

OlfactSpeech: Generating Odor Descriptors from Speech with LLMs

Mya Ei San^{1,2}, Manuel Aleixandre¹, Manabu Okumura^{1,2}, Takamichi Nakamoto¹

1) Institute of Integrated Research, Institute of Science Tokyo, Tokyo, Japan

2) Dept. of Information and Communications Engineering, Institute of Science Tokyo, Yokohama, Japan

ei.s.1ac5@m.isct.ac.jp

Recent studies suggest that LLMs are promising to represent olfactory experiences and that their predictions correlate, to some extent, with human perceptual judgments [1][2]. In our study, we develop a system that generates odor descriptors from users' spoken utterances describing the odors they want to experience. The system receives utterances via a microphone, converts them into the text, and finally generates a set of relevant odor descriptors. After that, these descriptors are used to reproduce the corresponding scent through an olfactory display, which allows users to physically experience their desired odors. An example illustration is as follow: **User Input Utterance:** "A sweet room with roses and ripe peaches." **Output Descriptors:** "sweet" (directly mentioned), "floral" and "fruity" (inferred from "roses" and "peaches")

To identify a robust method for generating descriptors, regardless of whether they are explicitly mentioned or must be inferred from the input utterances, we investigate the capabilities of LLMs and recent methods for leveraging LLMs, namely, Retrieval-Augmented Generation (RAG) and Chain-of-Thought (CoT) prompting. We experimented with these methods for generating descriptors using a prepared dataset organized into four difficulty levels: Easy, Medium, Medium-Hard, and Hard, with 10 sentences in each. Each dataset consists of odor-related sentences paired with ground-truth descriptor sets from a predefined list of 57 descriptors. We conduct experiments using GPT-4o and evaluate using the F1 score. Our findings suggest that few-shot RAG improves the generation of odor descriptors' performance on Easy sentences, in which descriptors are explicitly mentioned. For the more difficult categories, in which odor descriptors must be inferred indirectly from the descriptions of the input, few-shot RAG with CoT prompting yields further improvements.

Overall, our experimental results demonstrate that LLMs can generate odor descriptors to a promising extent when combined with RAG and CoT prompting approaches. However, the limited size of the dataset makes it difficult to fully exploit these approaches, as their effectiveness depends on the availability of high-quality demonstration examples. Our analysis of LLM's generated CoT reasoning also indicates that the olfactory knowledge in current LLMs remains insufficient for generating reliable perceptual evidence. To address these limitations, we are currently pursuing two directions: 1) automatically collecting odor-related explanations for odor terms from input odor sentences and incorporating them into RAG with CoT prompting; and 2) constructing a multi-agent framework in which a secondary LLM verifies the validity of the generated odor descriptor sets.

Acknowledgment

This work was supported in JST Agency Mirai program (JPMJMI25H1).

References

- [1] Kurfali, Murathan, et al. "Representations of smells: The next frontier for language models?." *Cognition* 264 (2025): 106243
- [2] Wen, Yunge, et al. "AromaGen: Interactive Generation of Rich Olfactory Experiences with Multimodal Language Models." *arXiv preprint arXiv:2604.01650* (2026).