Text2Smell: Emergent Representations of Human Olfactory Perception in Large Language Models

Anonymous authors
Paper under double-blind review

Abstract

Large language models (LLMs) trained exclusively on text have recently demonstrated emergent capacities to perceive the world as humans do, suggesting that linguistic co-occurrence statistics implicitly encode aspects of human sensory experience. Yet, whether such models capture the structure of olfactory perception, one of the most complex and least understood human senses, remains unknown. In this work, we investigate whether state-of-the-art LLMs can predict human smell perception purely from linguistic cues and how their representations compare to those of molecular transformer models explicitly trained on chemical structure. We prompt LLMs to provide perceptual olfactory ratings to odorants, and evaluate their outputs against human ratings across several datasets. Surprisingly, we find that LLMs exhibit strong alignment with human perceptual judgments, comparable to, and in most cases exceeding, the performance of specialized molecular transformers. These results indicate that linguistic knowledge alone carries rich latent structure about human olfaction, bridging the gap between language and chemical perception. Our findings position LLMs as powerful linguistically grounded perceptual models and open new directions for studying sensory grounding and cross-modal representation learning through language.

1 Introduction

Recent advances in large language models (LLMs) have shown that models trained purely on text can acquire surprisingly broad sensory knowledge, from visual attributes to sound semantics (Marjieh et al., 2024; Zhang et al., 2022a; Siedenburg & Saitis, 2023). Through large-scale language pretraining, these models develop internal representations that capture rich semantic, perceptual, and commonsense regularities present in human descriptions of the world Shiono et al. (2025). Although they are never exposed to raw sensory data, LLMs demonstrate an emergent capacity to reason about sensory concepts and make judgments that are consistent with human perception, for example, describing textures as "rough" or "smooth" Tu et al. (2025), or predicting emotional valence from speech Lalk et al. (2025). This suggests that the linguistic co-occurrence patterns embedded in massive text corpora implicitly encode perceptual relationships, providing a bridge between symbolic knowledge and the structure of human experience. Understanding the extent and limits of this implicit sensory grounding is therefore crucial for advancing language models toward genuine perceptual understanding and for uncovering how human sensory knowledge is reflected in language itself.

Among all human senses, *olfaction*, the sense of smell, represents a uniquely complex and challenging domain for computational modeling. Unlike vision or audition, where perceptual responses are grounded in well-characterized physical dimensions such as wavelength or frequency, the mapping between molecular structure and olfactory experience remains poorly understood (Saini & Ramanathan, 2022). Recent works in chemoinformatics, computational olfaction, and molecular transformer models have aimed to predict odor qualities from molecular features using handcrafted descriptors, graph-based neural networks, or large transformer-based encoders (Lee et al., 2023; Ravia et al., 2020; Taleb et al., 2024; Lötsch et al., 2019). While these approaches capture chemical commonalities, they often fail to fully reflect the subjective nature of human smell perception. For instance, two molecules with similar structural features may evoke vastly different perceptual impressions, and conversely, perceptually similar odors may arise from structurally dissimilar compounds Sharma et al. (2019); Sell (2006); Boesveldt et al. (2010). This long-standing "structure–odor

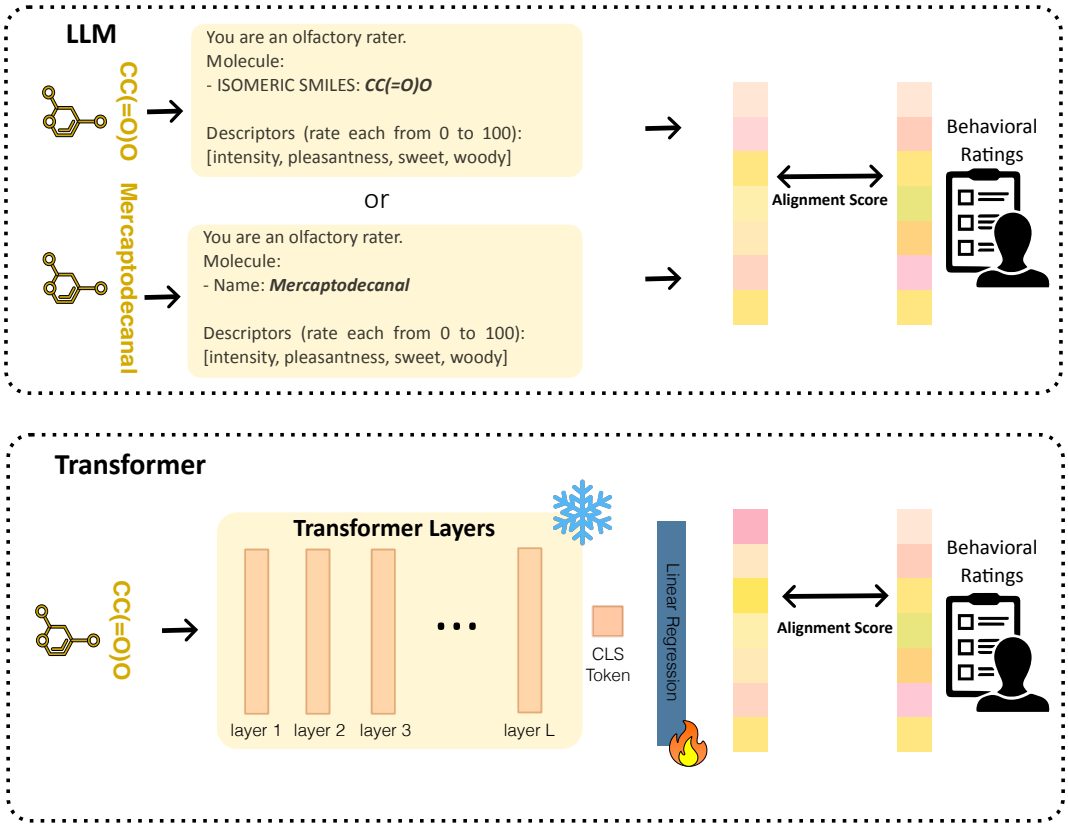

Figure 1: **Overview of the main methodology.** Top: LLMs are prompted with either the common names or SMILES representations of molecules and instructed to provide perceptual ratings similar to human judgments. Bottom: Representations of the same molecules are extracted from specialized molecular transformers trained on chemical structures.

