BENCHMARK FOR ASSESSING OLFACTORY PERCEPTION OF LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Here we introduce the Olfactory Perception (OP) benchmark, designed to assess the capability of large language models (LLMs) to reason about smell. The benchmark contains 1,010 questions across eight task categories spanning odor classification, odor primary descriptor identification, intensity and pleasantness judgments, multi-descriptor prediction, mixture similarity, olfactory receptor activation, and smell identification from real-world odor sources. Each question is presented in two prompt formats, compound names and isomeric SMILES, to evaluate the effect of molecular representations. Evaluating 21 model configurations across major model families, we find that compound-name prompts consistently outperform isomeric SMILES, with gains ranging from +2.4 to +18.9 percentage points (mean $\approx$ +7 points), suggesting current LLMs access olfactory knowledge primarily through lexical associations rather than structural molecular reasoning. The best-performing model reaches 64.4% overall accuracy, which highlights both emerging capabilities and substantial remaining gaps in olfactory reasoning. We further evaluate a subset of the OP across 21 languages and find that aggregating predictions across languages improves olfactory prediction, with AUROC = 0.86 for the best performing language ensemble model. LLMs should be able to handle olfactory and not just visual or aural information.

1 INTRODUCTION

The sense of smell occupies a unique position among human perceptual modalities (Mainland et al., 2015; Secundo et al., 2014). Unlike vision, where wavelengths map predictably to colors (Stockman & Sharpe, 2000), or audition, where frequencies correspond to pitch (Oxenham et al., 2011), olfaction operates through a complex interplay between molecular structure and receptor biology that remains incompletely understood (Mainland et al., 2015; Buck & Axel, 1991; Bushdid et al., 2014). Humans can distinguish over a trillion distinct odors Bushdid et al. (2014); err (2024), though estimates are still debated (Gerkin & Castro, 2015), yet predicting how a molecule will smell from its chemical structure alone has long eluded computational approaches (Keller et al., 2017). This structure-odor relationship represents one of the most challenging frontiers in sensory science Keller et al. (2017); Lee et al. (2023); Secundo et al. (2014).

Large language models have demonstrated remarkable capabilities across diverse domains, from mathematical reasoning to code generation Brown et al. (2020); Chen et al. (2021). Recent work has shown that these models, particularly those trained with explicit reasoning objectives (Han et al., 2024), can perform sophisticated tasks in chemistry, interpreting SMILES (Simplified Molecular Input Line Entry System) (Weininger, 1988) strings, predicting reaction products, and even elucidating molecular structures from NMR spectra Runcie et al. (2025); Mirza et al. (2025). These advances raise a natural question: can LLMs bridge the gap between molecular structure and sensory perception? Specifically, do these models possess knowledge about how molecules smell?

This question carries both scientific and practical significance. LLMs capable of olfactory reasoning could accelerate fragrance design Mao et al. (2018), flavor development Ge et al. (2025), and the identification of malodorous contaminants Bartsch et al. (2016). Yet despite growing interest in evaluating LLM sensory alignment across modalities including color, auditory perception, and taste Marjieh et al. (2024), olfaction has remained notably absent from systematic evaluation. Prior work has found limited human-AI alignment for smell Zhong et al. (2024), shown that LLMs capture

olfactory-semantic similarity but not factual accuracy Kurfalı et al. (2025), and demonstrated that specialized graph neural networks can achieve human-level odor prediction Keller et al. (2017); Lee et al. (2023). However, no existing work systematically evaluates whether general-purpose LLMs possess factual knowledge about olfactory properties through structured question-answering.

We address this gap by introducing the Olfactory Perception (OP) benchmark, comprising 1,010 questions across eight task categories that span the full complexity of olfactory perception: from basic odor detection Mayhew et al. (2022) and classification to intensity and pleasantness judgments Keller et al. (2017), from single-molecule descriptor identification International Fragrance Association (2020) to mixture similarity Bushdid et al. (2014); Snitz et al. (2013); Ravia et al. (2020); Satarifard et al. (2025), and from semantic odor labeling Lee et al. (2023) to receptor activation Lalis et al. (2024). Ground-truth answers are derived from established datasets in olfactory science. Motivated by cross-linguistic diversity in odor language Majid & Burenhult (2014); Majid et al. (2018); Majid (2021), we additionally evaluate a subset across 21 languages and report a cross-lingual majority-vote setting.

A distinctive feature of our approach is the dual-prompting strategy: each question is presented using both isomeric SMILES notation and common compound names. Because odor perception is stereospecific (e.g., (R)-carvone smells of spearmint while (S)-carvone smells of caraway), this design enables direct comparison of how molecular representation format affects olfactory reasoning Runcie et al. (2025), assessing whether models reason about molecular properties or merely retrieve lexical associations. Our evaluation encompasses 21 model configurations across six providers, including OpenAI's GPT and o3 series, Google's Gemini, Anthropic's Claude, Meta's Llama, xAI's Grok, and DeepSeek. We systematically vary reasoning budgets where applicable, enabling analysis of how extended deliberation affects olfactory task performance, an approach that has revealed substantial performance differences in chemistry tasks Runcie et al. (2025); Mirza et al. (2025). Our findings reveal that while current models achieve moderate success on certain tasks, substantial gaps remain, particularly for questions requiring genuine molecular reasoning rather than factual recall about well-known compounds.

The contributions of this work are threefold. First, we introduce a comprehensive benchmark for evaluating LLMs on olfactory reasoning, comprising 1,010 questions across eight task categories grounded in peer-reviewed olfactory science. Second, we establish baseline performance across state-of-the-art models, identifying both emerging capabilities and systematic failures. Third, via dual-prompting (compound names vs. isomeric SMILES), we provide new insight into how molecular representation shapes model performance, helping to distinguish structural reasoning from lexical association.

2 Related Work

Recent benchmarks have evaluated LLM chemical reasoning, including ChemBench Mirza et al. (2025), ChemIQ Runcie et al. (2025), LlaSMol Yu et al. (2024), GPQA Rein et al. (2024), MMLU Hendrycks et al. (2021), and a comprehensive eight-task evaluation Guo et al. (2023), but none test whether LLMs can reason about how molecules smell. More broadly, LLMs show substantial but incomplete alignment with human sensory judgments across modalities such as pitch, color, and taste Marjieh et al. (2024); Xu et al. (2025); Fukushima et al. (2025), with performance degrading for modalities less frequently described in language. Olfaction, notably absent from these evaluations, represents a natural test case.

Initial work on LLMs and olfaction has found limited human-AI alignment: SNIFF AI Zhong et al. (2024) achieved only 27.5% success in scent identification from descriptions, and Kurfalı et al. (2025) showed that GPT-4o captures olfactory-semantic similarity but evaluated similarity judgments rather than factual accuracy. Related multimodal work has explored image-text olfactory matching Kurfalı et al. (2023); Esteban-Romero et al. (2025). Meanwhile, specialized ML systems purpose-built for olfaction have achieved strong performance: the Principal Odor Map (POM) Lee et al. (2023) reaches AUROC = 0.89 for multi-label descriptor prediction on the GS-LF dataset (the same 138-descriptor vocabulary used in our RATA task), Dense Sense Saha et al. (2025) reaches AUROC = 0.875, Mol-PECO Zhang et al. (2023) achieves AUROC of 0.813, and POMMIX Tom et al. (2025) extends the POM to mixture similarity (Pearson $\rho = 0.779$). Two DREAM Challenges Keller et al. (2017);

Satarifard et al. (2025) provided foundational benchmarks for these systems. However, none of this work addresses whether general-purpose LLMs possess latent olfactory knowledge.

Our OP benchmark fills this gap as, to our knowledge, the first structured factual evaluation of LLM olfactory reasoning, comprising 1,010 questions across eight task categories with ground-truth answers derived from established olfactory science datasets (Mayhew et al., 2022; Keller et al., 2017; Lee et al., 2023; Lalis et al., 2024; Snitz et al., 2013; Bushdid et al., 2014; Ravia et al., 2020; International Fragrance Association, 2020; Kreissl et al., 2022). A dual-prompting strategy (isomeric SMILES vs. compound names) enables direct comparison of how molecular representation affects olfactory reasoning. A detailed discussion and side-by-side comparison with prior work is provided in Appendix B.

3 Olfactory Perception Benchmark

We introduce a unified benchmark of 1,010 olfaction questions spanning odor detectability, semantic odor description, perceptual judgments, mixture similarity, receptor activation, and smell identification test from mixtures. Figure 1 provides an overview of the benchmark; a detailed summary of question categories, data sources, and response formats is provided in Appendix A. Each item is presented in two equivalent formats: (i) isomeric SMILES (prompt 1), and (ii) compound names (prompt 2), to separate structure-based reasoning from name priors. Tasks are multiple choice from constrained lists of options to enable consistent automatic scoring across models.

Odor Classification (OC). Odor classification is a binary task that asks whether a given molecule is *Odorous* or *Odorless* (175 questions, 50% Odorous). This task assesses basic odor detectability prediction from chemical identity of a molecule. Molecules were obtained from a previously curated dataset Mayhew et al. (2022) and molecular weight was constrained to be ≤ 350.0 (so as to be in the smellable range); items were presented to LLMs in random order. We consider odor classification to be a simple task.

Odor Primary Descriptor (OPD). Odor primary descriptor is a multi-class task where the model selects the single descriptor for a molecule from a provided list of four options (175 questions). This task evaluates whether models can map chemicals to main semantic odor categories. The molecules were obtained from the 2020 version of the IFRA fragrance ingredient glossary (FIG) International Fragrance Association (2020). Besides the primary descriptor, FIG provides secondary and tertiary descriptors; we excluded these descriptors from the list of distractors to obtain a more robust evaluation. Selection of molecules was done on a stratified set of descriptors to represent less dominant descriptors, and a total of 29 descriptors are present among questions. We consider odor primary descriptor a simple task.

Odor Intensity and Pleasantness (OIn and OPl). To evaluate model assessment of two olfactory dimensions (*i.e.,* intensity and pleasantness), we use two paired-comparison tasks, where the model chooses which of two molecules is more intense or more pleasant (175 molecular pairs for intensity, and 175 molecular pairs for pleasantness comparisons). Ground truth data are obtained from prior work Keller et al. (2017) which provides mean human subjective ratings on a 0–100 scale. Molecular pairs were selected to be above and below median with a minimum score difference of 10. Furthermore, prompts ask the model to rate the intensity and pleasantness on a 0–100 scale. We consider odor intensity and pleasantness categories to be simple tasks.

Rate-All-That-Apply (RATA). To assess more complex olfactory perceptual capability, we employ a multilabel semantic profiling task (100 questions), where, given a molecule and a descriptor lexicon (138 odor descriptors), the model selects all descriptors that apply in describing the odor of the molecule. We selected 100 molecules from an integrated dataset of the Good Scents and Leffingwell Associates (GS-LF) Lee et al. (2023). We constrained our selection to molecules with 2 to 5 answers from 138 descriptors, with 25 questions for 2, 3, 4, and 5 descriptors, respectively. We consider RATA categorization as an intermediate-difficulty task.

Odor Similarity of Mixtures (OS). In odor similarity of mixtures, we evaluate a categorical mixture-perception task (100 questions) where the model compares two mixtures (each a set of 2–10 molecules) and predicts an ordinal similarity label (e.g., from strongly similar to strongly dissimilar). We selected 100 mixture pairs from various datasets Snitz et al. (2013); Bushdid et al. (2014); Ravia et al. (2020) and after standardization of perceptual distance, similarity values were categorized into

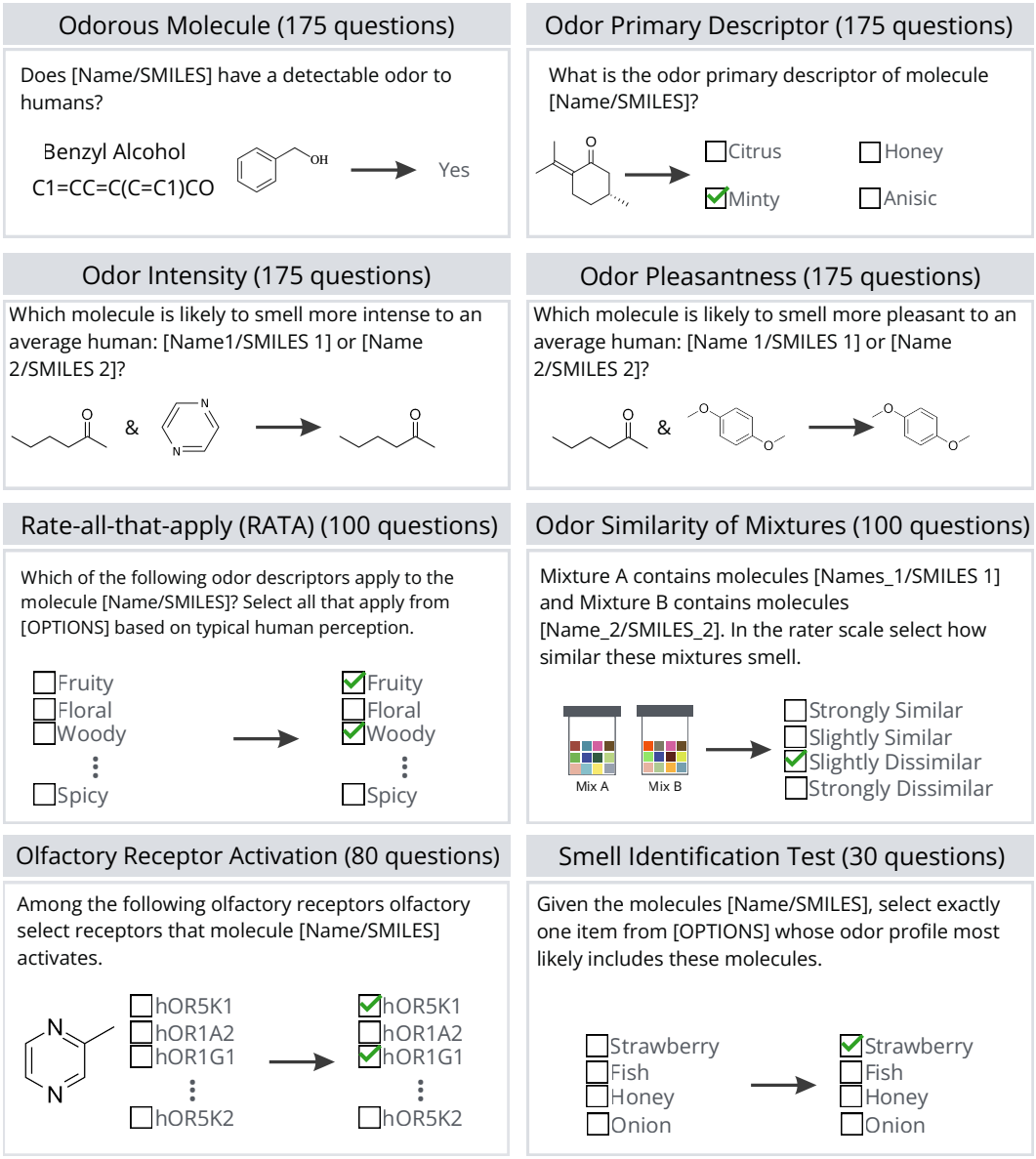

Figure 1: Olfactory Perception (OP) benchmark task description and output format across tasks.

four bins of a categorical rating scale, including: strongly similar, slightly similar, slightly dissimilar, and strongly dissimilar. This task deals with mixture-level olfactory perception, and we classify it as intermediate-difficulty.

