual similarity structures. However, when guided by metaphor-style prompts, even participants with less domain knowledge were able to generate descriptions that more closely matched their own perceptual judgments. This highlights the potential value of structured elicitation for improving the correspondence between language and perception in modeling applications.

4.3 Summary: Sensory Anchors for Perception-Aware NLP

Our findings demonstrate that Sensory Anchors provide an effective framework for analyzing how perceived similarity is shaped by domain knowledge. By examining sensory descriptions, perceptual similarity rankings, and the relationship between the two, we identify three key insights.

First, high-knowledge participants produced more specific and structured sensory descriptions, as evidenced by both lexical measures and annotation-based quality ratings. This was accompanied by more stable perceptual similarity rankings, particularly for categories strongly associated with Red Wine, such as Red Fruits and Oak. However, variation persisted in how individual items were interpreted, even within these categories, suggesting that domain knowledge does not fully eliminate interpretive diversity.

Second, descriptions produced by high-knowledge participants did not generalize well to the perceptual judgments of low-knowledge participants. Correlations across groups were low, challenging the assumption that language grounded in expert discourse can reliably explain perceptual similarity for less experienced users.

Third, metaphor-style scaffolding improved the correspondence between language and perception in both groups. Notably, participants with lower domain knowledge—who showed weaker alignment in the free-form condition—produced metaphor-style responses that more closely reflected their

Comparison	Spearman ρ	p-value
Free-Form Descriptions		
High-knowledge Text vs. High-knowledge Perceptual	0.536	0.215
Low-knowledge Text vs. Low-knowledge Perceptual	0.286	0.535
High-knowledge Text vs. Low-knowledge Perceptual	0.357	0.432
Free-Form + Metaphor-Style Responses		
High-knowledge Text vs. High-knowledge Perceptual	0.679	0.094
Low-knowledge Text vs. Low-knowledge Perceptual	0.679	0.094
High-knowledge Text vs. Low-knowledge Perceptual	0.500	0.253

Table 4: Spearman rank correlations between textual and perceptual similarity scores, computed over the seven Sensory Anchors for each language condition and participant group.

own perceptual judgments. This suggests that structured prompts can help elicit more perceptually grounded language, particularly when prior knowledge is limited.

Together, these results demonstrate that Sensory Anchors offer a useful framework for analyzing perceptual variation and its relationship to language across knowledge levels. They underscore the importance of domain knowledge and linguistic scaffolding in the design of perception-aware NLP systems.

While this study focused on Red Wine as a case domain, the Sensory Anchors framework is designed to be applicable to other food categories with structured sensory representations.

5 Conclusion & Future Work

This study investigated how domain knowledge shapes food perception and description, introducing Sensory Anchors as structured reference points for modeling perceptual similarity in language. Analyses of participants' descriptions and similarity rankings indicate that perceptual structures vary across knowledge levels, and that descriptions from high-knowledge participants may not generalize well to those with lower knowledge. Metaphor-style prompts improved alignment in both groups, highlighting the role of linguistic scaffolding in supporting consistent mappings between perception and language.

Sensory Anchors offer a flexible and interpretable framework for linking textual and perceptual representations in food-related NLP. Although this study focused on Red Wine and used wine knowledge as the primary axis of variation, the framework is not inherently limited to domain knowledge. It can extend to other sources of perceptual variation—such as sensory experience, cultural background, or affective associations. Importantly, our study deliberately targeted category-level rather than instance-level perception. This design allows us to investigate how people conceptualize and describe broad sensory categories (e.g., Red Wine) based on general experience, which is crucial for building scalable, knowledge-sensitive, and conceptually robust NLP systems. Applications include inclusive recommendation and retrieval systems, culturally adaptive food communication, food and beverage pairing support, and personalized sensory education tools—advancing the broader goal of aligning language with perception across diverse user groups.

Limitations. This study has several limitations. First, the sample size was relatively small ($N = 34$) and restricted to U.S.-based participants, limiting generalizability and cultural diversity. Second, participants were grouped by a median split (threshold = 23), which may obscure fine-grained differences near the cutoff. Third, while we assessed descriptive quality using structural metrics and human annotation, we did not evaluate semantic accuracy or domain-specific vocabulary usage, which could clarify how meaning varies with knowledge. Lastly, our exploratory correlation analyses did not include correction for multiple comparisons, raising the risk of spurious correlations.

Future Work. Future research could extend the framework to other food domains and investigate perception across cultural or linguistic groups. Incorporating finer group definitions (e.g., percentile-based or continuous modeling) and controlled experimental conditions may help disentangle different sources of perceptual variability. Additionally, integrating semantic evaluation techniques could further improve our understanding of how perceptual similarity is reflected in language.