gap" motivates an alternative perspective: rather than predicting perception solely from molecular structure, *can we leverage the rich perceptual knowledge embedded in language to predict how humans experience the smells of different odorants with diverse molecular structures?* In this work, we explore whether LLMs can predict human olfactory perception purely from linguistic cues, and how their representations compare to those of specialized chemical models. We prompt state-of-the-art LLMs (e.g., GPT-4, Gemini) to rate odorants along perceptual dimensions and semantic descriptors (e.g., pleasantness, sweet, woody, fishy), given either the odorant's common name or molecular representations. We then evaluate their outputs against human behavioral ratings from large-scale psychophysical datasets and against predictions derived from pre-trained molecular transformers. Our analyses reveal that LLMs exhibit robust alignment with human olfactory judgments, often outperforming the predictive performance of chemical structure–based models.

In this paper, we make three main contributions:

- We introduce a systematic framework for evaluating the alignment between LLMs and human olfactory perception, combining large-scale psychophysical data with model-prompted odor ratings
- We provide the first direct comparison between text-only LLMs and molecular transformer models in predicting perceptual ratings across multiple semantic dimensions.
- We demonstrate that LLMs, despite never being explicitly fine-tuned for this task, can serve as linguistically grounded perceptual models, capturing cross-modal regularities that emerge from human descriptions of sensory experience.

Dataset	# of Participants	# of Odorants	Rating Type	Rating Range	# of Descriptors
Keller (Keller & Vosshall, 2016)	55	420	Continuous	[0,100]	23
Sagar (Sagar et al., 2023)	3	125	Continuous	[-1,1]	15
Leffingwell (Sanchez-Lengeling et al., 2019)	1	3522	Binary	{0,1}	113
Snitz (Snitz et al., 2013)	139	83	Similarity	[0,1]	NA

Table 1: **All the datasets used in this study.** The number of odorants and perceptual descriptors shown corresponds to those retained for analysis following data cleaning and curation.

Together, these contributions bridge the gap between language and olfaction, highlighting that perceptual knowledge is implicitly encoded in linguistic co-occurrence patterns. Our work opens new directions for multimodal research on sensory grounding in language, suggesting that LLMs can function not only as generators of linguistic structures but also as proxies for human perceptual cognition.

2 Related Works

This study investigates how well LLMs align with human perception in the domain of olfaction. We situate our work at the intersection of three research threads: (i) using LLMs to model human perception and cognition, (ii) employing LLMs for molecular representation and reasoning, and (iii) modeling olfactory perception through both linguistic and chemical representations. After describing each of these three threads, we describe the research gap that we will address in this paper.

LLMs as Models of Human Perception. A growing body of research examines whether LLMs capture aspects of human perception and subjective experiences. Marjieh et al. (2024) show that LLMs can predict human similarity judgments across six sensory modalities, including color, pitch, timbre, loudness, taste, and speech sounds, revealing that language-only models recover canonical perceptual structures even without direct sensory grounding. However, Xu et al. (2025) systematically analyzed conceptual representations from non-sensorimotor dimensions (e.g., valence, dominance) to sensorimotor domains (e.g., auditory, visual, olfactory) across a broad range of concepts, and observed a gradual decline in similarity between LLM-derived and human-derived representations, with the strongest divergences emerging in sensorimotor domains. Similarly, Hicke et al. (2025) analyzes sensory language in model-generated narratives and reports systematic mismatches with human usage; all models generate stories that differ significantly from human usage of sensory language, but the direction of these differences varies considerably between model families.

LLMs for Molecular Representation. Recent efforts have extended LLMs beyond natural language to the chemical and molecular domains, enabling their application to chemistry-related tasks such as reaction prediction Liu et al. (2024); M. Bran et al. (2024). In more recent works, Zhuang et al. (2025) introduces a framework for molecular structure elucidation, extending LLM's coverage of the chemical structure space. In property prediction, Xian et al. (2025) presents MolRAG, a retrieval-augmented generation framework that grounds textual reasoning in molecular databases, achieving stronger generalization and interpretability. Extending to multi-objective optimization, Dey et al. (2025) introduce the first instruction-tuned family of LLMs for molecule optimization, which outperforms state-of-the-art baselines on both in- and out-of-domain tasks.

Modeling Olfactory Perception with LLMs Olfactory perception remains largely underexplored in machine learning and, in particular, in the study of LLMs. To the best of our knowledge, the only prior work examining LLM–human alignment in this domain is SniffAI Zhong et al. (2024). However, SniffAI focuses solely on textual odor descriptions, with no link to the underlying chemical or molecular structures, and is based on a small dataset of 20 odor objects from 4 families rated by 40 participants. In contrast, our work evaluates alignment using the actual molecular stimuli that constitute odorants, providing a more direct examination of how LLM representations relate to human olfactory perception.

Model	Input Type	# of Parameters	Base Model	Dataset	Layers
MoLFormer-XL (Ross et al., 2022)	SMILES	45M	Transformer	ZINC, PubChem	12
MTL-BERT (Zhang et al., 2022b)	SMILES	6M	BERT	ChEMBL	8
ChemBERTa (Chithrananda et al., 2020)	SMILES	44M	RoBERTa	PubChem	6

Table 2: **The pre-trained molecular transformer models used in this study and their characteristics**. All models are publicly available via HuggingFace. The "Dataset" column indicates the training dataset used for each model.