Olfactory Receptor Activation (ORA). In this task, we aim to assess model capability in identifying olfactory receptor (OR) activation for pure molecules. We use a multilabel task (80 questions) that asks which receptors (from a candidate set of 4–10 human OR gene IDs) are activated by a given molecule. The molecules-OR pairs were obtained from the M2OR dataset Lalis et al. (2024). We constrained our selection of molecules-OR pairs to cases where 4 to 10 ORs activation were experimentally studied, and 1 to 10 ORs were observed to be activated. We consider ORA a hard task.

Smell Identification Test (SIT). We curated a multiple-choice smell identification task (30 questions) with molecular mixture information. The model selects the most likely odor source from a four-option

set given a molecular profile. We obtained the list of volatile organic compounds for different food items from the Leibniz-lsbtum odorant database Kreissl et al. (2022). The items studied include the odor profile for mango, peanut, hazelnut, tomato, apple, walnut, raspberry, peach, honey, parsley, grapefruit, pineapple, strawberry, apricot, rice, grape, popcorn, orange, cheese, melon, leather, chocolate, coffee, onion, fish, beer, whisky, red wine, prawn, and bread. We consider the SIT a hard task.

Multilingual Translation. We generated multilingual versions of the RATA tasks by translating the English odor-descriptor vocabulary and the UI/instruction text into each target language using `GPT-5.2`. We prompted the model to translate in a perfumery/sensory context and to produce brief, natural descriptors; we used high-reasoning settings for disambiguation. Descriptor translations were normalized to lowercase for consistency, while UI strings were translated separately to preserve natural phrasing. To reduce hallucinated or awkward terms, we applied a lightweight two-step quality control check for any uncertain cases: the model proposes candidate single-word descriptors and then selects the best option based on whether it is a real word in the target language and whether it fits olfactory usage. If no suitable translation is found, we fall back to the closest candidate; if generation fails entirely, we retain the original English descriptor. For languages where compounds are standard (e.g., German, Swedish), compound forms were allowed as long as they remained whitespace-free.

4 EXPERIMENTS

To comprehensively evaluate olfactory reasoning capabilities across the current LLM landscape, we select models spanning multiple providers, architectures, and reasoning configurations. A key dimension of our evaluation is the effect of reasoning budget on olfactory task performance, motivated by findings in chemistry benchmarks showing that extended reasoning substantially improves molecular understanding Runcie et al. (2025). We organize our evaluation into two groups:

- **Closed-Source Models:** We evaluate models from five providers. From OpenAI Achiam et al. (2023), we test the reasoning models o3 (high) and o4-mini (high), the GPT-5 family Singh et al. (2025) at two reasoning levels (low, high), GPT-5 Pro, GPT-5.2 Pro, and GPT-OSS 120B Agarwal et al. (2025) at high reasoning settings. From Google, we evaluate Gemini 2.5 Pro Comanici et al. (2025) at three reasoning budgets (8K, 16K, and 32K tokens) to isolate the effect of reasoning depth. From xAI, we include Grok 3 Mini at low and high reasoning and Grok 4.1 Fast. From Anthropic, we evaluate Claude Sonnet 4.5, Claude Opus 4.5 Anthropic (2025), and Claude Opus 4.6 Anthropic (2026) at two reasoning budgets (high, max).
- **Open-Source Models:** We evaluate DeepSeek Reasoner Guo et al. (2025) at three reasoning budgets (8K, 16K, and 32K tokens) and Llama-3.3-70B-Instruct Dubey et al. (2024). These models provide a baseline for open-weight performance on olfactory tasks.

In total, we evaluate 21 model configurations across 6 providers. All models were queried through their respective provider APIs without enabling web search, tool use, or any external retrieval capabilities; by default, none of the APIs grant models access to external resources during inference. All models are prompted identically using both isomeric SMILES and compound name representations; full prompt templates and worked examples are provided in Appendix C. For reasoning models, we systematically vary the reasoning budget to analyze the relationship between computational deliberation and olfactory task accuracy.

Evaluation Metrics. We adopt task-appropriate metrics to capture performance across the diverse question formats in the OP benchmark. For single-answer tasks (OC, OPD, OIn, OPl, OS, SIT), we report *any-overlap accuracy*: a question is scored correct if the set of tokens extracted from the model's response intersects the ground-truth token set. All prompts explicitly instruct models to respond with only the selected answer and no additional commentary; a post-hoc cleaning step verified compliance and corrected the very few cases (<1% of responses, limited to Claude 4.6 and Grok Mini reasoning traces) where extraneous tokens appeared. For multi-answer tasks (RATA, ORA), we employ *per-question multilabel F1*, which provides partial credit when models correctly identify a subset of applicable labels and penalises both spurious and missing predictions. For the three continuous-rating tasks (OIn, OPl, OS), we additionally compute Pearson correlations between predicted and human psychophysical ratings. Overall accuracy is the unweighted arithmetic mean of the eight per-task scores. More details on the answer extraction pipeline are provided in Appendix D.

Table 1: Olfactory Perception (OP) benchmark results using compound name prompts. Accuracy (%) is reported per task and overall (unweighted mean). **OC** = Odor Classification (n=175), **OPD** = Odor Primary Descriptor (n=175), **OIn** = Odor Intensity (n=175), **OPl** = Odor Pleasantness (n=175), **RATA** = Rate-All-That-Apply (n=100), **OS** = Odor Similarity (n=100), **ORA** = Olfactory Receptor Activation (n=80), **SIT** = Smell Identification Test (n=30). **Bold** = best overall; underlined = second best.

		Simple (N=700)				**Intermediate** (N=200)		**Hard** (N=110)		
Model	**Reasoning**	OC	OPD	OIn	OPl	RATA	OS	ORA	SIT	**Overall**
Closed-Source										
GPT-5	high	89.7	73.7	66.3	71.4	36.4	34.0	40.8	76.7	61.1
GPT-5	low	90.3	72.0	70.9	71.4	30.8	28.0	40.4	73.3	59.6
GPT-5 Pro	high	**92.0**	73.1	71.4	70.9	36.1	29.0	42.5	**80.0**	61.9
GPT-5.2 Pro	high	88.6	76.0	72.6	72.0	34.6	28.0	**52.8**	73.3	62.2
GPT-OSS-120B	high	82.9	60.6	65.1	72.0	25.1	34.0	35.6	56.7	54.0
o3	high	89.1	70.9	68.0	70.3	31.5	32.0	42.4	70.0	59.3
o4-mini	high	88.6	65.7	69.1	73.7	29.0	32.0	40.5	73.3	59.0
Gemini 2.5 Pro	16K	89.7	78.9	68.0	72.6	30.3	31.0	39.7	66.7	59.6
Gemini 2.5 Pro	32K	87.4	78.3	66.9	72.0	34.0	27.0	42.4	70.0	59.7
Gemini 2.5 Pro	8K	88.6	**80.0**	65.7	73.7	31.5	29.0	37.9	63.3	58.7
Grok 3 Mini	high	81.7	72.0	68.0	73.7	37.0	22.0	41.9	66.7	57.9
Grok 3 Mini	low	81.7	73.1	66.3	72.6	36.0	18.0	37.2	73.3	57.3
Grok 4.1 Fast	default	88.6	67.4	66.9	70.3	35.5	33.0	31.1	73.3	58.3
Claude Opus 4.5	high	**92.0**	76.6	71.4	73.1	**42.2**	25.0	45.8	70.0	62.0
Claude Opus 4.6	high	91.4	78.3	71.4	**74.9**	40.0	26.0	49.6	73.3	63.1
Claude Opus 4.6	max	**92.0**	77.7	**74.9**	74.3	38.9	26.0	51.1	**80.0**	**64.4**
Claude Sonnet 4.5	—	89.1	67.4	66.9	71.4	34.9	29.0	38.4	**80.0**	59.6
Open-Source										
DeepSeek Reasoner	16K	79.4	69.7	68.6	**74.9**	36.0	25.0	29.1	70.0	56.6
DeepSeek Reasoner	32K	81.1	73.7	69.7	73.1	33.1	32.0	31.6	73.3	58.5
DeepSeek Reasoner	8K	80.6	70.3	71.4	72.6	34.7	**35.0**	30.5	63.3	57.3
Llama 3.3 70B	—	83.4	60.6	68.0	72.0	26.8	29.0	35.0	46.7	52.7

5 RESULTS

In this section, we evaluate a diverse set of state-of-the-art LLMs on the OP benchmark to quantify current olfactory reasoning ability and identify systematic failure modes. We report results across all task categories and compare prompting conditions (compound names vs. isomeric SMILES) to isolate the effect of molecular representation on performance.

Overall Performance Table 1 and Figure 2 present benchmark performance across all evaluated models. The best-performing configuration, **Claude Opus 4.6 (max)**, attains 64.4% overall accuracy with compound name prompts, followed by **Claude Opus 4.6 (high)** at 63.1%, **GPT-5.2 Pro** at 62.2%, and **Claude Opus 4.5** at 62.0%. These scores substantially exceed chance levels for each task category (Figure 2b, dashed lines), yet remain far from ceiling performance, indicating that, while frontier models encode meaningful olfactory knowledge, substantial room for improvement remains.

A clear performance hierarchy emerges across providers. Anthropic and OpenAI frontier models occupy the top positions, with all configurations except GPT-OSS-120B surpassing 59% accuracy. The progression within the Claude family—from Sonnet 4.5 (59.6%) through Opus 4.5 (62.0%) to Opus 4.6 (max) (64.4%)—reflects consistent improvements associated with increased model capability and reasoning depth. OpenAI's GPT-5 variants demonstrate competitive performance, with GPT-5.2 Pro trailing Claude Opus 4.6 (max) by 2.2 percentage points. The o-series reasoning models perform well, with o3 (high) at 59.3% and o4-mini (high) at 59.0%. Google's Gemini 2.5 Pro (58.7–59.7%) and xAI's Grok family (57.3–58.3%) occupy the mid-tier, while DeepSeek Reasoner reaches 56.6–58.5%. The open-source Llama 3.3 70B lags behind all proprietary alternatives at

52.7%, underscoring a persistent capability gap between open-weight and closed-source systems for specialized perceptual reasoning.

Performance varies considerably across task categories. Simple tasks (OC, OPD, OIn, OPl) produce the highest accuracy, with top models reaching 92.0% on odor classification and 80.0% on primary descriptor identification. Intermediate tasks (RATA, OS) prove more demanding, with best performance limited to 42.2% and 35.0% respectively. Hard tasks exhibit the widest variance: SIT reaches 80.0% for three models, benefiting from knowledge about foods and beverages, whereas ORA peaks at 52.8% (GPT-5.2 Pro), still a challenging task due to the specialized biochemistry knowledge required. A detailed question-level difficulty analysis is provided in Appendix E.1 (Figure 5) for single-label tasks and Appendix E.2 (Figure 6) for the multi-label tasks RATA and ORA.

Isomeric SMILES vs. Compound Name Prompts A consistent and substantial performance gap separates the two molecular representation formats (Figure 2a). Across all 21 model configurations, compound name prompts outperform isomeric SMILES notation, with improvements ranging from

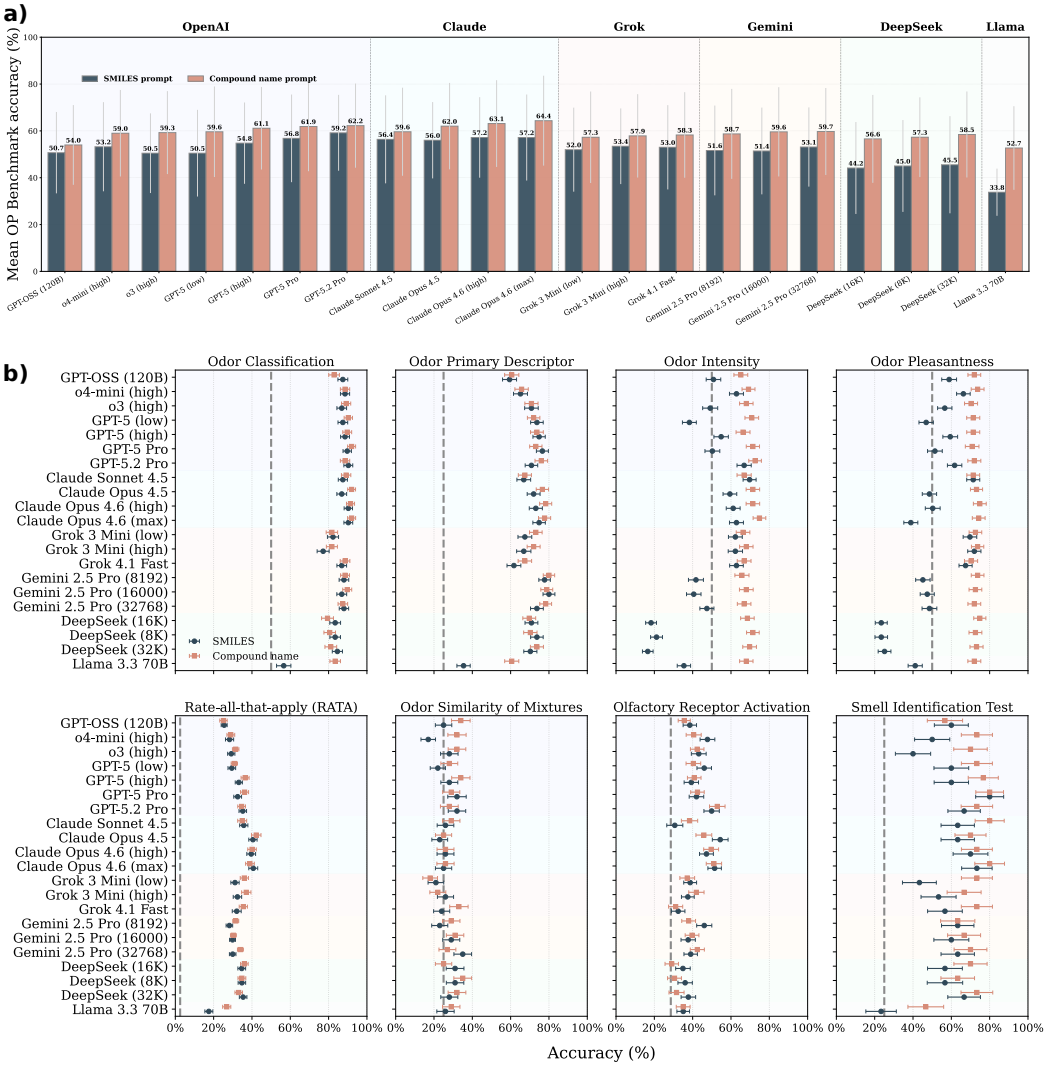

Figure 2: **Olfactory Perception (OP) benchmark performance.** **(a)** Overall mean accuracy per model for isomeric SMILES and compound name prompts. Error bars: 95% confidence intervals. **(b)** Per-category accuracy for each model under both prompts. Horizontal error bars are bootstrap standard deviations of the mean accuracy. Gray dashed lines indicate chance baselines.