Ethical and Societal Implications

Ethical Considerations and Limitations

Our dataset was collected through crowd-sourced tasks involving perceptual similarity judgments and textual descriptions. While quality control measures were implemented on MTurk, such as minimum approval rates and task completion thresholds, the participant sample may still be biased toward specific demographic groups. This limits the generalizability of our findings and highlights the need for broader participant recruitment in future studies (Ross et al., 2010; Snow et al., 2008).

Additionally, our approach relies on structured Sensory Anchors that draw from expert-oriented taxonomies, such as those defined by the Wine & Spirit Education Trust (WSET, 2020). While these frameworks offer consistency and interpretability, they may not fully capture culturally diverse interpretations of food perception (Prescott, 1998; Spence, 2015). Future work could expand the design of Sensory Anchors by incorporating regionally and culturally grounded descriptors to support more inclusive modeling of perceptual variability.

Although participants also provided confidence ratings alongside their perceptual similarity judgments, we excluded these scores from the current analysis due to their subjective nature and the complexity of modeling inter-individual calibration. Future work may leverage confidence information for weighting similarity rankings, interpreting alignment strength, or identifying perceptual uncertainty, particularly in low-knowledge populations.

Overall, our study underscores the importance of considering both participant diversity and the conceptual framing of perceptual categories when designing perception-aware NLP systems.

Societal Impact and Accessibility

This research contributes to more equitable and accessible food-related NLP systems by modeling perceptual variability across users. Representing food perception in a structured way can improve the quality and clarity of textual food descriptions, which is particularly valuable for individuals with olfactory or gustatory impairments. Prior studies have shown that sensory disorders can significantly affect dietary decisions, quality of life, and food-related communication (Croy et al., 2014; Miwa et al., 2001). By enabling the generation and retrieval of interpretable descriptions that reflect user-specific sensory expectations, our approach supports more personalized and inclusive recommendation systems.

In addition, NLP and AI-driven structured knowledge representation have been explored in accessibility applications, including assistive recommendation systems (Gavat et al., 2023; Christensen et al., 2019). Recent work on knowledge graph–based systems has shown that structured information can improve retrieval for health-related queries, including those related to smell and taste disorders (Tauqeer et al., 2023). Our research contributes to this direction by modeling perceptual similarity in a structured format, enabling the identification of perceptual gaps across user groups. This facilitates the collection of more inclusive and user-tailored food descriptions, making food-related NLP systems better equipped to accommodate diverse sensory profiles.

Our findings demonstrate that incorporating perceptual similarity into food-related NLP can help structure sensory information in ways that are more interpretable and actionable. This improves usability across users with varying needs, preferences, and sensory capabilities.

Environmental Considerations

As NLP systems become increasingly integrated into food-related domains, it is important to consider their environmental impact. Large language models (LLMs) offer powerful capabilities but often require resource-intensive fine-tuning and inference. While our study does not directly evaluate computational efficiency, it contributes toward more sustainable NLP practices by introducing a framework that leverages lightweight, structured inputs—such as perceptual rankings and targeted textual prompts—to reduce reliance on large-scale model adaptation.

In particular, the structured nature of Sensory Anchors enables in-context learning and few-shot adaptation, which can reduce the need for full retraining and minimize computational overhead. This aligns with broader efforts to develop environmentally responsible AI systems (Strubell et al., 2019; Schwartz et al., 2020). Future research may explore the integration of perceptual data into prompt-based learning strategies, further advancing the efficiency and scalability of food-related NLP applications.

References

Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 2012. SemEval-2012 task 6: A pilot on semantic textual similarity. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 385–393, Montréal, Canada. Association for Computational Linguistics.

Yong-Yeol Ahn, Sebastian E. Ahnert, James P. Bagrow, and Albert-László Barabási. 2011. Flavor network and the principles of food pairing. *Scientific Reports*, 1(1):196.

Aliya, Shi Liu, Danni Zhang, Yufa Cao, Jinyuan Sun, Shui Jiang, and Yuan Liu. 2024. Research on the Evaluation of Baijiu Flavor Quality Based on Intelligent Sensory Technology Combined with Machine Learning. *Chemosensors*, 12(7):125.

Lampros Androutsos, Lorenzo Pallante, Agorakis Bompotas, Filip Stojceski, Gianvito Grasso, Dario Piga, Giacomo Di Benedetto, Christos Alexakos, Athanasios Kalogeras, Konstantinos Theofilatos, Marco A. Deriu, and Seferina Mavroudi. 2024. Predicting multiple taste sensations with a multiobjective machine learning method. *npj Science of Food*, 8(1):47.