Research Gap Although understanding sensory perception in LLMs is important, current research is still limited. Studies on LLMs and human perception have focused primarily on visual, auditory, or tactile modalities, without examining the uniquely complex case of olfaction. In parallel, molecular transformer models achieve strong performance in structural and physicochemical prediction tasks but lack access to the subjective and linguistic dimensions of human smell perception (Taleb et al., 2024). Previous attempts to connect language and olfaction, such as SniffAI (Zhong et al., 2024), rely solely on textual odor descriptions and small-scale datasets, without linking model predictions to actual molecular stimuli or human psychophysical data. Our work bridges these gaps by directly comparing LLM-derived and molecule-based representations of olfactory perception, providing the first large-scale evaluation of linguistic sensory alignment in the olfactory domain.

3 Background

3.1 Olfactory Perception Task

Olfactory perception refers to the subjective experience of smell and is typically examined through psychophysical experiments. Existing datasets have been collected using three main approaches. In the first, participants smell individual odorants and provide continuous ratings along perceptual dimensions (e.g., pleasantness, intensity, sweetness, and woodiness). In the second approach, participants provide binary ratings for a predefined set of descriptors, indicating whether a perceptual quality (e.g., fruity, sweet) is present (1) or absent (0) for each odorant. In the third approach, participants, given a range, judge the similarity between pairs of odorants, assigning numerical scores that quantify how similar two odors smell.

Throughout this paper, we refer to these responses collectively as perceptual or behavioral ratings. Together, these ratings characterize how humans perceive molecular stimuli and form the foundation for modeling perceptual similarity and structure in the olfactory domain.

3.2 Molecular Representations

Each odorant can consist of a single molecule, referred to as a mono-molecular odorant, or a combination of multiple molecules, forming a mixture. In this work, we focus exclusively on odorants composed of a single molecule. Each such odorant can be represented either by its common name (e.g., phenylethyl alcohol) or by string-based molecular encodings such as SMILES(Simplified Molecular Input Line Entry System) Weininger (1988) . SMILES encodes a molecule's structure as a linear string describing its atoms and bonds. Chemical models typically rely on these structural representations to infer molecular properties, whereas LLMs can process both textual names and symbolic strings. This dual encoding enables us to investigate how linguistic and structural information each contributes to predicting human olfactory perception.

3.3 Datasets

We use the publicly available datasets summarized in Table 1. These datasets contain detailed information about odorants and their constituent molecules, along with corresponding behavioral ratings from human

participants. Each dataset was collected using one of the psychophysical approaches described in Section 3.1. For all analyses, we use the average response across repetitions and participants for each odorant.

For the `Keller` (Keller & Vosshall, 2016), `Sagar` (Sagar et al., 2023), and `Leffingwell` (Sanchez-Lengeling et al., 2019) datasets, this yields a behavioral rating matrix $\mathbf{R}^{\text{behavioral}} \in \mathbb{R}^{N \times P}$, where N denotes the number of odorants and P the number of perceptual descriptors in that dataset. For `Keller` dataset, we included only the odorants that were rated by all participants and when multiple concentrations were available, we selected the behavioral ratings corresponding to the concentration of 0.001%. Missing (`NaN`) values were replaced with zeros, and the final behavioral rating for each odorant was obtained by averaging across all participants. For the `Sagar` dataset, we included only odorants that consisted of single molecules (mono-molecular stimuli). Additionally, we retained only the odorants and descriptors that were common across all participants. For `leffingwell` dataset, we used it as provided in (Sanchez-Lengeling et al., 2019), without additional curation or filtering. All odorant and descriptor associations were retained in their original binary form.

For the `Snitz` (Snitz et al., 2013) dataset, which are based on pairwise similarity judgments, we construct a sparse similarity matrix $\mathbf{S}^{\text{behavioral}} \in \mathbb{R}^{N \times N}$ where each entry S_{ij} represents the perceived similarity between odorants i and j for the pairs that were evaluated by participants. For this dataset, we included only pairs of odorants in which both stimuli were mono-molecular. Not all odorant pairs have associated similarity ratings, resulting in partially observed matrices. All datasets used in this study are publicly available and can be accessed either through their corresponding GitHub repositories or via the Pyrfume library (Hamel et al., 2024).

4 Methodolgy

Our objective is to evaluate the alignment between human and LLMs' olfactory responses when perceiving odorants, and to compare this alignment with representations derived from pre-trained molecular models of chemical structure. In this section, we introduce the models used in this study and how we measure the alignment according to the data provided. We access GPT and Gemini via their APIs, and OLMo via HuggingFace. All other models used in our work are linear, which makes both their training and inference computationally lightweight.

All the datasets used in this study are publicly available through their corresponding GitHub repositories or via the Pyrfume library (Hamel et al., 2024). GPT and Gemini are accessed through their API and OLMo is accessed through the Hugging Face repository. The code and instructions to reproduce all experiments and results are available at `https://anonymous.4open.science/r/Mol2Smell-FE70/`.

4.1 Generating Model Responses

Model responses can be obtained through two primary approaches: (i) prompting large language models (LLMs) to generate textual responses, and (ii) extracting internal embeddings by providing the models with structured prompts or molecular inputs.

For proprietary LLMs such as Gemini and GPT, we obtained model outputs directly as textual responses generated from the prompts described above. In contrast, for OLMo, which is open-source and provides access to internal representations, we extracted hidden embeddings instead of relying on generated text.

For transformer-based molecular models, we provided each model with string-based representations of individual molecules in the form of SMILES. From the resulting hidden states, we extracted the classification token (`[CLS]`) from the final transformer layer as the molecular representation. For each model $m \in M$, this procedure yields a representation matrix $\mathbf{R}_m^{\text{transformer}} \in \mathbb{R}^{N \times D_m}$, where N denotes the number of odorants and D_m is the embedding dimensionality of model m.