2.4 percentage points (Claude Opus 4.5: 59.6% $\rightarrow$ 62.0%) to 18.9 percentage points (Llama 3.3 70B: 33.8% $\rightarrow$ 52.7%). DeepSeek Reasoner (8K) shows a 12.3-point improvement (45.0% $\rightarrow$ 57.3%). The mean improvement across models is approximately 7 percentage points, suggesting that current LLMs access olfactory knowledge primarily through lexical associations rather than structural molecular reasoning. Frontier reasoning models display the narrowest gaps, with GPT-5 Pro, GPT-5.2 Pro, and Claude Opus 4.5 each retaining over 95% of their compound name accuracy under SMILES prompts. Conversely, Llama 3.3 70B exhibits the largest disparity, with isomeric SMILES accuracy barely exceeding chance. Per-task analysis (Figure 2b) reveals that the isomeric SMILES/name gap varies substantially by task category. OC exhibits the smallest gaps, potentially because odorousness correlates with molecular properties (volatility, molecular weight) that can be partially inferred from isomeric SMILES structure. OIn, OPl, and SIT show the largest gaps, consistent with tasks where identifying the target molecule or food source is critical and far easier from a name than from structural notation. In contrast, multi-label tasks (RATA, ORA) show modest gaps, suggesting that descriptor-level knowledge is similarly accessible, or similarly limited, under both representations.

Effect of Reasoning Budget Systematic variation of reasoning token budgets reveals consistent but modest performance gains. Claude Opus 4.6 shows clear scaling from high (63.1%) to max (64.4%), and the GPT-5 family improves from 59.6% (low) to 62.2% (GPT-5.2 Pro). Similar patterns appear across other providers, though DeepSeek Reasoner shows a non-monotonic pattern, suggesting optimal reasoning budgets may be task-dependent. Overall, no model gains more than approximately 2 percentage points from extended reasoning, contrasting with chemistry benchmarks where reasoning provided substantially larger gains Runcie et al. (2025).

Fine-Grained Performance Analysis Beyond categorical accuracy, we assess whether models capture the continuous structure of human olfactory perception. Figure 3a presents Pearson correlations between model-predicted ratings and human psychophysical measurements: the best models reach $r \approx 0.55$ for OIn, approaching specialized model performance Keller et al. (2017); Satarifard et al. (2025) (red dashed line); OPl correlations are higher ($r \approx 0.60$); and OS exhibits the weakest correlations ($r \approx 0.35$), confirming that models struggle to integrate perceptual information across molecules. Isomeric SMILES and compound name prompts produce similar correlations across all three dimensions, with a slightly larger gap for OIn. Turning to the multi-label tasks, RATA and ORA exhibit distinct difficulty profiles: RATA F1 scores are mainly bell-shaped, indicating partial credit on most questions, while ORA is bimodal: models either possess the relevant receptor-ligand knowledge or lack it entirely (Appendix E.2, Figure 6). Per-label analysis (Figure 3b) reveals a clear divide for RATA: descriptors with well-established functional-group associations (e.g., sulfurous, floral, fruity Genva et al. (2019)) achieve the highest mean F1, while descriptors such as spicy, fresh, and tropical prove nearly impossible. A detailed case study of this failure mode is provided in Appendix E.3. For ORA, per-label difficulty varies sharply across receptors (Figure 3b): the wildtype `hOR2W1` (appearing in 30 label assignments) is reliably predicted at F1 above 0.6, and the `M81V` variant is moderately well-predicted, but `hOR2W1_D296N` (24) and `hOR52D1` (5) sit near zero for most models. The `hOR52D1` gap likely reflects data scarcity, while `D296N` reveals a specific knowledge error: Claude Opus 4.6 (max) never predicts this receptor across all 24 label appearances, whereas GPT-5.2 Pro correctly identifies it in 19 cases. Inter-model variance is substantially higher for receptor labels than for semantic descriptors, indicating that receptor knowledge is more idiosyncratic across model families; a detailed analysis of these knowledge gaps is provided in Appendix F.2 (Figure 10).

Systematic Failure Modes Two qualitatively distinct failure mechanisms underlie the hardest categories. For OS, models use molecular overlap as a proxy for perceptual similarity: accuracy reaches $\sim$85% when similar mixtures share many molecules but drops to near 0% when similar mixtures share few, and all models exhibit a systematic bias toward predicting dissimilarity; Claude models assign "Slightly Dissimilar" to the vast majority of mixtures, even though perceptually similar mixtures often share zero molecules. This heuristic renders mixture-level olfactory similarity fundamentally beyond current LLM capabilities; a detailed analysis is provided in Appendix F.1 (Figures 8, 7, 9). Additionally, Claude Opus 4.6's safety filter refuses OC questions about hazardous compounds (e.g., nerve agents), illustrating a tension between safety alignment and scientific evaluation (Appendix F.3).

Multilingual Evaluation To probe whether olfactory knowledge is language-specific, we translated the RATA task prompts into 21 languages spanning over six language families and evaluated seven models per language. Figure 4a shows that English achieves the highest mean per-question F1, followed closely by French, Spanish, and Russian, while non-Indo-European languages (Korean,

Chinese, Swahili) cluster at the lower end. Figure 4b presents language-family AUROC curves for DeepSeek (32K): Germanic and Romance languages achieve the highest family-level AUROC (0.778 and 0.774, respectively), while the remaining families score lower (0.725–0.749). An ensemble aggregating votes across all 21 languages achieves AUROC = 0.828, suggesting that multilingual aggregation captures complementary olfactory knowledge. Figure 4c shows per-model AUROC using an all-language vote-fraction ensemble: Gemini 2.5 Pro (32K) leads at 0.864, and a cross-model

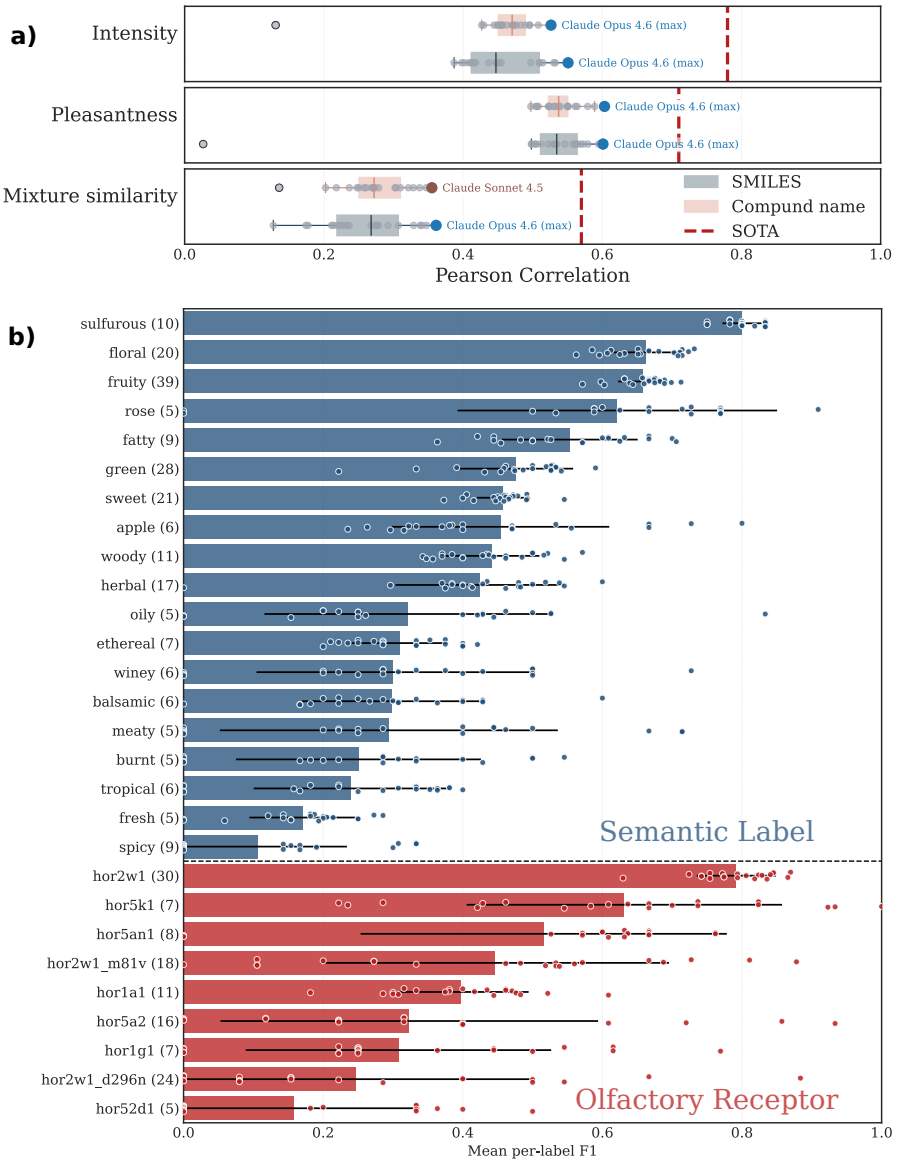

Figure 3: **Predictions correlations and performance difficulty across olfactory tasks. (a)** Pearson correlations across models for three continuous-rating categories (odor intensity, odor pleasantness, and mixture similarity). Each row shows the distribution of model correlations for isomeric SMILES and Compound name prompts; gray points represent individual models, and the best-performing model per prompt/category is highlighted with a colored circle. State-of-the-art reference performance is indicated by red dashed lines. **(b)** Combined label-difficulty ranking for RATA (blue) and olfactory-receptor activation (red). For each label, per-model F1 scores are computed in a multilabel setting; the bar shows the mean F1 across models with an error bar (standard deviation), while points represent individual model values.

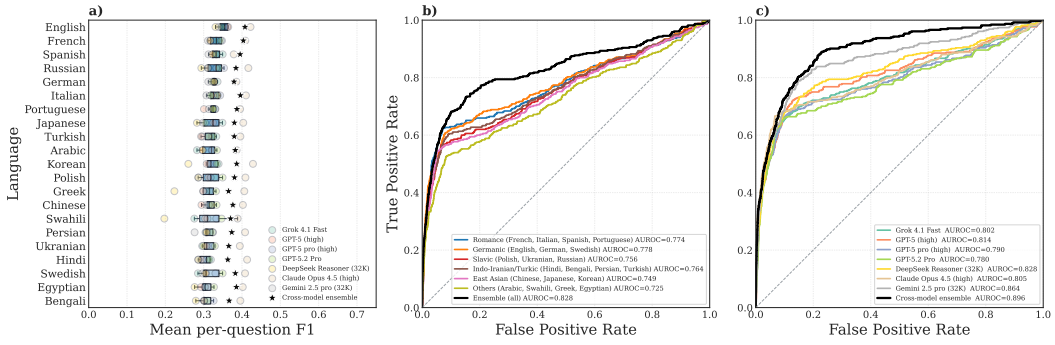

Figure 4: **Multilingual RATA performance for compound name prompt.** **(a)** Per-language distributions of model performance, summarized as mean per-question multilabel F1, colored points denote individual models and the black star denotes a cross-model ensemble. **(b)** DeepSeek (32K) AUROC by language family using per-label vote fractions across languages within each family (East Asian and Others are not same language families), black curve represents an ensemble of all languages. **(c)** AUROC for each model using an all-language majority vote ensemble, plus a cross-model ensemble pooling all models from all languages. Dashed diagonal indicates chance.

ensemble combining all seven models reaches 0.896, indicating that model diversity further improves multilingual olfactory prediction beyond language diversity alone.

6 Conclusion, Limitations, and Future Work

Here, we introduce the Olfactory Perception (OP) benchmark, a structured evaluation suite for testing whether large language models can reason about smell from molecular information and real-world odor sources. Across a broad set of olfactory tasks, we observe emerging but incomplete capabilities: the best-performing systems achieve moderate accuracy, yet performance varies substantially by task type. A consistent pattern is the gap between prompts using compound names versus isomeric SMILES, suggesting current LLMs often succeed via lexical associations rather than robust structural reasoning. Taken together, our results position olfaction as a challenging and underexplored modality for LLM evaluation, and provide a concrete benchmark for measuring progress toward models that can connect chemistry to sensory perception. The benchmark makes deliberate design trade-offs: we use standardized descriptor vocabularies and discrete response formats, which enable reproducible automatic evaluation but do not capture the full richness of odor perception (e.g., contextual effects, individual and cultural variation, nuanced free-form descriptions). As with other instruction-following benchmarks, results can be influenced by prompting and output formatting sensitivity, despite using consistent templates. Multilingual variants rely on automated translation with validation, which may introduce subtle connotation shifts. Benchmark accuracy should be interpreted as agreement with established labels rather than as evidence of mechanistic olfactory understanding. Several directions could strengthen future work. On the data side, expanding to mixtures, concentration effects, temporal dynamics, broader receptor-level coverage, and multimodal inputs (e.g., verbal descriptions, facial expressions during smelling) would better reflect real olfaction. Incorporating human evaluation and cross-cultural annotation could capture partial correctness and reduce brittleness to surface forms. On the modeling side, hybrid systems combining LLMs with cheminformatics tools, methods encouraging structure-based reasoning (e.g., adversarial splits, representation-invariant prompting), and low-cost strategies such as prompt repetition Leviathan et al. (2025) are promising directions. Finally, extending to generative settings, such as proposing molecules with target odor profiles under safety and synthesizability constraints, would connect benchmark performance to applications in fragrance, flavor, and environmental monitoring.