Jean-Christophe Barbe, Justine Garbay, and Sophie Tempère. 2021. The Sensory Space of Wines: From Concept to Evaluation and Description. A Review. *Foods*, 10(6):1424.

Chris Callison-Burch. 2009. Fast, Cheap, and Creative: Evaluating Translation Quality Using Amazon's Mechanical Turk. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 286–295, Singapore. Association for Computational Linguistics.

Heidi Christensen, Kristy Hollingshead, Emily Prud'hommeaux, Frank Rudzicz, and Keith Vertanen, editors. 2019. *Proceedings of the Eighth Workshop on Speech and Language Processing for Assistive Technologies*. Association for Computational Linguistics, Minneapolis, Minnesota.

Ilja Croijmans, Iris Hendrickx, Els Lefever, Asifa Majid, and Antal Van Den Bosch. 2020. Uncovering the language of wine experts. *Natural Language Engineering*, 26(5):511–530.

Ilja Croijmans and Asifa Majid. 2016. Not All Flavor Expertise Is Equal: The Language of Wine and Coffee Experts. *PLoS ONE*, 11(6):e0155845.

Ilona Croy, Steven Nordin, and Thomas Hummel. 2014. Olfactory disorders and quality of life–an updated review. *Chemical Senses*, 39(3):185–194.

Inge Gavat, Andreea Griparis, and Svetlana Segarceanu. 2023. Natural language processing in assistive technologies. *The Romanian Journal of Technical Sciences. Applied Mechanics.*, 68(2-3):129–140.

Leah Hamilton. 2022. Translating Sensory Perceptions: Existing and Emerging Methods of Collecting and Analyzing Flavor Data.

Leah M. Hamilton, Clinton L. Neill, and Jacob Lahne. 2023. Flavor language in expert reviews versus consumer preferences: An application to expensive American whiskeys.

Tenghao Huang, Donghee Lee, John Sweeney, Jiatong Shi, Emily Steliotes, Matthew Lange, Jonathan May, and Muhao Chen. 2024. FoodPuzzle: Developing Large Language Model Agents as Flavor Scientists. *Preprint*, arXiv:2409.12832.

Georgios Iatropoulos, Pawel Herman, Anders Lansner, Jussi Karlgren, Maria Larsson, and Jonas K. Olofsson. 2018. The language of smell: Connecting linguistic and psychophysical properties of odor descriptors. *Cognition*, 178:37–49.

Sohyun Jeong and Jeehyun Lee. 2021. Effects of cultural background on consumer perception and acceptability of foods and drinks: A review of latest cross-cultural studies. *Current Opinion in Food Science*, 42:248–256.

Svetlana Kiritchenko and Saif Mohammad. 2017. Best-Worst Scaling More Reliable than Rating Scales: A Case Study on Sentiment Intensity Annotation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 465–470, Vancouver, Canada. Association for Computational Linguistics.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Harry T. Lawless and Hildegarde Heymann. 2010. *Sensory Evaluation of Food: Principles and Practices*. Food Science Text Series. Springer, New York, NY.

Xuehui Leng, Masanao Ochi, Takeshi Sakaki, Junichiro Mori, and Ichiro Sakata. 2019. A cross-lingual analysis on culinary perceptions to understand the cross-cultural difference. In *Proceedings of the Symposium Interpretable AI for Well-being: Understanding Cognitive Bias and Social Embeddedness co-located with Association for the Advancement of Artificial Intelligence 2019 Spring Symposium (AAAI-Spring Symposium 2019), Stanford, CA, March 25-27, 2019*, volume 2448 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Asifa Majid. 2021. Human Olfaction at the Intersection of Language, Culture, and Biology. *Trends in Cognitive Sciences*, 25(2):111–123.

Kana Maruyama and Michael Spranger. 2022. Interpretable relational representations for food ingredient recommendation systems. In *Proceedings of the 13th International Conference on Computational Creativity, ICCC 2022, Bozen-Bolzano, Italy, June 27 - July 1, 2022*, pages 271–275. Association for Computational Creativity (ACC).

Julian McAuley and Jure Leskovec. 2013. From Amateurs to Connoisseurs: Modeling the Evolution of User Expertise through Online Reviews. *Preprint*, arXiv:1303.4402.

Chreston Miller, Leah Hamilton, and Jacob Lahne. 2021. Sensory Descriptor Analysis of Whisky Lexicons through the Use of Deep Learning. *Foods*, 10(7):1633.

Weiqing Min, Shuqiang Jiang, Linhu Liu, Yong Rui, and Ramesh Jain. 2019. A survey on food computing. *ACM Comput. Surv.*, 52(5).