4.2 Models

LLMs We employ three state-of-the-art LLMs including Gemini-2.5-flash[1], GPT-5-mini[2], and OLMoE-1B-7B-0125-Instruct[3] to generate perceptual ratings corresponding to those provided by humans. Depending on the dataset type (absolute ratings or similarity comparisons, see Section 3.3), we designed two types of prompt templates, which were filled with the molecular representations, the target rating range, and the relevant perceptual descriptors (e.g., intensity, pleasantness, fruity). Descriptor sets and rating ranges are derived from the original human datasets, with a single exception: in the `Leffingwell` dataset the human labels are binary, but we prompt the LLMs to return probabilities in the range [0,1] for each descriptor. Our motivation for this choice was to allow greater flexibility in designing evaluation metrics that are more meaningful, especially since converting probabilities to binary values is straightforward, whereas the reverse is not. For each dataset, 3 independent repetitions were generated by each LLM for every human rating, and the results were aggregated by computing their mean. These outputs were then represented as numerical matrices, where for each LLM $l \in L$, we obtain a matrix $\mathbf{R}_l^{\text{LLM}} \in \mathbb{R}^{N \times P_d}$, with N denoting the number of odorants and P_d the number of perceptual descriptors for dataset d. This matrix subsequently serves as the basis for downstream aggregation and alignment analyses with human perceptual ratings. Regarding the temperature for LLMs, GPT-5 models do not support temperature settings and use a default hard-coded temperature. We set the temperature equal to 1 for Gemini.

Transformers We use three state-of-the-art encoder-only transformer models (as summarized in Table 2) pre-trained on large-scale chemical datasets, to extract computational representations of odorants. These models, publicly available via HuggingFace[4], are trained using self-supervision and have demonstrated strong performance on tasks such as property and reaction prediction.

4.3 Evaluation

Alignment Evaluation for Generated Responses When the LLMs were prompted to produce numerical ratings, their alignment with human perceptual responses can be directly assessed. For datasets containing continuous ratings or similarity judgments between pairs of odorants, we measure alignment using the *Pearson correlation coefficient (r)* between LLM-derived and human-derived ratings. Otherwise, for the `Lefingwell` dataset, where human responses are provided as binary labels for predefined descriptors, we assess alignment using the *AUC-ROC* score.

Alignment Evaluation for Embeddings Since it is not possible to directly measure alignment between embeddings and human ratings, linear models (Yamins & DiCarlo, 2016) are widely employed in the machine learning community (d'Ascoli et al., 2025; Zhang et al., 2025; Toneva & Wehbe, 2019) to assess this alignment. These models minimize the influence of additional model complexity, enabling a more direct and interpretable evaluation of representational correspondence. Following this standard approach, when the absolute continuous or binary ratings for odorants are provided, we first train linear models to predict the continuous or binary multi-target ratings assigned to odorants. We then use the trained models to predict held-out test targets, aggregate the predictions across the test sets of all 10 folds. Finally, we compute the *Pearson correlation coefficient (r)* for datasets containing continuous ratings and the *AUC-ROC* score for datasets with binary ratings, comparing the targets predicted from embeddings with human ratings, matching the same evaluation procedure used for the LLMs. For the similarity datasets, instead of training linear models, we compute cosine similarity between pairs of extracted representations and then evaluate alignment by correlating these similarities with those provided by humans to obtain the final *Pearson correlation coefficient (r)*.

Unified Alignment Score. Throughout this paper, we use the term **alignment score** to denote both the AUC-ROC and Pearson correlation coefficient metrics, as each quantifies the correspondence between model

[1] `https://deepmind.google/models/gemini/flash/`
[2] `https://platform.openai.com/docs/models/gpt-5-mini`
[3] `https://huggingface.co/allenai/OLMo-7B`
[4] `https://huggingface.co/models`

Dataset	Metric	Transformers			LLMs		
		MoLFormer	MTL-BERT	ChemBERTa	GPT	Gemini	OLMo
Keller	r	0.19 ± 0.03	0.20 ± 0.03	0.16 ± 0.03	0.35 ± 0.04	$\mathbf{0.39 \pm 0.04}$	0.21 ± 0.03
Sagar	r	0.21 ± 0.04	0.20 ± 0.04	0.18 ± 0.04	0.27 ± 0.05	$\mathbf{0.32 \pm 0.04}$	0.27 ± 0.03
Snitz	r	0.33 ± 0.001	0.60 ± 1.66	0.11 ± 0.31	0.72 ± 1.74	$\mathbf{0.76 \pm 9.91}$	0.14 ± 0.00
Leffingwell	ROC-AUC	$\mathbf{0.84 \pm 0.01}$	$\mathbf{0.84 \pm 0.01}$	0.79 ± 0.01	0.75 ± 0.01	0.76 ± 0.01	$\mathbf{0.84 \pm 0.01}$

Table 3: **Alignment scores across models and datasets.** LLMs outperform in similarity-based tasks and in predicting continuous ratings over a limited set of descriptors. In contrast, specialized molecular transformers achieve higher performance when richer chemical information is available. For details of alignment score, see Section 4.3. Higher values indicate better alignment.