REFERENCES

Erratum for the report humans can discriminate more than 1 trillion olfactory stimuli by c. bushdid et al. *Science*, 383(6685):eado6457, 2024. doi: 10.1126/science.ado6457. URL https://www.science.org/doi/abs/10.1126/science.ado6457.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, et al. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*, 2025.

Anthropic. Claude Opus 4.5 system card. Technical report, 2025. URL https://www.anthropic.com/claude-opus-4-5-system-card.

Anthropic. Claude Opus 4.6 system card. Technical report, 2026. URL https://www-cdn.anthropic.com/0dd865075ad3132672ee0ab40b05a53f14cf5288.pdf.

Jennifer Bartsch, Erik Uhde, and Tunga Salthammer. Analysis of odour compounds from scented consumer products using gas chromatography-mass spectrometry and gas chromatography-olfactometry. *Analytica chimica acta*, 904:98–106, 2016. URL https://api.semanticscholar.org/CorpusID:7550203.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020. URL https://arxiv.org/abs/2005.14165.

Linda B. Buck and Richard Axel. A novel multigene family may encode odorant receptors: a molecular basis for odor recognition. *Cell*, 65 1:175–87, 1991. URL https://api.semanticscholar.org/CorpusID:6548928.

Caroline Bushdid, Marcelo O Magnasco, Leslie B Vosshall, and Andreas Keller. Humans can discriminate more than 1 trillion olfactory stimuli. *Science*, 343(6177):1370–1372, 2014.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code, 2021. URL https://arxiv.org/abs/2107.03374.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv e-prints*, pp. arXiv–2407, 2024.

Sergio Esteban-Romero, Iván Martín-Fernández, Manuel Gil-Martín, and Fernando Fernández-Martínez. Synthesizing olfactory understanding: Multimodal language models for image–text smell matching. *Symmetry*, 17(8):1349, 2025.

Makoto Fukushima, Shusuke Eshita, and Hiroshige Fukuhara. Advancements and limitations of llms in replicating human color-word associations. *Discover Artificial Intelligence*, 5(1):64, 2025.

Xinyu Ge, Yongjie Zhou, Qing Li, Yuqing Tan, Yongkang Luo, and Hui Hong. Machine learning for food flavor prediction and regulation: models, data integration, and future perspectives. *Journal of advanced research*, 2025. URL https://api.semanticscholar.org/CorpusID:282185617.

Manon Genva, Tierry Kenne Kemene, Magali Deleu, Laurence Lins, and Marie-Laure Fauconnier. Is it possible to predict the odor of a molecule on the basis of its structure? *International journal of molecular sciences*, 20(12):3018, 2019.

Richard C Gerkin and Jason B Castro. The number of olfactory stimuli that humans can discriminate is still unknown. *eLife*, 4:e08127, jul 2015. ISSN 2050-084X. doi: 10.7554/eLife.08127. URL https://doi.org/10.7554/eLife.08127.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Taicheng Guo, Bozhao Nan, Zhenwen Liang, Zhichun Guo, Nitesh Chawla, Olaf Wiest, Xiangliang Zhang, et al. What can large language models do in chemistry? a comprehensive benchmark on eight tasks. *Advances in Neural Information Processing Systems*, 36:59662–59688, 2023.

Yang Han, Ziping Wan, Lu Chen, Kai Yu, and Xin Chen. From generalist to specialist: A survey of large language models for chemistry, 2024. URL https://arxiv.org/abs/2412.19994.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

International Fragrance Association. The ifra fragrance ingredient glossary (fig), 2020. URL https://d3t14p1xronwr0.cloudfront.net/docs/ifra-fragrance-ingredient-glossary-april-2020.pdf.

Andreas Keller, Richard C Gerkin, Yuanfang Guan, Amit Dhurandhar, Gabor Turu, Bence Szalai, Joel D Mainland, Yusuke Ihara, Chung Wen Yu, Russ Wolfinger, et al. Predicting human olfactory perception from chemical features of odor molecules. *Science*, 355(6327):820–826, 2017.

J. Kreissl, V. Mall, P. Steinhaus, and M. Steinhaus. Leibniz-lsb@tum odorant database, version 1.2, 2022. URL https://www.leibniz-lsb.de/en/databases/leibniz-lsbtum-odorant-database.

Murathan Kurfalı, Jonas K Olofsson, and Thomas Hörberg. Enhancing multimodal language models with olfactory information. 2023.

Murathan Kurfalı, Pawel Herman, Stephen Pierzchajlo, Jonas Olofsson, and Thomas Hörberg. Representations of smells: The next frontier for language models? *Cognition*, 264:106243, 2025.

Maxence Lalis, Matej Hladiš, Samar Abi Khalil, Loïc Briand, Sébastien Fiorucci, and Jérémie Topin. M2or: a database of olfactory receptor–odorant pairs for understanding the molecular mechanisms of olfaction. *Nucleic Acids Research*, 52(D1):D1370–D1379, 2024.

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.

Yaniv Leviathan, Matan Kalman, and Yossi Matias. Prompt repetition improves non-reasoning llms. *arXiv preprint arXiv:2512.14982*, 2025.

Joel Mainland, Yun Li, Ting Zhou, Wen Liu, and Hiroaki Matsunami. Human olfactory receptor responses to odorants. *Scientific Data*, 2:150002, 02 2015. doi: 10.1038/sdata.2015.2.

Asifa Majid. Human olfaction at the intersection of language, culture, and biology. *Trends in Cognitive Sciences*, 25(2):111–123, 2021. ISSN 1364-6613. doi: https://doi.org/10.1016/j.tics.2020.11.005. URL https://www.sciencedirect.com/science/article/pii/S1364661320302771.

Asifa Majid and Niclas Burenhult. Odors are expressible in language, as long as you speak the right language. *Cognition*, 130(2):266–270, 2014. ISSN 0010-0277. doi: https://doi.org/10.1016/j.cognition.2013.11.004. URL https://www.sciencedirect.com/science/article/pii/S001002771300214X.

Asifa Majid, Niclas Burenhult, Marcus Stensmyr, Josje Valk, and Bill Hansson. Olfactory language and abstraction across cultures. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 373, 06 2018. doi: 10.1098/rstb.2017.0139.

Haitao Mao, Linlin Liu, Jian Du, and Rafiqul Gani. A machine learning based computer-aided molecular design/screening methodology for fragrance molecules. *Computers & Chemical Engineering*, 115, 04 2018. doi: 10.1016/j.compchemeng.2018.04.018.

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.

Emily J Mayhew, Charles J Arayata, Richard C Gerkin, Brian K Lee, Jonathan M Magill, Lindsey L Snyder, Kelsie A Little, Chung Wen Yu, and Joel D Mainland. Transport features predict if a molecule is odorous. *Proceedings of the National Academy of Sciences*, 119(15):e2116576119, 2022.

Adrian Mirza, Nawaf Alampara, Sreekanth Kunchapu, Martiño Ríos-García, Benedict Emoekabu, Aswanth Krishnan, Tanya Gupta, Mara Schilling-Wilhelmi, Macjonathan Okereke, Anagha Aneesh, et al. A framework for evaluating the chemical knowledge and reasoning abilities of large language models against the expertise of chemists. *Nature Chemistry*, pp. 1–8, 2025.

Andrew Oxenham, Christophe Micheyl, Michael Keebler, Adam Loper, and Sébastien Santurette. Pitch perception beyond the traditional existence region of pitch. *Proceedings of the National Academy of Sciences of the United States of America*, 108:7629–34, 05 2011. doi: 10.1073/pnas.1015291108.

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.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. Gpqa: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*, 2024.

Nicholas T Runcie, Charlotte M Deane, and Fergus Imrie. Assessing the chemical intelligence of large language models. *Journal of Chemical Information and Modeling*, 2025.

Pinaki Saha, Mrityunjay Sharma, Sarabeshwar Balaji, Aryan Amit Barsainyan, Ritesh Kumar, Volker Steuber, and Michael Schmuker. Dense sense: a novel approach utilizing electron density augmented machine learning paradigm to understand the complex odour landscape. *Digital Discovery*, 4(11):3339–3350, 2025.

Vahid Satarifard, Laura Sisson, Yikun Han, Pedro Ilídio, Matej Hladiš, Maxence Lalis, Xuebo Song, Wenjie Yin, Aharon Ravia, CiCi Xingyu Zheng, et al. High-fidelity tuning of olfactory mixture distances in the perceptual space of smell through a community effort. *biorxiv*, pp. 2025–12, 2025.

Lavi Secundo, Kobi Snitz, and Noam Sobel. The perceptual logic of smell. *Current opinion in neurobiology*, 25C:107–115, 01 2014. doi: 10.1016/j.conb.2013.12.010.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

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.

Andrew Stockman and Lindsay T. Sharpe. The spectral sensitivities of the middle- and long-wavelength-sensitive cones derived from measurements in observers of known genotype. *Vision research*, 40 13:1711–37, 2000. URL https://api.semanticscholar.org/CorpusID:7886523.

Gary Tom, Cher Tian Ser, Ella M Rajaonson, Stanley Lo, Hyun Suk Park, Brian K Lee, and Benjamin Sanchez-Lengeling. From molecules to mixtures: Learning representations of olfactory mixture similarity using inductive biases. *arXiv preprint arXiv:2501.16271*, 2025.

David Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.*, 28:31–36, 1988. URL https://api.semanticscholar.org/CorpusID:5445756.

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*, 9(9):1871–1886, 2025.

Botao Yu, Frazier N Baker, Ziqi Chen, Xia Ning, and Huan Sun. Llasmol: Advancing large language models for chemistry with a large-scale, comprehensive, high-quality instruction tuning dataset. *arXiv preprint arXiv:2402.09391*, 2024.

Mengji Zhang, Yusuke Hiki, Akira Funahashi, and Tetsuya J Kobayashi. Mol-peco: a deep learning model to predict human olfactory perception from molecular structures. *arXiv preprint arXiv:2305.12424*, 2023.

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.

A Benchmark Question Categories

Table 2 summarizes the eight question categories in the OP benchmark, along with their response formats, data sources, and difficulty levels. Categories are grouped into three tiers: simple tasks (binary or paired-comparison judgments), intermediate tasks (multi-label profiling and mixture-level perception), and hard tasks (receptor biology and real-world odor source identification). Question counts range from 30 (Smell Identification) to 175 (each of the four simple categories), totaling 1,010 questions. Each category draws ground-truth answers from a distinct peer-reviewed dataset, as indicated in the reference column.

Table 2: Question categories and data sources used in the benchmark.

Question Category	Options	n	Note	Type	Reference
Odor Classification	Odorous/Odorless	175	Molecular Weight $\leq$ 350.0 50/50	Pure	Mayhew et al. (2022)
Odor Descriptor	1 answer, 3 distractors	175	Stratified to represent less dominant descriptors, 29 Descriptors total	Pure	International Fragrance Association (2020)
Odor Intensity	High/Low	175	Above and below median, with score difference of 10	Pure	Keller et al. (2017)
Odor Pleasantness	High/Low	175	Above and below median, with score difference of 10	Pure	Keller et al. (2017)
Rate-all-that-apply	2-5 answers from 138 distractors	100	25 questions for 2,3,4, and 5 descriptors	Pure	Lee et al. (2023)
Odor Similarity of Mixtures	Rater scale	100	Mixtures with 2-10 molecules	Mixture	Snitz et al. (2013); Bushdid et al. (2014); Ravia et al. (2020)
Olfactory Receptor Activation	1-10 answers from 4-10 ORs	80	Human OR, 4-10 OR studied	Pure	Lalis et al. (2024)
Smell Identification Test	1 answer, 3 distractors	30		Mixture	Kreissl et al. (2022)

B Detailed Comparison with Prior Work

This appendix provides an expanded discussion of related work, complementing the summary in Section 2, followed by a side-by-side comparison table.

LLM benchmarks for Chemistry and Molecular Understanding. Recent benchmarks have evaluated the chemical reasoning capabilities of large language models. ChemBench Mirza et al. (2025) established a comprehensive chemistry benchmark with over 2,700 questions across eight sub-categories including organic chemistry, toxicity, and medicinal chemistry, finding that leading models outperform human chemists on average. ChemIQ Runcie et al. (2025) focuses specifically on molecular comprehension with 816 questions testing SMILES interpretation, atom counting, reaction prediction, and NMR structure elucidation-reasoning models (o3-mini, Gemini Pro 2.5); it achieved 50-57% accuracy, substantially outperforming non-reasoning models (3-7%). Additionally, LlaSMol Yu et al. (2024) introduced SMolInstruct, a large-scale instruction tuning dataset with over 3 million samples across 14 chemistry tasks, and demonstrated that fine-tuned open-source LLMs can substantially outperform GPT-4 on chemistry tasks, with SMILES representations outperforming SELFIES for molecular understanding. Other benchmarks include GPQA Rein et al. (2024) with expert-written chemistry questions, MMLU Hendrycks et al. (2021) which includes chemistry among its 57 subjects, and a comprehensive eight-task benchmark Guo et al. (2023).

While these benchmarks comprehensively assess structural and general chemical understanding, none evaluate whether LLMs can reason about the perceptual properties of molecules, in particular how they smell. Our work addresses this gap by testing the translation from molecular structure to olfactory perception.

LLMs and Sensory Perception. A growing body of work investigates whether LLMs align with human sensory judgments. It has been shown that GPT models produce similarity judgments significantly correlated with human data across six modalities: pitch, loudness, colors, consonants, taste, and timbre Marjieh et al. (2024). Notably, olfaction was excluded from their evaluation, leaving an open question about LLM capabilities in this domain. This line of inquiry is extended Xu et al. (2025) by comparing LLM representations of ∼4,442 lexical concepts against human norms across non-sensorimotor, sensory, and motor domains, finding that alignment decreases systematically from non-sensorimotor to sensory domains and is minimal for motor concepts; adding visual training improves sensory but not motor alignment. These findings highlights a grounding gap that is particularly acute for embodied perception. Even within well-studied modalities, limitations persist. For example, GPT-3 through GPT-4o were tested on color-word associations against over 10,000 Japanese participants Fukushima et al. (2025) and found that GPT-4o peaked at approximately 50% accuracy, with strong variation across word categories. Together, these findings suggest that LLMs capture substantial but incomplete perceptual structure from text, with performance degrading for modalities that are less frequently described in language. Olfaction, rarely discussed in explicit perceptual terms, represents a natural test case for these limitations.