Takaki Miwa, Mitsuru Furukawa, Toshiaki Tsukatani, Richard M. Costanzo, Laurence J. DiNardo, and Evan R. Reiter. 2001. Impact of Olfactory Impairment on Quality of Life and Disability. *Archives of Otolaryngology–Head & Neck Surgery*, 127(5):497–503.

W. V. Parr, M. Mouret, S. Blackmore, T. Pelquest-Hunt, and I. Urdapilleta. 2011. Representation of complexity in wine: Influence of expertise. *Food Quality and Preference*, 22(7):647–660.

John Prescott. 1998. Comparisons of taste perceptions and preferences of Japanese and Australian consumers: Overview and implications for cross-cultural sensory research. *Food Quality and Preference*, 9(6):393–402.

John Prescott. 2015. Multisensory processes in flavour perception and their influence on food choice. *Current Opinion in Food Science*, 3:47–52.

Xiaoxiao Qi, Wen Chang, Anyu Liu, Jie Sun, and Mengyu Fan. 2024. Exploring the influence of emotionality and expertise on online wine reviews: Does greater knowledge lead to less review? *International Journal of Contemporary Hospitality Management*, 36.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Anggie V. Rodríguez-Mendoza, Santiago Arbeláez-Parra, Rafael Amaya-Gómez, and Nicolas Ratkovich. 2024. Flavor Wheel Development from a Machine Learning Perspective. *Foods*, 13(24):4142.

Eleanor Rosch, Carolyn Mervis, Wayne Gray, David Johnson, and Penny Braem. 1976. Basic objects in natural categories. *Cognitive Psychology - COG PSYCHOL*, 8:382–439.

Joel Ross, Lilly Irani, M. Six Silberman, Andrew Zaldivar, and Bill Tomlinson. 2010. Who are the crowdworkers? shifting demographics in mechanical turk. In *CHI '10 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '10, pages 2863–2872, New York, NY, USA. Association for Computing Machinery.

Roy Schwartz, Jesse Dodge, Noah A. Smith, and Oren Etzioni. 2020. Green AI. *Commun. ACM*, 63(12):54–63.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Ng. 2008. Cheap and Fast – But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 254–263, Honolulu, Hawaii. Association for Computational Linguistics.

Laura J. Speed and Asifa Majid. 2020. Grounding language in the neglected senses of touch, taste, and smell. *Cognitive Neuropsychology*, 37(5-6):363–392.

Charles Spence. 2015. Multisensory Flavor Perception. *Cell*, 161(1):24–35.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and Policy Considerations for Deep Learning in NLP. *Preprint*, arXiv:1906.02243.

Xiaoxia Su, Miao Yu, Simin Wu, Mingjuan Ma, Hongxu Su, Fei Guo, Qi Bian, and Tianyi Du. 2022. Sensory lexicon and aroma volatiles analysis of brewing malt. *npj Science of Food*, 6(1):20.

Amar Tauqeer, Ismaheel Hammid, Sareh Aghaei, Parvaneh Parvin, Elbrich M. Postma, and Anna Fensel. 2023. Smell and Taste Disorders Knowledge Graph: Answering Questions Using Health Data. *Expert Systems with Applications*, 234:121049.

Damir D. Torrico. 2025. The Potential Use of ChatGPT as a Sensory Evaluator of Chocolate Brownies: A Brief Case Study. *Foods (Basel, Switzerland)*, 14(3):464.

Christoph Trattner and David Elsweiler. 2017. Food Recommender Systems: Important Contributions, Challenges and Future Research Directions. *Preprint*, arXiv:1711.02760.

Natalia Velikova, Roy D. Howell, and Tim Dodd. 2015. The development of an objective wine knowledge scale: The item response theory approach. *International Journal of Wine Business Research*, 27(2):103–124.

WSET. 2020. Wines of the world. Accessed: 2025-03-04.

A Metrics Calculation

To quantitatively assess textual characteristics and sensory coverage, we employed the following measures:

Word Count: The total number of words in each participant's response. Stopwords were not removed to reflect natural language use.

Mean Segmental Type-Token Ratio (MSTTR): A measure of lexical diversity, calculated by dividing the text into fixed-length segments and computing the average Type-Token Ratio (TTR) across all segments.

Normalized Shannon Entropy: A measure of information richness, computed as follows:

$$H_{\text{norm}} = \frac{-\sum_i p_i \log_2 p_i}{\log_2 N} \quad (1)$$

where p_i represents the probability of each unique word, and N is the total number of words in the description. This normalization ensures comparability across varying text lengths.