Dataset	GPT		Gemini		OLMo	
	SMILES	Name	SMILES	Name	SMILES	Name
Keller	0.35 ± 0.04	0.38 ± 0.04	0.39 ± 0.04	0.39 ± 0.03	0.21 ± 0.03	0.2 ± 0.02
Sagar	0.27 ± 0.05	0.32 ± 0.04	0.32 ± 0.04	0.33 ± 0.03	0.27 ± 0.03	0.21 ± 0.03
Snitz	0.71 ± 1.74	0.71 ± 3.99	0.76 ± 9.91	0.73 ± 2.89	0.14 ± 0.00	0.44 ± 0.00
Leffingwell	0.75 ± 0.01	0.77 ± 0.01	0.76 ± 0.01	0.76 ± 0.01	0.84 ± 0.01	0.81 ± 0.01

Table 4: **Alignment scores across LLMs and datasets for two different input types.** The difference between the two input formats is more pronounced for *GPT*, whereas for *Gemini*, the standard error decreases when using the name-based input format.

predictions and human responses, tailored to the respective output type. When interpreting the results, it should be noted that the chance levels of the two metrics are not the same. AUC-ROC has an expected chance value of 0.5, whereas Pearson correlation has a chance value of 0.

5 Results

In this section, we explore the alignment between LLMs and humans through three research questions:
RQ1: How well do LLM-generated perceptual ratings align with human ratings, compared to molecular transformers across multiple datasets and output representation types?
RQ2: Does the type of input representation, chemical structure (SMILES) versus common molecular name, affect the perceptual outputs of LLMs?
RQ3: Where do the observed alignment differences originate within LLMs across different input types?

Alignment between LLMs and Humans. Table 3 summarizes the alignment results across all datasets and models. As shown, *GPT* and *Gemini* consistently outperform the molecular transformers across three datasets `Sagar`, `Keller`, and `Snitz`, where continuous or similarity rating is provided. This result is interesting given that LLMs have never been fine-tuned to perform such a task, yet their inferred perceptual ratings show a substantially higher correspondence with human olfactory judgments. The OLMo model shows performance comparable to that of the transformer models on these three datasets. One possible explanation is that OLMo is a smaller model, which may limit its capacity to capture complex perceptual patterns. However, a direct comparison is difficult because the other two models are not open-source, preventing full transparency regarding their architecture and training details.

For the `Leffingwell` dataset, the performance of GPT and Gemini, while comparable to that of the transformer models, is slightly lower. However, for OLMo models, it is similar to what is reported among the transformer models. One common aspect between this model and the other two transformers is that for all of them, we extracted embeddings instead of direct responses, which is the opposite of what we did for Gemini and GPT. The `Leffingwell` dataset also includes 113 perceptual descriptors (i.e., classes), approx-

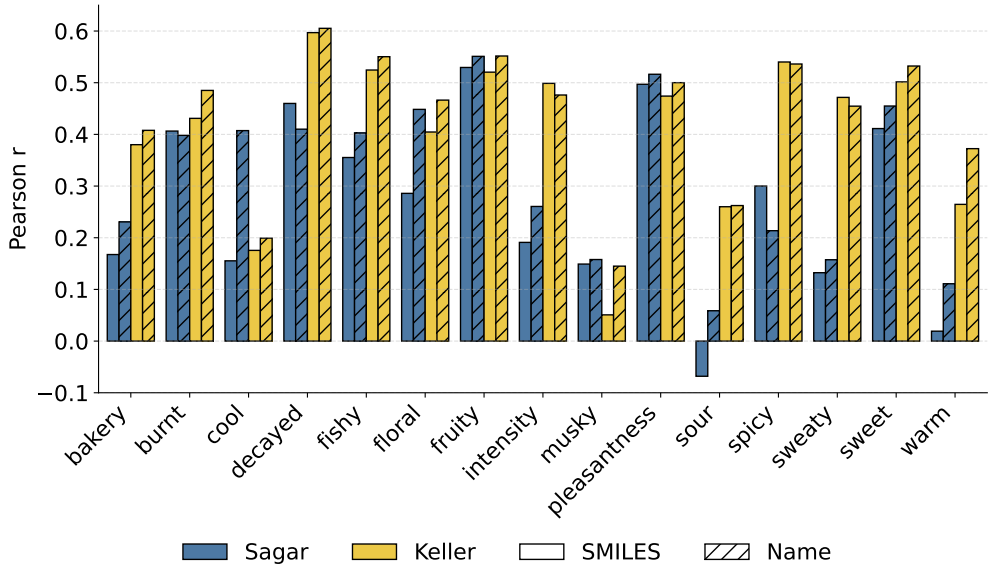

Figure 2: **Descriptor-level alignment patterns between human olfactory ratings and GPT-5-mini predictions across two input types**: molecular names and SMILES strings . Name-based inputs yield higher alignment for positive and concrete descriptors (e.g., fruity, floral, bakery, sweet, pleasantness, warm).

imately five times more than the other datasets, resulting in a much more fine-grained perceptual space. Such granularity may exceed the level of detail that LLMs can reliably infer from textual representations. In addition, the human ratings in this dataset are binary, which complicates the evaluation process and may reduce the sensitivity of the metrics to subtle perceptual differences. Therefore, one might conclude that for such tasks, it may be better to extract embeddings rather than rely on direct responses.

Together, these results suggest that LLMs encode a rich, linguistically grounded understanding of olfactory perception that generalizes across diverse datasets and rating schemes. They capture meaningful perceptual regularities that align with human judgments, highlighting the potential of language-based models as complementary tools to chemically grounded approaches.

Impact of Input Representation: Common Name vs SMILES. Table 4 compares the performance of LLMs in predicting human olfactory responses based on two input formats: common molecular names and SMILES strings. Across GPT and Gemini, prompts using odorant names consistently achieve comparable or higher alignment with human perceptual ratings than those using SMILES strings, particularly for datasets with continuous or binary rating formats (i.e., `Leffingwell`, `Sagar`, and `Keller`). This pattern indicates that LLMs can effectively leverage linguistic knowledge embedded in chemical or common names, which often implicitly encode information about chemical families (e.g., “aldehyde,” “musk,” “citrus”) or sensory associations. Interestingly, this suggests that natural language alone can provide an amount of perceptually relevant information comparable to that extracted from explicit molecular structures by specialized chemical transformers. Whereas structure-based models learn such relationships through atom-level encodings, LLMs appear to infer them from the co-occurrence statistics of names and descriptive contexts in large text corpora. Thus, name-based prompts reveal an emergent form of semantic grounding in LLMs that bridges linguistic and perceptual knowledge domains. For the OLMo model, in the `Snitz` dataset, the model performs substantially better when the molecule name is provided compared to when SMILES is used as input. In contrast, for the other datasets, the model slightly underperforms when the name is provided compared to SMILES.