LLMs and Olfaction. Concurrent work has begun exploring the olfaction gap specifically. SNIFF AI Zhong et al. (2024) investigated human-AI perceptual alignment for smell through user studies where participants described scents and an LLM embedding model attempted identification. Their findings revealed limited alignment, with biases toward certain scents (e.g., lemon, peppermint) and systematic failures on others (e.g., rosemary), achieving only 27.5% success in their scent description task.

Most recently, three generations of language models are systematically evaluated Kurfalı et al. (2025), from static word embeddings (Word2Vec, FastText) to encoder-based models (BERT) and decoder-based LLMs (GPT-4o, Llama 3.1), on their ability to recover olfactory information from natural language. Testing under nearly 200 training configurations across three odor datasets, they found that GPT-4o excels at simulating olfactory-semantic relationships, particularly on tasks where odor similarities are derived from word-based assessments. However, their evaluation focuses on semantic similarity judgments rather than factual accuracy about olfactory properties. Related multimodal work has explored image-text olfactory matching Kurfalı et al. (2023); Esteban-Romero et al. (2025).

Our benchmark complements these approaches by testing whether LLMs possess correct factual knowledge about odor classification, perceptual attributes, and biological mechanisms through structured question-answering with objective ground-truth answers.

Machine Learning for Olfactory Perception. Parallel to LLM development, specialized machine learning models have made significant progress in computational olfaction. Two DREAM Olfaction Prediction Challenges Keller et al. (2017); Satarifard et al. (2025) established foundational work on predicting semantic descriptors of single molecules and olfactory perceptual similarity of mixtures.

In addition, a Principal Odor Map has been introduced Lee et al. (2023) using graph neural networks, and it achieved human-level proficiency in describing odor qualities across 500,000 potential scent molecules, with AUROC = 0.89 for multi-label odor descriptor prediction on the GS-LF dataset (the same 138-descriptor vocabulary used in our RATA task). More recently, Mol-PECO Zhang et al. (2023) extended graph neural network approaches with Coulomb matrix representations to predict 118 odor descriptors from molecular structure, achieving AUROC of 0.813. Dense Sense Saha et al. (2025) further augmented graph neural networks with quantum-mechanical electron density features, reaching AUROC = 0.875 on the same odor descriptor prediction task. POMMIX Tom et al. (2025) further extended the POM to olfactory mixture similarity, combining attention-based mixture representations with inductive biases to reach Pearson $\rho = 0.779$ on human perceptual similarity data, directly relevant to the mixture-level tasks we evaluate in this benchmark.

These specialized systems demonstrate that machine learning can capture structure-odor relationships when purpose-built for olfactory tasks. However, they do not address whether general-purpose language models that are primarily trained on text without explicit olfactory supervision possess latent knowledge about smell.

Positioning of Our Work. Our Olfactory Perception (OP) benchmark is, to our knowledge, the first comprehensive evaluation of LLM olfactory reasoning through structured factual question-answering. While ChemIQ and ChemBench focus on molecular structure understanding, and SNIFF AI evaluates embedding-space alignment through subjective descriptions, our benchmark tests discrete factual knowledge across eight task categories: odor classification, odor primary descriptors, intensity, pleasantness, multi-descriptor identification, mixture similarity, receptor activation, and smell identification. Ground-truth answers are derived from established olfactory science datasets and resources (Mayhew et al., 2022; Keller et al., 2017; Lee et al., 2023; Lalis et al., 2024; Snitz et al., 2013; Bushdid et al., 2014; Ravia et al., 2020; International Fragrance Association, 2020; Kreissl et al., 2022). Furthermore, our dual-prompting strategy (isomeric SMILES vs. compound names) enables direct comparison of how different molecular representations affect olfactory reasoning-an experimental design not explored in prior work. Table 3 summarizes the positioning of our benchmark relative to prior work.

Table 3: Comparison of related work on LLMs, chemistry, and olfaction.

Work	Input	Specific Tasks	What is Tested
Chemistry LLM benchmarks			
ChemBench Mirza et al. (2025)	Chemistry questions	General, organic, analytical, toxicity chemistry (2,700 Q)	Structural & reaction knowledge
ChemIQ Runcie et al. (2025)	SMILES	Carbon counting, ring counting, SMILES-to-IUPAC, NMR elucidation, reaction prediction (816 Q)	Molecular comprehension
LlaSMol Yu et al. (2024)	SMILES/SELFIES	Name conversion, property prediction, synthesis (14 tasks, 3M samples)	Fine-tuned LLM chemistry
LLMs and Sensory Perception			
Marjieh et al. Marjieh et al. (2024)	Sensory stimuli	Pairwise similarity judgments across pitch, loudness, color, consonants, taste, timbre	6 modalities (olfaction excluded)
Xu et al. Xu et al. (2025)	Lexical concepts	Conceptual similarity across sensorimotor domains ($\sim$4,442 concepts)	Grounding gap in sensory domains
Fukushima et al. Fukushima et al. (2025)	Color stimuli	Color-word associations (17 colors $\times$ 80 words)	Color perception alignment
LLMs and Olfaction			
SNIFF AI Zhong et al. (2024)	Human descriptions	Human describes scent, LLM identifies source (27.5% success)	Description-to-scent mapping
Kurfalı et al. Kurfalı et al. (2025)	Odor word pairs	Pairwise similarity ratings across 3 odor datasets	Semantic similarity of odor words
Kurfalı et al. Kurfalı et al. (2023)	Image + text	Detect if image and text share olfactory source	Multimodal matching
Esteban-Romero et al. Esteban-Romero et al. (2025)	Image + text	Image-text smell matching (F1=0.76)	Multimodal matching
ML for Olfactory Perception (non-LLM)			
First DREAM Challenge Keller et al. (2017)	Molecular features	Predict intensity, pleasantness, 19 semantic descriptors (476 odorants)	Specialized ML models
Second DREAM Challenge Satarifard et al. (2025)	Molecular features	Predict olfactory perceptual distance of mixtures	Specialized ML models
Principal Odor Map Lee et al. (2023)	Molecular graphs	Predict odor qualities, similarity (500K molecules)	GNN on molecules
Mol-PECO Zhang et al. (2023)	Molecular graphs (Coulomb matrix)	Predict 118 odor descriptors (8,503 molecules)	Deep learning for QSOR
POMMIX Tom et al. (2025)	Molecular graphs (GNN + attention)	Predict olfactory mixture similarity	Mixture representation learning
OP benchmark (Ours)	**Isomeric SMILES / Compound Name**	**Odor classification, primary descriptor, intensity, pleasantness, rate-all-that-apply, mixture similarity, receptor activation, smell identification (1,010 Q)**	**LLM olfactory knowledge**

C PROMPT TEMPLATES

This appendix presents the exact prompt templates used for each question category in the OP benchmark, together with one real example drawn directly from the benchmark dataset. Each task is evaluated using two prompt variants: one using isomeric SMILES molecular representations and one using common compound names. All prompts instruct models to respond without additional commentary to facilitate automated answer extraction.

C.1 ODOR CLASSIFICATION

The odor classification task tests whether models can determine if a molecule has a detectable odor to humans ($n = 175$).

Template – Isomeric SMILES Prompt

```
Does [{SMILES}] have a detectable odor to humans? Only
respond with Odorous or Odorless. Do not write any comments.
```

Template – Name Prompt

```
Does [{compound_name}] have a detectable odor to humans?
Only respond Odorous or Odorless. Do not write any comments.
```

Example: Phenacetin (Odorless)

Isomeric SMILES prompt:

```
Does [CCOC1=CC=C(C=C1)NC(=O)C] have a detectable odor to humans?
Only respond with Odorous or Odorless. Do not write any comments.
```

Name prompt:

```
Does [Phenacetin] have a detectable odor to humans? Only respond
Odorous or Odorless. Do not write any comments.
```

Ground-truth answer: Odorless

Model predictions (compound name prompt):

Gemini 2.5 Pro	Odorless	✓
GPT-5 Pro	Odorless	✓
o3 (high)	Odorless	✓

C.2 ODOR PRIMARY DESCRIPTOR

This task requires models to identify the single odor primary descriptor for a molecule from four options ($n = 175$). The four options consist of the correct answer and three distractor descriptors.

Template – Isomeric SMILES Prompt

```
What is the odor primary descriptor of molecule [{SMILES}]?
Select only one from the [{OPTIONS}]. Do not write any
comments.
```

Template – Name Prompt

```
What is the odor primary descriptor of molecule
[{compound_name}]? Select only one from the [{OPTIONS}].
Do not write any comments.
```

Example: Octanal diethyl acetal

Isomeric SMILES prompt:

```
What is the odor primary descriptor of molecule [CCCCCCCC(OCC)OCC]?
Select only one from the [Green;Honey;Herbal;Marine]. Do not write
any comments.
```

Name prompt:

```
What is the odor primary descriptor of molecule [Octanal diethyl
acetal]? Select only one from the [Green;Honey;Herbal;Marine]. Do
not write any comments.
```

Ground-truth answer: Green

Model predictions (compound name prompt):

Gemini 2.5 Pro	Green	✓
GPT-5 Pro	Green	✓
o3 (high)	Green	✓

C.3 Odor Intensity

This task presents pairs of molecules and asks models to identify which has higher perceived odor intensity, plus provide numerical intensity estimates ($n = 175$).

Template – Isomeric SMILES Prompt

```
Which molecule is likely to smell more intense to humans:
{SMILES_1} or {SMILES_2}? Select only one of the SMILES.
If you had to rate the intensity of these molecules from
0 (extremely low) to 100 (highly intense), what would you
assign to each? Respond with the selected compound name (the
one with higher intensity), followed by two intensity values
in the order the molecules are listed. Use semicolons (;) as
separators. Do not write any comments.
```

Template – Name Prompt

```
Which molecule is likely to smell more intense to humans:
{compound_name_1} or {compound_name_2}? Select only one
of the compound names. If you had to rate the intensity
of these molecules from 0 (extremely low) to 100 (highly
intense), what would you assign to each? Respond with the
selected compound name (the one with higher intensity),
followed by two intensity values in the order the molecules
are listed. Use semicolons (;) as separators. Do not write
any comments.
```

Example: hexan-2-one vs. pyrazine

Isomeric SMILES prompt:

```
Which molecule is likely to smell more intense to humans:
CCCCC(=O)C or C1=CN=CC=N1? Select only one of the SMILES. If you
had to rate the intensity of these molecules from 0 (extremely low)
to 100 (highly intense), what would you assign to each? Respond
with the selected compound name (the one with higher intensity),
followed by two intensity values in the order the molecules are
listed. Use semicolons (;) as separators. Do not write any
comments.
```

Name prompt:
```
Which molecule is likely to smell more intense to humans:
hexan-2-one or pyrazine?  Select only one of the compound names.
If you had to rate the intensity of these molecules from 0
(extremely low) to 100 (highly intense), what would you assign
to each?  Respond with the selected compound name (the one with
higher intensity), followed by two intensity values in the order
the molecules are listed.  Use semicolons (;) as separators.  Do
not write any comments.
```

Ground-truth answer: hexan-2-one (intensity = 72.4 vs. 20.8)
Expected response format: `hexan-2-one;72;21`

Model predictions (compound name prompt):

Gemini 2.5 Pro	`Pyrazine; 45; 95`	✗
GPT-5 Pro	`pyrazine;30;70`	✗
o3 (high)	`pyrazine;30;85`	✗

All three models incorrectly select pyrazine as more intense.

C.4 Odor Pleasantness

This task presents pairs of molecules and asks models to identify which smells more pleasant, plus provide numerical pleasantness estimates ($n = 175$).

Template – Isomeric SMILES Prompt

```
Which molecule is likely to smell more pleasant to humans:
{SMILES_1} or {SMILES_2}?  Select only one of the SMILES. If
you had to rate the pleasantness of these molecules from 0
(extremely unpleasant) to 100 (highly pleasant), what would
you assign to each?  Respond with the selected compound
name (the one with higher pleasantness), followed by two
pleasantness values in the order the molecules are listed.
Use semicolons (;) as separators.  Do not write any comments.
```

Template – Name Prompt

```
Which molecule is likely to smell more pleasant to humans:
{compound_name_1} or {compound_name_2}?  Select only one of
the compound names.  If you had to rate the pleasantness of
these molecules from 0 (extremely unpleasant) to 100 (highly
pleasant), what would you assign to each?  Respond with the
selected compound name (the one with higher pleasantness),
followed by two pleasantness values in the order the
molecules are listed.  Use semicolons (;) as separators.  Do
not write any comments.
```

Example: 2-hydroxybenzaldehyde vs. ethyl hexanoate

Isomeric SMILES prompt:
```
Which molecule is likely to smell more pleasant to humans:
C1=CC=C(C(=C1)C=O)O or CCCCCC(=O)OCC? Select only one of the SMILES.
If you had to rate the pleasantness of these molecules from 0
(extremely unpleasant) to 100 (highly pleasant), what would you
assign to each?  Respond with the selected compound name (the one
with higher pleasantness), followed by two pleasantness values
in the order the molecules are listed.  Use semicolons (;) as
separators.  Do not write any comments.
```

Name prompt:
```
Which molecule is likely to smell more pleasant to humans:
2-hydroxybenzaldehyde or ethyl hexanoate?  Select only one of
the compound names.  If you had to rate the pleasantness of
these molecules from 0 (extremely unpleasant) to 100 (highly
pleasant), what would you assign to each?  Respond with the
selected compound name (the one with higher pleasantness), followed
by two pleasantness values in the order the molecules are listed.
Use semicolons (;) as separators.  Do not write any comments.
```

Ground-truth answer: ethyl hexanoate (pleasantness = 79.2 vs. 29.7)
Expected response format: `ethyl hexanoate;30;79`

Model predictions (compound name prompt):

Gemini 2.5 Pro	`Ethyl hexanoate; 65; 90`	✓
GPT-5 Pro	`ethyl hexanoate;55;90`	✓
o3 (high)	`ethyl hexanoate;35;85`	✓

All models correctly select ethyl hexanoate, though numerical estimates vary.

C.5 RATE-ALL-THAT-APPLY (RATA)

This multi-label classification task requires models to select all applicable odor descriptors from a comprehensive list of 138 possible descriptors ($n = 100$).