Sensory Category Coverage Ratio: The proportion of predefined sensory categories (Section 3.3) mentioned in each description, calculated as:

$$\text{Coverage} = \frac{\text{Unique sensory categories mentioned}}{\text{Total predefined sensory categories}} \quad (2)$$

These measures provide a structured approach for analyzing how knowledge levels influence food descriptions at different linguistic and perceptual levels. The results from this section establish the foundation for the perceptual similarity analysis in Section 4.2.

B Metaphor-Style Prompt List

Participants completed the following sentence templates for each sensory anchor and food item:

- The overall taste of the [food] is like ___.
- The sweetness of the [food] is like ___.
- The saltiness of the [food] is like ___.
- The sourness of the [food] is like ___.
- The bitterness of the [food] is like ___.
- The umami of the [food] is like ___.
- The smoothness of the [food] is like ___.
- The potency of the [food] is like ___.
- The acidity of the [food] is like ___.

C Additional Tables and Figures

Sensory Category	Example Words
Acidity	*little tangy, balances the acidity*
Aroma	*earthy, floral*
Aftertaste	*dry finish*
Flavor	*dark fruits, roasted nuts*
Taste	*sweet, deep, slightly bitter*
Weight	*rich, bold, full-bodied*
Texture	*smooth, creamy, velvety*

Table 5: Annotated Sensory Categories—Examples of sensory-related words.

Knowledge Level	High (%)	Mid (%)	Low (%)
High-Knowledge	20.8	77.1	2.1
Low-Knowledge	20.3	55.5	24.2

Table 6: Distribution of Description Quality by Knowledge Level

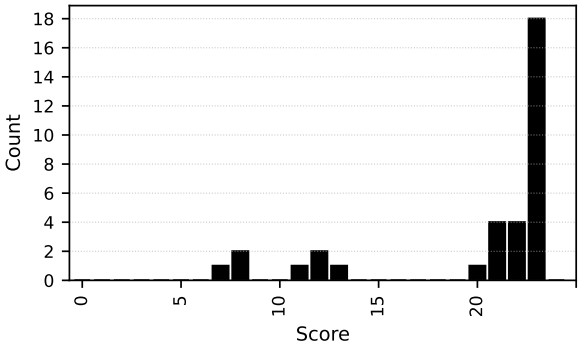

Figure 2: Distribution of participant scores (out of 24) on the wine knowledge test.

Sensory Anchor Category	Example Food Items
Green Fruits	Apple, Gooseberry, Pear, Grape
Citrus Fruits	Grapefruit, Lemon, Lime, Orange
Stone Fruits	Peach, Apricot, Nectarine
Tropical Fruits	Banana, Lychee, Mango, Melon, Passion Fruit, Pineapple
Red Fruits	Redcurrant, Cranberry, Raspberry, Strawberry, Red Cherry, Red Plum
Black Fruits	Blackcurrant, Blackberry, Blueberry, Black Cherry, Black Plum
Oak	Vanilla, Cloves, Coconut, Chocolate, Coffee

Table 7: Each Sensory Anchor Category and its corresponding items. One item was randomly selected from each category.

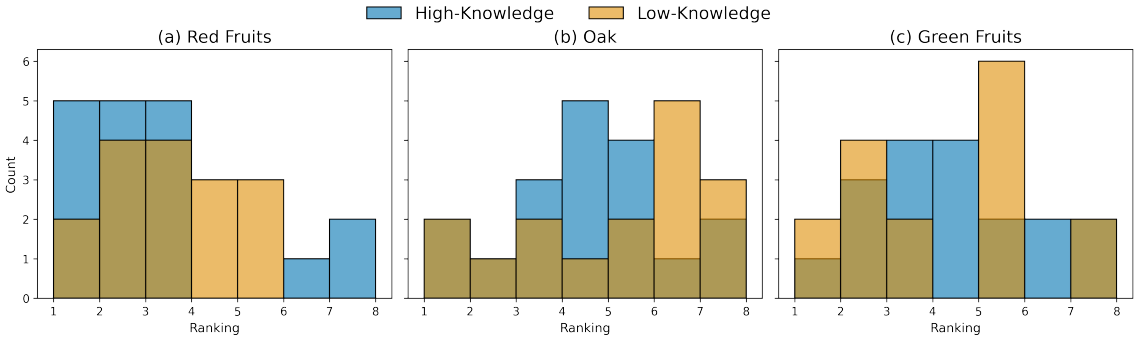

Figure 3: Distribution of perceptual similarity rankings for three sensory categories (Red Fruits, Oak, and Green Fruits) across knowledge groups.Each subplot displays the frequency of each assigned rank (1 = most similar) within each group.