Descriptor-Level Alignment Patterns Across Input Types. We conducted a final analysis to examine whether the alignment between LLMs and human judgments follows specific patterns at the level of individual perceptual descriptors. Figure 2 compares the alignment between human olfactory ratings and

the predictions of GPT-5-mini across a set of shared perceptual descriptors between two datasets, using two molecular input formats: common molecular names and SMILES strings. We specifically selected GPT-5 for this analysis because, as shown in Table 4, the performance gap between these two input formats is particularly pronounced for this model. This comparison allows us to examine how linguistic versus structural representations influence the model's alignment with human olfactory perception.

The analysis reveals that for positive and concrete descriptors, those easily linked to familiar odor sources, such as fruity, floral, bakery, sweet, pleasantness, and warm, using the name-based input leads to a clear improvement in performance. In contrast, for negative or more abstract descriptors such as decayed, sour, sweaty, or burnt, the difference between the two input types is smaller and more consistent across input types. This asymmetry may reflect the fact that name-based representations carry richer semantic cues for positive and familiar smells, whereas negative or ambiguous odors are less frequently and less distinctly represented in everyday language.

6 Discussion

The complexity of olfactory perception, from the perspectives of psychology, chemistry, neuroscience, and language, is undeniable. Despite decades of research, there is still no well-grounded, digitized framework for describing odorants, nor a fully developed linguistic system that captures the richness of olfactory experience. Humans, however, have found ways to communicate odors: sometimes by relating them to familiar sources (e.g., "rose-like," "burnt," "citrus"), and at other times through abstract perceptual dimensions that have emerged through language, such as pleasantness, intensity, or familiarity. In either case, the space of olfactory description remains deeply complex, and intertwined with the sciences of chemistry, cognition, and language.

In this work, we present a systematic analysis aimed at understanding the interaction between language, chemistry, and odor perception. To do so, we leveraged human perceptual datasets that describe how specific monomolecular odorants are experienced. We then prompted LLMs to provide similar perceptual judgments and compared their predictions with human data. Importantly, our analysis was not limited to LLMs trained exclusively on linguistic data, we also evaluated specialized molecular transformers trained directly on chemical structures, in order to examine whether the perceptual reasoning capabilities of LLMs extend beyond their linguistic training.

The results show that LLMs trained solely on natural language often outperformed chemically specialized transformers in predicting how humans perceive the smell of molecules. This finding suggests that a considerable amount of perceptual knowledge about odorants is implicitly encoded within human language. Moreover, we observed systematic differences in performance depending on the form of input, when odorants were presented through chemical structures (e.g., SMILES) versus through their common names, revealing how linguistic and structural representations capture distinct but complementary aspects of olfactory meaning.

Together, these results point to a promising direction toward bridging the long-standing structure–odor gap: leveraging the rich semantic priors embedded in human language to model, interpret, and ultimately unify chemical and perceptual representations of smell.

Future Work. Future research should aim to address the current limitations of available olfactory datasets by generating larger perceptual datasets to include a wider range of odorants, containing both single molecules and complex mixtures. Incorporating mixtures is essential, as real-world olfactory experiences often emerge from interactions among multiple compounds that produce emergent perceptual qualities. Beyond data, an important direction lies in developing multimodal modeling frameworks that integrate chemical, linguistic, and behavioral information. Each modality provides a unique and complementary perspective: chemical representations capture the physicochemical basis of odor, linguistic descriptions encode human semantic and cultural understanding of smells, and behavioral responses reflect perceptual and emotional aspects of olfaction. Leveraging these modalities jointly could enable cross-domain supervision, where, for instance, linguistic embeddings guide molecular models toward human-relevant features, and chemical structure constrains language models to remain physically grounded. Such multimodal alignment holds promise for

building unified models that not only predict human olfactory perception more accurately but also bridge the gap between symbolic, sensory, and chemical representations of smell.

7 Limitations

Our work is constrained by the limited availability of perceptual olfactory data. Only a few datasets currently provide human olfactory ratings, and these datasets contain a relatively narrow range of odorants. Furthermore, we restricted our analysis to odorants composed of single, well-defined molecular species (monomolecules), excluding odorant mixtures that consist of multiple compounds combined to produce complex scents (e.g., perfumes or natural extracts). This choice ensures that each stimulus has a clear chemical representation that can be directly mapped to a molecular structure, allowing for a fair and interpretable comparison between LLMs and molecular transformer models. Finally, to assess the intrinsic capability of LLMs and ensure fairness in the information provided to both model types, we intentionally avoided using any specialized prompting strategies.

References

Sanne Boesveldt, Mats J Olsson, and Johan N Lundström. Carbon chain length and the stimulus problem in olfaction. *Behavioural brain research*, 215(1):110–113, 2010.

Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. Chemberta: large-scale self-supervised pretraining for molecular property prediction. *arXiv preprint arXiv:2010.09885*, 2020.

Stéphane d'Ascoli, Jérémy Rapin, Yohann Benchetrit, Hubert Banville, and Jean-Rémi King. Tribe: Trimodal brain encoder for whole-brain fmri response prediction. *arXiv preprint arXiv:2507.22229*, 2025.