Template – Isomeric SMILES Prompt

```
Which of the following odor descriptors apply to molecule
{SMILES}?  Select all that apply from [{138 descriptors}].
Do not write any comments.
```

Template – Name Prompt

```
Which of the following odor descriptors apply to molecule
{compound_name}?  Select all that apply from [{138
descriptors}].  Do not write any comments.
```

Example: 2,3-Dimethylpentanal

Isomeric SMILES prompt:
```
Which of the following odor descriptors apply to molecule
CCC(C)C(C)C=O? Select all that apply from [alcoholic; aldehydic;
alliaceous; almond; amber; animal; anisic; apple; apricot; aromatic;
balsamic; banana; beefy; bergamot; berry; bitter; black currant;
brandy; burnt; buttery; cabbage; camphoreous; caramellic; cedar;
celery; chamomile; cheesy; cherry; chocolate; cinnamon; citrus;
clean; clove; cocoa; coconut; coffee; cognac; cooked; cooling;
cortex; coumarinic; creamy; cucumber; dairy; dry; earthy; ethereal;
fatty; fermented; fishy; floral; fresh; fruit skin; fruity; garlic;
gassy; geranium; grape; grapefruit; grassy; green; hawthorn;
hay; hazelnut; herbal; honey; hyacinth; jasmin; juicy; ketonic;
lactonic; lavender; leafy; leathery; lemon; lily; malty; meaty;
medicinal; melon; metallic; milky; mint; muguet; mushroom; musk;
musty; natural; nutty; odorless; oily; onion; orange; orangeflower;
orris; ozone; peach; pear; phenolic; pine; pineapple; plum; popcorn;
potato; powdery; pungent; radish; raspberry; ripe; roasted; rose;
rummy; sandalwood; savory; sharp; smoky; soapy; solvent; sour;
spicy; strawberry; sulfurous; sweaty; sweet; tea; terpenic; tobacco;
tomato; tropical; vanilla; vegetable; vetiver; violet; warm; waxy;
weedy; winey; woody].  Do not write any comments.
```

Name prompt:
`Which of the following odor descriptors apply to molecule 2,3-Dimethylpentanal? Select all that apply from [...same 138 descriptors ...]. Do not write any comments.`

Ground-truth answer: ethereal; green
Expected response format: `ethereal;green`

Model predictions (compound name prompt):

Gemini 2.5 Pro	`aldehydic; chocolate; citrus; cocoa; cortex; fatty; green; malty; oily; woody`	∼
GPT-5 Pro	`aldehydic; citrus; fresh; fruity; green; fatty; oily; apple`	∼
o3 (high)	`aldehydic; burnt; chocolate; cocoa; malty; nutty; roasted`	✗

Gemini and GPT-5 Pro partially overlap with ground truth (green), but also predict many false positives. o3 misses both ground-truth descriptors entirely.

C.6 Odor Similarity of Mixtures

This task evaluates models' ability to judge perceptual similarity between two odor mixtures containing 2–10 molecules each ($n = 100$).

Template – Isomeric SMILES Prompt

```
Mixture A contains molecules [{SMILES_list_A}] and mixture B
contains [{SMILES_list_B}]. On the scale [Strongly Similar;
Slightly Similar; Slightly Dissimilar; Strongly Dissimilar],
select how similar these mixtures smell. If you had to rate
the olfactory perceptual distance on a 0.0-1.0 scale (0.0 =
identical, 1.0 = completely different), what distance do you
assign? Respond with your selection from [Strongly Similar;
Slightly Similar; Slightly Dissimilar; Strongly Dissimilar],
followed by the distance value with two-decimal precision.
Use semicolons (;) as separators. Do not write any comments.
```

Template – Name Prompt

```
Mixture A contains molecules [{name_list_A}] and mixture B
contains [{name_list_B}]. On the scale [Strongly Similar;
Slightly Similar; Slightly Dissimilar; Strongly Dissimilar],
select how similar these mixtures smell. If you had to rate
the olfactory perceptual distance on a 0.0-1.0 scale (0.0 =
identical, 1.0 = completely different), what distance do you
assign? Respond with your selection from [Strongly Similar;
Slightly Similar; Slightly Dissimilar; Strongly Dissimilar],
followed by the distance value with two-decimal precision.
Use semicolons (;) as separators. Do not write any comments.
```

Example: 4-molecule mixtures (Strongly Similar)

Mixture A (4 molecules):
Names: 4-propan-2-ylbenzaldehyde; 2-ethylpyrazine; 3,5,5-trimethylcyclohex-2-en-1-one; toluene
SMILES: `CC(C)C1=CC=C(C=C1)C=O; CCC1=NC=CN=C1; CC1=CC(=O)CC(C1)(C)C; CC1=CC=CC=C1`

Mixture B (4 molecules):
Names: 1-phenylethanone; benzaldehyde; [(1S,2R,4S)-1,7,7-trimethyl-2-bicyclo[2.2.1]heptanyl] acetate; methyl 2-aminobenzoate

SMILES: `CC(=O)C1=CC=CC=C1; C1=CC=C(C=C1)C=O; CC(=O)OC1CC2CCC1(C2(C)C)C; COC(=O)C1=CC=CC=C1N`

Name prompt:
```
Mixture A contains molecules [4-propan-2-ylbenzaldehyde;
2-ethylpyrazine; 3,5,5-trimethylcyclohex-2-en-1-one; toluene]
and mixture B contains [1-phenylethanone; benzaldehyde;
[(1S,2R,4S)-1,7,7-trimethyl-2-bicyclo[2.2.1]heptanyl] acetate;
methyl 2-aminobenzoate]. On the scale [Strongly Similar; Slightly
Similar; Slightly Dissimilar; Strongly Dissimilar], select how
similar these mixtures smell. [...]
```

Ground-truth answer: Strongly Similar (experimental perceptual distance = 0.37)

Model predictions (compound name prompt):

Gemini 2.5 Pro	`Strongly Dissimilar;0.85`	✗
GPT-5 Pro	`Strongly Dissimilar;0.80`	✗
o3 (high)	`Slightly Similar;0.42`	∼

Despite the ground truth being "Strongly Similar," most models predict dissimilarity, illustrating the systematic failure mode discussed in Appendix F.1.

C.7 Olfactory Receptor Activation

This multi-label task tests whether models can identify which human olfactory receptors are activated by a given odorant molecule ($n = 80$). Each question provides 4–10 candidate receptor identifiers.

Template – Isomeric SMILES Prompt

```
Among the following [{OR_options}] olfactory receptor gene
IDs choose all olfactory receptors that molecule [{SMILES}]
activates. Do not write any comments.
```

Template – Name Prompt

```
Among the following [{OR_options}] olfactory receptor
gene IDs choose all olfactory receptors that molecule
[{compound_name}] activates. Do not write any comments.
```

Example: helional

Isomeric SMILES prompt:
```
Among the following [hOR1A2;nan;hOR3A1_R125Q;hOR1A1;hOR1D2;hOR1G1;
hOR52D1] olfactory receptor gene IDs choose all olfactory receptors
that molecule [CC(CC1=CC2=C(C=C1)OCO2)C=O] activates. Do not write
any comments.
```

Name prompt:
```
Among the following [hOR1A2;nan;hOR3A1_R125Q;hOR1A1;hOR1D2;hOR1G1;
hOR52D1] olfactory receptor gene IDs choose all olfactory receptors
that molecule [helional] activates. Do not write any comments.
```

Ground-truth answer: hOR1A2; hOR1A1; hOR52D1
Expected response format: `hOR1A2;hOR1A1;hOR52D1`

Model predictions (compound name prompt):

Gemini 2.5 Pro	`hOR1A1`	∼
GPT-5 Pro	`hOR1A2;hOR1A1`	∼
o3 (high)	`hOR1A2;hOR1A1;hOR1D2`	∼

All models identify some correct receptors but none predicts the full set; hOR52D1 is missed by all three.

C.8 Smell Identification Test

This task presents a mixture of molecules that constitute the aroma of a real-world food or object, and asks models to identify the source from four options ($n = 30$). The test covers 30 real-world odor sources: mango, peanut, hazelnut, tomato, apple, walnut, raspberry, peach, honey, parsley, grapefruit, pineapple, strawberry, apricot, rice, grape, popcorn, orange, cheese, melon, leather, chocolate, coffee, onion, fish, beer, whisky, red wine, prawn, and bread. The number of constituent molecules per item ranges from 1 (onion) to 88 (chocolate).

Template – Isomeric SMILES Prompt

```
Given the molecules {SMILES_list}, select exactly one item
from {options} whose odor profile most likely includes these
molecules.  Reply with the item name only.  Do not write any
comments.  DO NOT SEARCH ONLINE
```

Template – Name Prompt

```
Given the molecules {compound_name_list}, select exactly one
item from {options} whose odor profile most likely includes
these molecules.  Reply with the item name only.  Do not
write any comments.
```

Example: Melon (9 molecules)

Isomeric SMILES prompt:

```
Given the molecules CCC(C)C(=O)OC, CCCC(=O)OCC, CCCCCC/C=C/C=O,
CC/C=C\CC/C=C/C=O, CC/C=C\CC(=O)C=C, CCCCCC=O, CCC/C=C/C=O,
CC/C=C\CC=O, CCOC(=O)C(C)C, select exactly one item from
cheese;melon;bread;strawberry whose odor profile most likely
includes these molecules.  Reply with the item name only.  Do not
write any comments.  DO NOT SEARCH ONLINE
```

Name prompt:

```
Given the molecules methyl 2-methylbutanoate, ethyl butanoate,
(E)-2-nonenal, (E,Z)-2,6-nonadienal, (Z)-1,5-octadien-3-one,
hexanal, (E)-2-hexenal, (Z)-3-hexenal, ethyl 2-methylpropanoate,
select exactly one item from cheese;melon;bread;strawberry whose
odor profile most likely includes these molecules.  Reply with the
item name only.  Do not write any comments.
```

Ground-truth answer: melon

Model predictions (compound name prompt):

Gemini 2.5 Pro	`melon`	✓
GPT-5 Pro	`melon`	✓
o3 (high)	`melon`	✓

All three models correctly identify melon from compound names.

D Answer Extraction

All model responses are converted from free-form text into structured predictions using a two-stage pipeline: a universal tokenisation step shared across tasks, followed by task-specific interpretation rules. The full extraction code is released with the benchmark.

D.1 Universal Tokenisation

Every model response undergoes the same preprocessing regardless of task category:

1. **Splitting.** The raw response string is split on semicolons (`;`), newlines, tabs, non-numeric commas, the connective "and", and dash-separated phrases (`[;\n\r\t]+`, `,(?!\d)`, `\s+-\s+`, and `\s+and\s+`). Bullet-point prefixes (`-`, `*`, or numbered list markers) are stripped before splitting.
2. **Normalisation.** Each resulting token is lowercased; leading/trailing brackets, quotes, punctuation, and whitespace are removed; and internal whitespace is collapsed.
3. **Filtering.** Tokens that are purely numeric, empty, or begin with the prefix `desc_count` (an internal annotation artefact) are discarded. If splitting yields no valid tokens, the entire (normalised) response is treated as a single token.
4. **Empty / refusal handling.** Responses that are empty, NaN, or equal to the sentinel strings `"nan"`, `"none"`, or `"null"` yield an empty token set and are scored as incorrect ($s_i = 0$) for categorical tasks and as missing for continuous-rating tasks.

D.2 Task-Specific Interpretation

The universal token set is then interpreted per task:

- **Binary classification (OC).** The token set is checked for the presence of "odorous" or "odorless"; if either matches a ground-truth token, the question is scored correct via any-overlap (Section 4).
- **Multiple choice (OPD, SIT).** The token set is matched against the provided answer options. Success requires extracted token to overlap with the ground-truth option.
- **Compound selection with ratings (OIn, OPl).** Semicolon-separated responses are split as above; the first non-numeric token is the compound selection (scored via any-overlap), while the last two numeric values are extracted as paired ratings for the two stimuli in each question. These numeric predictions are used for Pearson correlation analysis (Figure 3a).
- **Multi-label selection (RATA, ORA).** All extracted tokens are matched against the valid option set; the resulting predicted label set $\hat{Y}_i$ is scored against the ground-truth set Y_i using per-question multilabel F1 (Section 4).
- **Mixture similarity (OS).** The categorical selection (Strongly Similar, Slightly Similar, Slightly Dissimilar, Strongly Dissimilar) is extracted for any-overlap scoring. A numerical distance value is extracted using a fallback chain: (i) first number after an equals sign, (ii) first number after a colon, (iii) last number in the response. This value is used for Pearson correlation analysis.

D.3 Multilingual Answer Extraction

For the multilingual RATA evaluation (Section 5), the extraction pipeline is extended to handle non-Latin scripts:

1. **Unicode transliteration.** Full-width and language-specific delimiters are mapped to ASCII equivalents before splitting.
2. **Fuzzy label matching.** If direct token matching against the translated option set yields no hits, a substring search is performed: each option (sorted longest-first to avoid partial prefix collisions) is sought within the normalised response text. For options containing ASCII characters, word-boundary matching is applied; for non-ASCII options (e.g., CJK characters), simple substring containment is used.

This two-level matching ensures robust extraction across languages with diverse punctuation conventions and tokenisation properties.

E Detailed Performance Analysis

This appendix provides detailed analyses of question difficulty and task-specific performance, supporting Section 5.

E.1 Single-Label Task Difficulty

Figure 5 shows the distribution of question difficulty for the six single-label tasks (830 of 1,010 questions; the multi-label tasks RATA and ORA are analyzed separately in Section E.2), measured as the percentage of the 21 evaluated models answering correctly (compound name prompts). Of these 830 questions, 390 (47.0%) are solved by every model and 113 (13.6%) by none, indicating that nearly half the single-label benchmark is saturated while a substantial tail remains universally unsolved.

OC clusters above 80%, with 116 of 175 questions solved by every model. OPl and OIn are strongly bimodal: OPl places 104 questions at 100% yet 25 at 0%, and OIn places 81 at 100% yet 27 at 0%, indicating that paired-comparison items are either trivially easy or universally hard with little middle ground. OS concentrates near 0%, with 43 of its 100 questions (43%) unsolved by any model and only 5 solved by all. SIT clusters in the upper range despite its "hard" designation.

The universally unsolved questions are not uniformly distributed across tasks: OS accounts for 43, OIn and OPl contribute 27 and 25 respectively (compound pairs where every model systematically selects the wrong molecule), and OPD contributes 14 disproportionately involving rare descriptors (Powdery, Amber, Animal-like, Tobacco-like). OC contributes only 3 and SIT only 1.

