

Generating Odor Descriptors with LLMs

Mya Ei San, Manuel Aleixandre, Manabu Okumura, Takamichi Nakamoto

Institute of Integrated Research, Institute of Science Tokyo,
Yokohama, Japan
oku@lr.first.iir.isct.ac.jp

To generate odors desired by users from their utterances, it is necessary to obtain a set of odor descriptors from the utterances. For example, from a sentence 'A freshly mown garden full of intoxicating jasmine,' we would like to get "floral," "fresh," and "green." Here, please note that the system must be able to output not only descriptors that appear explicitly in the utterance (e.g., "fresh") but also those that do not (e.g., "floral," "green"). Therefore, in this work, we develop a method for generating a set of odor descriptors from utterances using Large Language Models (LLMs). Specifically, we first investigate and evaluate whether odor descriptors can be generated by leveraging existing LLMs and related technologies, such as Retrieval-Augmented Generation (RAG) and Chain-of-Thought (CoT) prompting.

First, we prepared evaluation datasets - categorized as Easy, Medium, and Hard (with 10 sentences each) - consisting of pairs of input sentences and sets of odor descriptors, utilizing 57 odor descriptors. We conducted experiments using GPT-4o as the LLM and the F1 score - the harmonic mean of recall and precision - as the evaluation metric. We found that for the "Easy" dataset, where odor descriptors tend to explicitly appear in the input text, few-shot RAG without CoT is sufficient; in contrast, for the "Medium" and "Hard" datasets, few-shot RAG combined with CoT yields the highest performance. We also found that, when GPT-4o was used as the LLM, an F1 score of around 0.7 was achieved for the dataset as a whole.

These results demonstrate that, by leveraging recent advancements such as RAG and CoT, it is possible to generate odor descriptors using LLMs to a certain extent. At the same time, it has become clear that these technologies are not being fully utilized. For example, we have only a small amount of examples (a dataset) provided as "shots." Further, while CoT relies on LLMs themselves to provide explanations regarding odors, their knowledge of odors cannot necessarily be considered sufficient. Thus, we are now working on the following:

1. Automatically collecting odor-related explanations for words and utilizing them via RAG during Chain-of-Thought (CoT) prompting; and
2. Constructing a multi-agent framework that incorporates a secondary LLM - distinct from the one generating the odor descriptors - to verify the validity of the generated descriptor sets.