Vishal Dey, Xiao Hu, and Xia Ning. GeLLM^{3}O: Generalizing large language models for multi-property molecule optimization. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 25192–25221, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1225. URL https://aclanthology.org/2025.acl-long.1225/.

Elizabeth A Hamel, Jason B Castro, Travis J Gould, Robert Pellegrino, Zhiwei Liang, Liyah A Coleman, Famesh Patel, Derek S Wallace, Tanushri Bhatnagar, Joel D Mainland, et al. Pyrfume: A window to the world's olfactory data. *Scientific Data*, 11(1):1220, 2024.

Rebecca M. M. Hicke, Sil Hamilton, and David Mimno. The zero body problem: Probing LLM use of sensory language. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=vv1ZyQF8LD.

Andreas Keller and Leslie B Vosshall. Olfactory perception of chemically diverse molecules. *BMC neuroscience*, 17:1–17, 2016.

Christopher Lalk, Kim Targan, Tobias Steinbrenner, Jana Schaffrath, Steffen Eberhardt, Brian Schwartz, Antonia Vehlen, Wolfgang Lutz, and Julian Rubel. Employing large language models for emotion detection in psychotherapy transcripts. *Frontiers in Psychiatry*, 16:1504306, 2025.

Brian K Lee, Emily J Mayhew, Benjamin Sanchez-Lengeling, Jennifer N Wei, Wesley W Qian, Kelsie A Little, Matthew Andres, Britney B Nguyen, Theresa Moloy, Jacob Yasonik, et al. A principal odor map unifies diverse tasks in olfactory perception. *Science*, 381(6661):999–1006, 2023.

Shengchao Liu, Jiongxiao Wang, Yijin Yang, Chengpeng Wang, Ling Liu, Hongyu Guo, and Chaowei Xiao. Conversational drug editing using retrieval and domain feedback. In *The twelfth international conference on learning representations*, 2024.

Jörn Lötsch, Dario Kringel, and Thomas Hummel. Machine learning in human olfactory research. *Chemical senses*, 44(1):11–22, 2019.

Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D White, and Philippe Schwaller. Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, 6(5):525–535, 2024.

Raja Marjieh, Ilia Sucholutsky, Pol van Rijn, Nori Jacoby, and Thomas L Griffiths. Large language models predict human sensory judgments across six modalities. *Scientific Reports*, 14(1):21445, 2024.

Aharon Ravia, Kobi Snitz, Danielle Honigstein, Maya Finkel, Rotem Zirler, Ofer Perl, Lavi Secundo, Christophe Laudamiel, David Harel, and Noam Sobel. A measure of smell enables the creation of olfactory metamers. *Nature*, 588(7836):118–123, 2020.

Jerret Ross, Brian Belgodere, Vijil Chenthamarakshan, Inkit Padhi, Youssef Mroueh, and Payel Das. Large-scale chemical language representations capture molecular structure and properties. *Nature Machine Intelligence*, 4(12):1256–1264, 2022.

Vivek Sagar, Laura K Shanahan, Christina M Zelano, Jay A Gottfried, and Thorsten Kahnt. High-precision mapping reveals the structure of odor coding in the human brain. *Nature neuroscience*, pp. 1–8, 2023.

Kushagra Saini and Venkatnarayan Ramanathan. Predicting odor from molecular structure: a multi-label classification approach. *Scientific reports*, 12(1):13863, 2022.

Benjamin Sanchez-Lengeling, Jennifer N Wei, Brian K Lee, Richard C Gerkin, Alán Aspuru-Guzik, and Alexander B Wiltschko. Machine learning for scent: Learning generalizable perceptual representations of small molecules. *arXiv preprint arXiv:1910.10685*, 2019.

Charles S Sell. On the unpredictability of odor. *Angewandte Chemie International Edition*, 45(38):6254–6261, 2006.

Anju Sharma, Rajnish Kumar, Imlimaong Aier, Rahul Semwal, Pankaj Tyagi, and Pritish Varadwaj. Sense of smell: structural, functional, mechanistic advancements and challenges in human olfactory research. *Current neuropharmacology*, 17(9):891–911, 2019.

Daiki Shiono, Ana Brassard, Yukiko Ishizuki, and Jun Suzuki. Evaluating model alignment with human perception: A study on shitsukan in llms and lvlms. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 11428–11444, 2025.

Kai Siedenburg and Charalampos Saitis. The language of sounds unheard: Exploring musical timbre semantics of large language models. *arXiv preprint arXiv:2304.07830*, 2023.

Kobi Snitz, Adi Yablonka, Tali Weiss, Idan Frumin, Rehan M Khan, and Noam Sobel. Predicting odor perceptual similarity from odor structure. *PLoS computational biology*, 9(9):e1003184, 2013.

Farzaneh Taleb, Miguel Vasco, Antonio Ribeiro, Mårten Björkman, and Danica Kragic. Can transformers smell like humans? *Advances in Neural Information Processing Systems*, 37:72032–72060, 2024.

Mariya Toneva and Leila Wehbe. Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). *Advances in neural information processing systems*, 32, 2019.

Jiahang Tu, Hao Fu, Fengyu Yang, Hanbin Zhao, Chao Zhang, and Hui Qian. Texttoucher: Fine-grained text-to-touch generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 7455–7463, 2025.

David Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, 28(1):31–36, 1988.

Ziting Xian, Jiawei Gu, Lingbo Li, and Shangsong Liang. Molrag: unlocking the power of large language models for molecular property prediction. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15513–15531, 2025.

Qihui Xu, Yingying Peng, Samuel A Nastase, Martin Chodorow, Minghua Wu, and Ping Li. Large language models without grounding recover non-sensorimotor but not sensorimotor features of human concepts. *Nature human behaviour*, pp. 1–16, 2025.

Daniel LK Yamins and James J DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, 19(3):356–365, 2016.