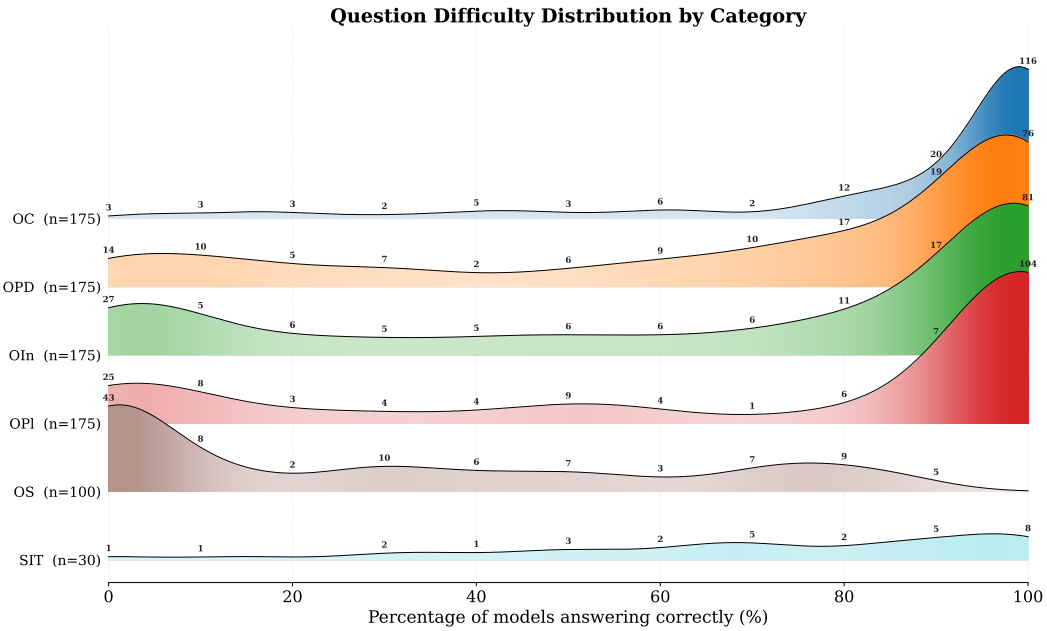

Figure 5: **Question difficulty distribution.** Each ridge shows the density of questions at a given difficulty level, measured as the percentage of the 21 evaluated models answering correctly (compound name prompts); ridge heights are proportional to the number of questions per category and annotated numbers indicate bin counts. OC clusters at the right (easy), OS at the left (hard), and OIn/OPl exhibit bimodal distributions with mass at both extremes.

E.2 Multi-Label Task Difficulty (RATA and ORA)

Figure 6 provides a question-level view of the two multi-label tasks. For RATA, the F1 distribution across the top four models is roughly bell-shaped, peaking in the 0.4–0.6 range, with 15–18 questions per model scoring $F1 = 0$ and fewer than 5 reaching $F1 > 0.8$. ORA exhibits a bimodal distribution: questions cluster either at $F1 = 0$ ($\sim$20–25 questions) or above 0.6 ($\sim$25–30 questions), with a gap in the 0.2–0.4 range.

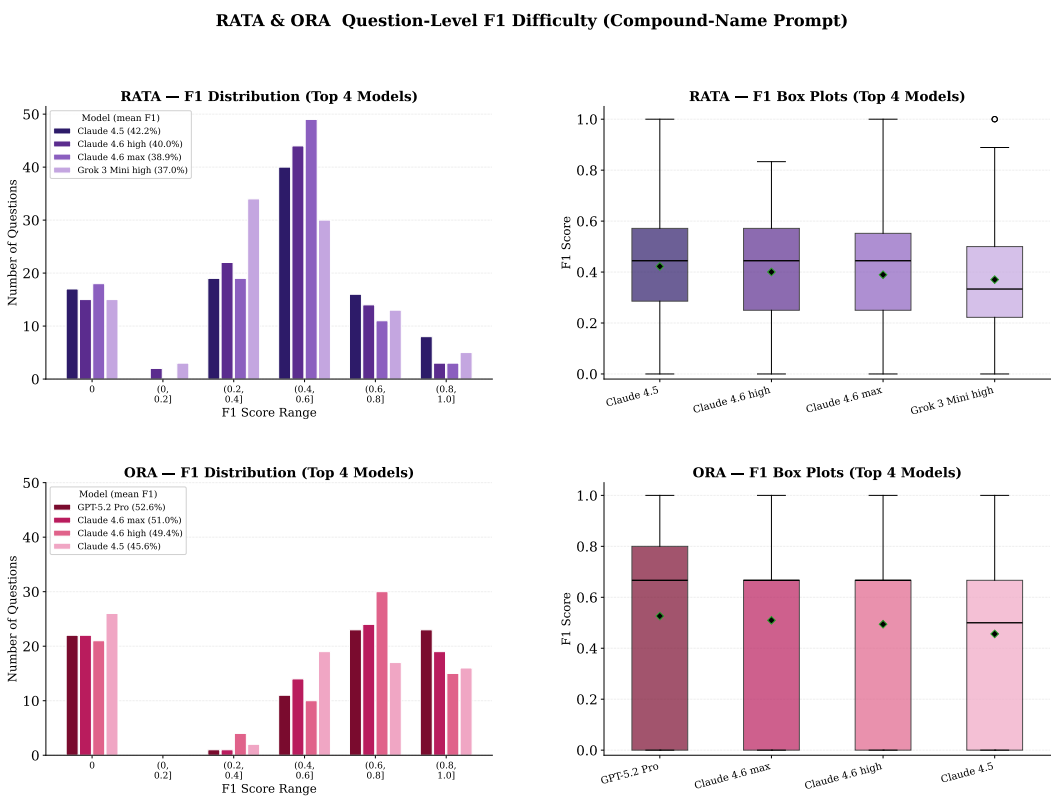

Figure 6: **RATA and ORA question-level F1 difficulty** (compound name prompts, top 4 models per task). **Left:** Histograms of per-question F1 scores. RATA shows a roughly bell-shaped distribution peaking at 0.4–0.6, while ORA is bimodal with clusters at F1 = 0 and F1 > 0.6. **Right:** Box plots summarize the F1 distributions; diamonds indicate means. The bimodal ORA pattern indicates that models either possess the relevant receptor–ligand knowledge or lack it entirely.

The best RATA model (Claude Opus 4.5, 42.2%) and best ORA model (GPT-5.2 Pro, 52.6%) represent different families, suggesting that receptor biology and semantic descriptor knowledge draw on partially independent capabilities.

E.3 Per-Label Difficulty and the "Spicy" Case Study

Among RATA descriptors, *sulfurous* achieves the highest mean F1 (≈0.55), mapping reliably to sulfur-containing functional groups. *Floral* and *fruity* also rank highly, benefiting from identifiable structural motifs. At the opposite extreme, *spicy* (mean F1 < 0.10), *fresh*, and *tropical* prove nearly impossible. By analysing the "spicy" case (9 ground-truth questions), 11 of 21 models achieve 0% recall. In 4 out of 5 examined reasoning traces, the word "spicy" never appears as a candidate. The two successful predictions occur only when the compound name enables associative retrieval (e.g., 1,2-dihydroperillaldehyde → perillaldehyde → cumin-like spiciness). This pattern generalizes: descriptors that depend on holistic molecular shape (spicy, musty, creamy, fermented) are systematically underrepresented in model predictions, which default to high-frequency alternatives (sweet, floral, green).

F Systematic Failure Analysis

This appendix presents detailed failure analyses for the two hardest task categories (OS and ORA), including reasoning traces, prediction biases, and case studies.

F.1 Odor Similarity: Prediction Bias and Molecular Overlap

All models exhibit a systematic bias toward predicting dissimilarity. Figure 7 shows that models overwhelmingly select "Slightly Dissimilar" or "Dissimilar" regardless of the ground-truth label: Claude models assign "Slightly Dissimilar" to 77–84 of 100 mixtures, while other families show similar but less extreme patterns. "Strongly Similar" is almost never predicted, even for the 25 ground-truth "Strongly Similar" pairs. This is consistent with models analyzing each molecule independently and finding surface-level differences, rather than integrating perceptual features into a holistic similarity judgment.

Table 4 provides a per-subcategory breakdown. No model exceeds chance (25%) on the combined Similar categories. DeepSeek Reasoner (8K) achieves the highest overall accuracy (35%) not through better perceptual modeling but by being more willing to predict "Strongly Dissimilar" (27 predictions vs. Claude's 1–5), which happens to capture more of the dissimilar ground truths.

The molecular overlap heuristic further explains this failure. Figure 8 plots accuracy against the number of shared molecules between mixtures: accuracy reaches ∼85% when mixtures share 9 molecules but drops to near 0% when similar mixtures share 0–2. Crucially, perceptual similarity and molecular overlap are poorly correlated in the ground truth: 9 of 25 "Strongly Similar" pairs (36%) share zero molecules. Reasoning traces confirm that models enumerate shared and unshared molecules as their primary strategy (Figure 9), a heuristic that systematically biases toward dissimilarity predictions when mixtures have low molecular overlap.

Table 4: Odor Similarity (OS) per-subcategory accuracy using compound name prompts. The ground truth is balanced with 25 questions per category (100 total). SS = Strongly Similar, SlS = Slightly Similar, SlD = Slightly Dissimilar, SD = Strongly Dissimilar. Combined Similar accuracy is computed as (SS + SlS correct) / 50. Chance level is 25% (random selection among four equiprobable categories). No model exceeds chance on the combined Similar categories.

	Similar			Dissimilar		
Model	SS (/25)	SlS (/25)	**Combined**	SlD (/25)	SD (/25)	**Overall**
Closed-Source						
GPT-5 (high)	2	2	4/50 (8%)	14	16	34%
GPT-5 (low)	2	5	7/50 (14%)	12	9	28%
GPT-5 Pro	2	2	4/50 (8%)	9	16	29%
GPT-5.2 Pro	2	1	3/50 (6%)	10	15	28%
GPT-OSS-120B	2	7	9/50 (18%)	20	5	34%
o3 (high)	1	2	3/50 (6%)	16	13	32%
o4-mini (high)	1	6	7/50 (14%)	14	11	32%
Gemini 2.5 Pro (16K)	3	0	3/50 (6%)	11	17	31%
Gemini 2.5 Pro (8K)	3	2	5/50 (10%)	7	17	29%
Gemini 2.5 Pro (32K)	2	1	3/50 (6%)	7	17	27%
Grok 3 Mini (low)	2	9	11/50 (22%)	6	1	18%
Grok 3 Mini (high)	2	2	4/50 (8%)	15	3	22%
Grok 4.1 Fast	2	5	7/50 (14%)	18	8	33%
Claude Sonnet 4.5	2	2	4/50 (8%)	20	5	29%
Claude Opus 4.5	2	1	3/50 (6%)	22	0	25%
Claude Opus 4.6 (high)	2	0	2/50 (4%)	21	3	26%
Claude Opus 4.6 (max)	2	1	3/50 (6%)	19	4	26%
Open-Source						
DeepSeek Reasoner (8K)	2	1	3/50 (6%)	21	11	**35%**
DeepSeek Reasoner (16K)	1	1	2/50 (4%)	16	7	25%
DeepSeek Reasoner (32K)	2	2	4/50 (8%)	18	10	32%
Llama 3.3 70B	0	6	6/50 (12%)	19	4	29%

Odor similarity case study. Figure 9 presents a representative failure where the ground truth is "Strongly Similar" (perceptual distance = 0.23) despite only 15% molecular overlap. Claude Opus 4.6

predicts "Slightly Dissimilar" (distance = 0.58–0.62) on both prompt formats. The reasoning trace reveals three key errors: the model assumes low molecular overlap implies perceptual dissimilarity; it overweights the sulfurous compound in Mixture B as a "significant character driver," whereas human perception integrates this differently; and it analyzes molecules individually rather than predicting the emergent perceptual characteristics.

F.2 Olfactory Receptor Activation: The D296N Knowledge Gap

For ORA, the wildtype `hOR2W1` (30 of 80 questions) is the most reliably predicted receptor, with several models achieving F1 above 0.6. The `M81V` variant is moderately well-predicted. However, `hOR2W1_D296N` (24 questions) and `hOR52D1` (5 questions) occupy the bottom of the per-label ranking, with most models near zero. Inter-model variance is substantially higher for receptor labels than for semantic descriptors, indicating that receptor knowledge is more idiosyncratic across model families than descriptor knowledge.

Claude models never predict `hOR2W1_D296N` across all 80 ORA questions (0/24 ground-truth appearances), despite this receptor being activated by numerous compounds in the M2OR database Lalis et al. (2024). Claude's reasoning consistently invokes a loss-of-function hypothesis, assuming the D296N substitution "disrupts a critical DRY motif-like region" or "eliminates activation," and reinforces this assumption each time without revision. In contrast, GPT-5.2 Pro correctly predicts `hOR2W1_D296N` in 19 of 24 cases. Remarkably, GPT-5.2 Pro's reasoning traces reveal the *same* initial misconception about D296N, but the model self-corrects during chain-of-thought, challenging its own structural assumptions and ultimately including D296N. This single knowledge gap accounts for approximately 19 missed true positives and largely explains GPT-5.2 Pro's ORA advantage (52.6% vs. Claude Opus 4.6 max's 51.0%). Both model families correctly recognize the wildtype and `M81V` variant, indicating that the `D296N` gap is a specific factual error rather than general unfamiliarity with the receptor family.

The `hOR52D1` gap likely reflects data scarcity: with only 5 ground-truth appearances, this receptor may be too rare in training corpora for any model to have learned its ligand profile. Figure 10 illustrates the D296N failure mode on a representative question.

F.3 Safety Alignment and Benchmark Accuracy

An unexpected finding concerns the interaction between safety alignment and benchmark accuracy. Claude Opus 4.6 refuses to answer questions about certain hazardous compounds. On odor classification, Claude 4.6 refuses both isomeric SMILES and compound name prompts for Tabun (nerve agent GA). For ethyl phosphonothioic dichloride (chemical weapons precursor), Claude 4.6 answers correctly from isomeric SMILES but refuses when given the compound name, suggesting the name triggers the safety filter while the isomeric SMILES notation does not. Claude Opus 4.6 (max) additionally refuses Thiofanox (organophosphate pesticide). Claude Opus 4.5 shows zero safety refusals on these compounds, indicating the filter was strengthened in the 4.6 update. No other model family (GPT, Gemini, Grok, DeepSeek, Llama) refuses any of these questions; all correctly answer "Odorous." These refusals cost Claude Opus 4.6 (max) two correct answers on odor classification.

While the safety considerations are understandable, "Does Tabun have a detectable odor?" represents legitimate toxicology knowledge: nerve agent detection is critical for protective equipment design and exposure assessment. This finding illustrates a tension between safety alignment and scientific utility that warrants consideration in benchmark design and model deployment. These findings are similar to ChemBench Mirza et al. (2025), where safety filters similarly reduced model performance on toxicity-related chemistry questions.

Additionally, Claude Sonnet 4.5 exhibits 28 epistemic refusals on olfactory receptor activation, responding "I cannot determine which receptors…" rather than providing predictions. While honest, this conservative behavior reduces its ORA score to 38.2% compared to Claude Opus 4.6 (max)'s 51.0%.

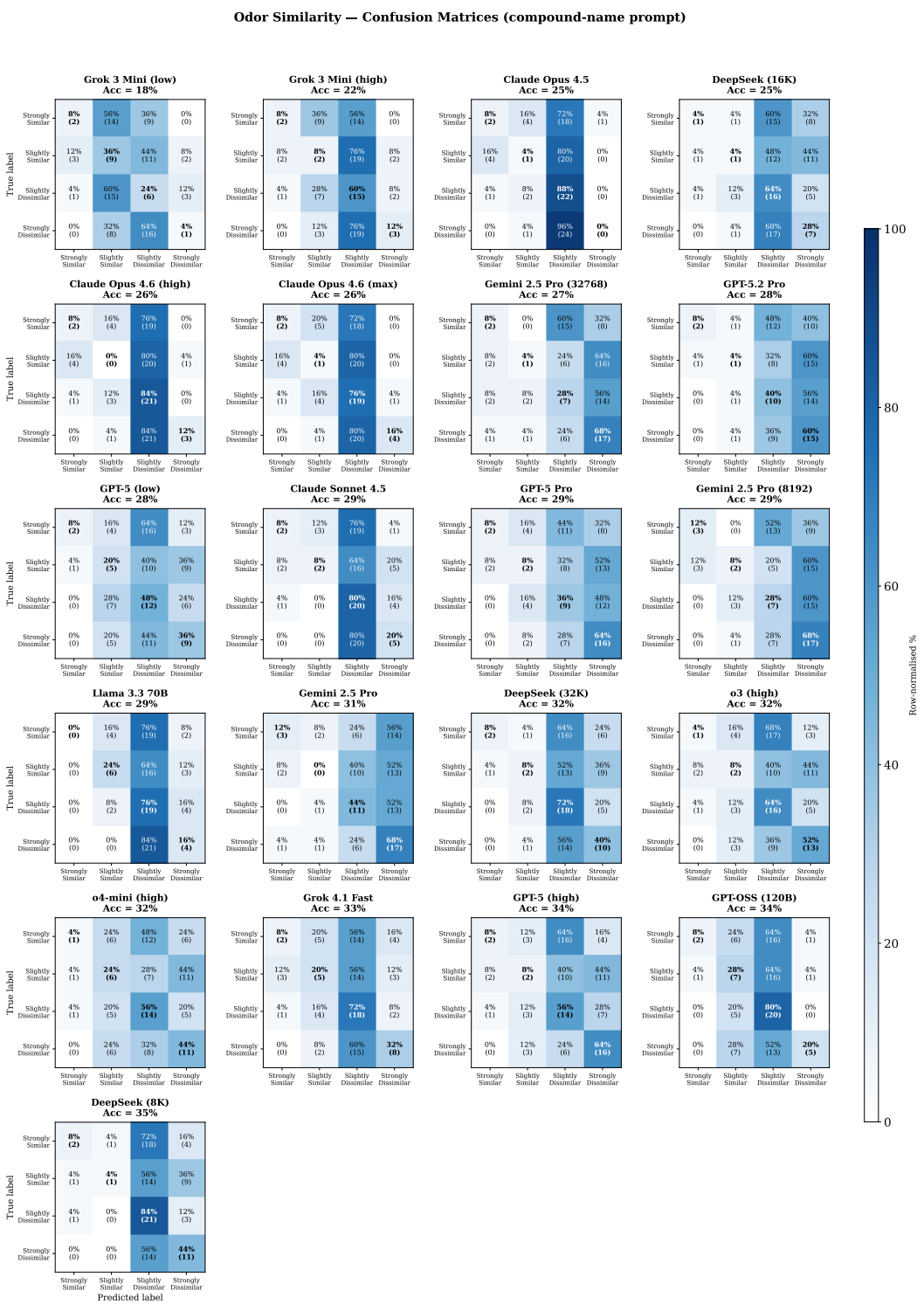

Figure 7: **Confusion matrices for odor similarity predictions** (compound name prompts). Each matrix shows row-normalized percentages (true label → predicted label). Models are ordered by overall accuracy (18–35%). Claude models exhibit extreme "Slightly Dissimilar" bias: Claude Opus 4.5 predicts this category for 72–96% of questions regardless of ground truth (third column). In contrast, DeepSeek (8K) achieves the highest accuracy by distributing predictions more broadly, including 44% correct on "Strongly Dissimilar." No model exceeds 12% accuracy on either Similar category (top two rows, diagonal cells), indicating that predicting perceptual similarity remains beyond current LLM capabilities.

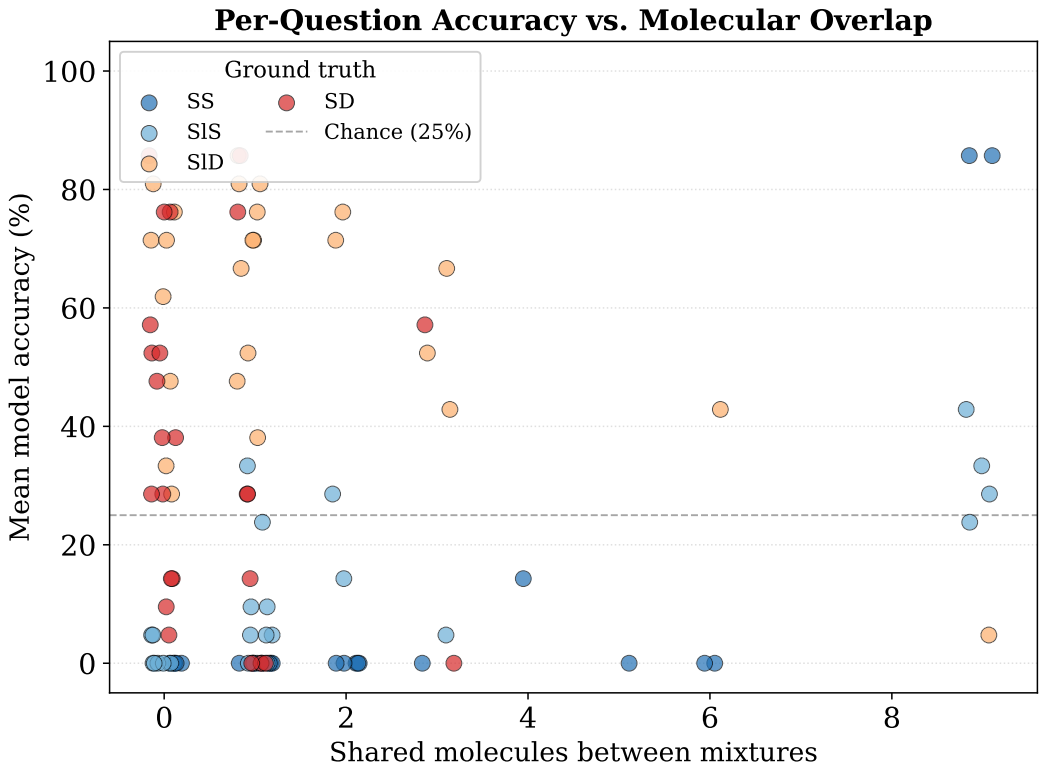

Figure 8: **Per-question model accuracy versus molecular overlap for Odor Similarity.** Each point represents one of 100 mixture pairs, colored by ground-truth category (SS = Strongly Similar, SlS = Slightly Similar, SlD = Slightly Dissimilar, SD = Strongly Dissimilar). Dashed line indicates chance (25%). Models achieve above-chance accuracy on "Strongly Similar" only when mixtures share $\geq$9 molecules (top right). For mixtures with 0–2 shared molecules, accuracy on Similar categories (blue points) clusters near 0%, while Dissimilar categories (orange/red) achieve 30–80%. This reveals that models use molecular overlap as a proxy for perceptual similarity; they cannot recognize that perceptually similar mixtures may share few or no molecules.

Case Study: Odor Similarity Failure Mode

Prompt (Compound Name)

Mixture A contains molecules [4-hydroxybenzaldehyde; 6-nonyloxan-2-one; 1-naphthalen-2-ylethanone; 2,3-di(butanoyloxy)propyl butanoate; 3-methylbutyl (E)-3-phenylprop-2-enoate; [(Z)-hex-3-enyl] 2-hydroxypropanoate] and mixture B contains [1-(3,5,5,6,8,8-hexamethyl-6,7-dihydronaphthalen-2-yl)ethanone; phenylmethoxymethylbenzene; 1-naphthalen-2-ylethanone; ethyl 3-hydroxyhexanoate; 1,3-diphenylpropan-2-one; [(Z)-hex-3-enyl] 2-phenylacetate; [(Z)-hex-3-enyl] 2-hydroxypropanoate; 2-methoxy-4-methyl-1-propan-2-ylbenzene; (benzyldisulfanyl)methylbenzene].

On the scale [`Strongly Similar; Slightly Similar; Slightly Dissimilar; Strongly Dissimilar`], select how similar these mixtures smell. If you had to rate the olfactory perceptual distance on a 0.00–1.00 scale (0.00 = identical, 1.00 = completely different), what distance do you assign?

Respond with your selection from [`Strongly Similar; Slightly Similar; Slightly Dissimilar; Strongly Dissimilar`], followed by the distance value with two-decimal precision. Use semicolons (;) as separators. Do not write any comments.

Ground Truth

Answer: Strongly Similar **Experimental Perceptual Distance:** 0.23

Note: Mixtures share only 2 of 13 unique molecules (15% overlap), yet are perceptually "Strongly Similar", different molecules produce similar odor profiles.

Model Prediction (Claude Opus 4.6)

Prompt Format	Prediction	Predicted Distance
Isomeric SMILES notation	Slightly Dissimilar ✗	0.58
Compound names	Slightly Dissimilar ✗	0.62

Model Reasoning (Claude 4.6)

*"Both mixtures share the orange blossom and floral notes from 2-acetonaphthone, plus the green fruity profile from cis-3-hexenyl lactate. Mixture A leans toward creamy, buttery sweetness with vanilla undertones, while Mixture B introduces musky amber warmth and herbal-spicy elements, with an unusual sulfurous note that creates a distinct character difference... The 2 shared components represent roughly 15% of the combined unique ingredients, and the sulfurous compound in B functions as a significant character driver that further distinguishes it from A's sweeter profile. This points to a **slightly dissimilar** rating with a distance estimate closer to 0.62."*

Failure Analysis

The model correctly identifies shared molecules and individual odor characteristics, yet predicts "Slightly Dissimilar" (distance 0.62) when ground truth is "Strongly Similar" (distance 0.23). Key errors:

- **Molecular overlap fallacy**: Assumes low molecular overlap (15%) implies perceptual dissimilarity.
- **Overweighting outliers**: Fixates on the "sulfurous note" as a "significant character driver," but human perception integrates this differently.
- **Lack of perceptual model**: Analyzes molecules individually rather than predicting emergent perceptual gestalt.

Figure 9: Representative failure case on Odor Similarity (OS). Ground truth indicates "Strongly Similar" (perceptual distance = 0.23) despite only 15% molecular overlap. Claude Opus 4.6 predicts "Slightly Dissimilar" (distance = 0.58–0.62) on both prompt formats, revealing systematic reliance on molecular composition rather than perceptual integration.

Case Study: D296N Knowledge Gap in Olfactory Receptor Activation

Prompt (Compound Name)

Among the following [hOR2W1; hOR2W1_D296N; hOR2W1_M81V; hOR5K1; hOR2M3; hOR11A1; hOR8H1; hOR5M3; hOR1A1; hOR8D1] olfactory receptor gene IDs choose all olfactory receptors that molecule [**Methyl butanoate**] activates. Do not write any comments.

Isomeric SMILES prompt: same question with [`CCCC(=O)OC`] instead of compound name.

Ground Truth

Answer: hOR2W1; hOR2W1_D296N; hOR2W1_M81V

Note: All three hOR2W1 variants—wildtype, D296N, and M81V—are activated by methyl butanoate according to the M2OR database Lalis et al. (2024). The D296N mutation does not abolish receptor function.

Model Predictions

Model	Prompt	Prediction
GPT-5.2 Pro	Isomeric SMILES	hOR2W1; hOR2W1_D296N; hOR2W1_M81V
	Compound	hOR2W1; hOR2W1_D296N; hOR2W1_M81V
Claude Opus 4.6 max	Isomeric SMILES	hOR2W1; hOR2W1_M81V
	Compound	hOR2W1; hOR1A1 ✗

Model Reasoning

Claude Opus 4.6 max correctly identifies hOR2W1 as the primary receptor but applies a loss-of-function hypothesis to D296N that it never revises, despite revisiting the question at least 10 times across both prompts:

"The D296N mutation disrupts a critical DRY motif-like region in OR2W1, which would eliminate its response to methyl butyrate, while the M81V mutation is more conservative and might preserve some function."

GPT-5.2 Pro starts with the same misconception but self-corrects during reasoning:

"D296 is quite close to the C-terminus, likely impacting G-protein coupling. The D296N mutant may not respond to ligands like methyl butyrate, so I should probably exclude it."

"Since D296 isn't a conserved residue, it might not be a classic knock-out. That leads me to think D296N could still be functional, so I'll include all three OR2W1 constructs." → **F1 = 1.00**

Failure Analysis

Both models share the same initial misconception about D296N, yet arrive at opposite conclusions. Key errors:

- **Systematic knowledge gap**: Claude *never* predicts hOR2W1_D296N across all 80 ORA questions (0/24 ground-truth appearances), accounting for ∼19 missed true positives and largely explaining GPT-5.2 Pro's ORA advantage (52.6% vs. 51.0%).
- **Loss-of-function assumption**: Claude's reasoning consistently invokes "critical DRY motif" disruption—a plausible but incorrect inference that it reinforces each time.
- **Self-correction vs. confirmation bias**: GPT-5.2 Pro overrides its initial assumption (19/24 correct); Claude reinforces it without revision. Identical domain misconceptions lead to divergent outcomes depending on whether the model's reasoning supports self-correction.

Figure 10: Case study: D296N knowledge gap in olfactory receptor activation (ORA). GPT-5.2 Pro achieves perfect F1 on both prompts, while Claude Opus 4.6 max systematically excludes hOR2W1_D296N due to an incorrect loss-of-function assumption. Both models share the same initial misconception, but only GPT-5.2 Pro self-corrects during reasoning. Question compound: Methyl butanoate (Isomeric SMILES: `CCCC(=O)OC`).