Chenyu Zhang, Benjamin Van Durme, Zhuowan Li, and Elias Stengel-Eskin. Visual commonsense in pretrained unimodal and multimodal models. *arXiv preprint arXiv:2205.01850*, 2022a.

Mingfang Zhang, Jarod Lévy, Stéphane d'Ascoli, Jérémy Rapin, F Alario, Pierre Bourdillon, Svetlana Pinet, Jean-Rémi King, et al. From thought to action: How a hierarchy of neural dynamics supports language production. *arXiv preprint arXiv:2502.07429*, 2025.

Xiao-Chen Zhang, Cheng-Kun Wu, Jia-Cai Yi, Xiang-Xiang Zeng, Can-Qun Yang, Ai-Ping Lu, Ting-Jun Hou, and Dong-Sheng Cao. Pushing the boundaries of molecular property prediction for drug discovery with multitask learning bert enhanced by smiles enumeration. *Research*, 2022:0004, 2022b.

Shu Zhong, Zetao Zhou, Christopher Dawes, Giada Brianz, and Marianna Obrist. Sniff ai: Is my'spicy'your'spicy'? exploring llm's perceptual alignment with human smell experiences. *arXiv preprint arXiv:2411.06950*, 2024.

Xiang Zhuang, Bin Wu, Jiyu Cui, Kehua Feng, Xiaotong Li, Huabin Xing, Keyan Ding, Qiang Zhang, and Huajun Chen. Boosting llm's molecular structure elucidation with knowledge enhanced tree search reasoning. *arXiv preprint arXiv:2506.23056*, 2025.

Dataset	Metric	GPT		Gemini	
		SMILES	Name	SMILES	Name
Keller	r	0.36 ± 0.03	$\mathbf{0.40 \pm 0.03}$	0.35 ± 0.03	0.40 ± 0.03
Sagar	r	0.25 ± 0.04	$\mathbf{0.32 \pm 0.04}$	0.30 ± 0.04	0.36 ± 0.03
Snitz	r	0.70 ± 0.00	$\mathbf{0.73 \pm 0.00}$	0.71 ± 0.00	0.70 ± 1.71

Table 5: **Alignment scores across LLMs and datasets for two different input types for the second template.** The difference between the two input formats is more pronounced for *Gemini*, whereas for *GPT*, the standard error decreases when using the name-based input format. Temperature for Gemini=1

A Prompt Templates and Ablations

To obtain perceptual ratings from LLMs, we designed structured text prompts that instruct the model to act as an olfactory rater.

Each prompt specifies the *molecule identity*, provided either as a common name or as a SMILES string, a list of *perceptual descriptors* to be rated (e.g., *sweet*, *woody*, *fishy*), and a predefined *rating range*. For similarity judgments, the prompt instead instructs the model to provide a numerical similarity score between two odorants.

The model is explicitly instructed to return only a valid JSON object containing numerical ratings for each descriptor, without any additional text or explanation. This structured output format ensures reliable parsing and consistent downstream analysis.

To evaluate the robustness of model responses to prompt framing, we conducted an ablation study using an alternative prompt template. In this variant, instead of instructing the model to *act as an olfactory rater*, the prompt asks the model to *imagine smelling the odorants*. This formulation more closely mirrors the instructions given to human subjects in psychophysical experiments.

Below, we provide both prompt templates for descriptor ratings and similarity judgments. The results obtained using the alternative template are reported in Table 5. All main results presented in the paper are based on the primary template.

A.1 Prompt Templates

We designed two prompt templates to obtain perceptual judgments from LLMs. Each template was evaluated on two tasks: i) descriptor ratings and (ii) similarity ratings.

A.1.1 Template 1: Olfactory Rater Instruction

Case 1: Descriptor Ratings

```
System message:
You are an olfactory rater. Output ONLY valid JSON.

User message:
Molecule:
- Name: vanillin

Descriptors (rate each from 0 to 100): ["sweet", "woody", "floral"]

Output rules:
- Return ONLY a single valid JSON object. No prose, no markdown.
- Keys must match the descriptor list exactly.
- Values must be numbers in [0,100].
```

Case 2: Similarity Ratings

```
System message:
You are an olfactory rater. Output ONLY valid JSON.

User message:
Molecule A:
- Name: vanillin

Molecule B:
- Name: ethyl vanillin

Task:
Rate the perceptual similarity between Molecule A and Molecule B.

Similarity scale:
- 0 = completely different odor
- 100 = identical odor

Output rules:
- Return ONLY a single valid JSON object. No prose, no markdown.
- Use exactly this key: "similarity"
- Value must be a number in [0,100].
```

A.1.2 Template 2: Human Smelling Instruction (Ablation)

Case 1: Descriptor Ratings

```
System message:
Imagine you are smelling an odorant as a human participant in an olfactory experiment.
Output ONLY valid JSON.

User message:
Molecule:
- Name: vanillin

Descriptors (rate each from 0 to 100): ["sweet", "woody", "floral"]

Instructions:
Provide ratings as a human would based on perceived smell.

Output rules:
- Return ONLY a single valid JSON object. No prose, no markdown.
- Keys must match the descriptor list exactly.
- Values must be numbers in [0,100].
```

Case 2: Similarity Ratings

```
System message:
Imagine you are smelling two odorants as a human participant in an olfactory experiment.
Output ONLY valid JSON.

User message:
Molecule A:
- Name: vanillin
```

```
Molecule B:
- Name: ethyl vanillin

Task:
Smell both odorants and rate their perceptual similarity as a human would.

Similarity scale:
- 0 = completely different odor
- 100 = identical odor

Output rules:
- Return ONLY a single valid JSON object. No prose, no markdown.
- Use exactly this key: "similarity"
- Value must be a number in [0,100].
```

Expected model output:

```
{
  "sweet": 90,
  "woody": 45,
  "floral": 72
}
```