Enhancing Low-Resource Text Classification with LLM-Generated Corpora : A Case Study on Olfactory Reference Extraction

Anonymous ACL submission

Abstract

Extracting sensory information from text, particularly olfactory references, is challenging due to limited annotated datasets and the implicit, subjective nature of sensory experiences. This study investigates whether GPT-4o-generated data can complement or replace human annotations. We evaluate human- and LLM-labeled corpora on two tasks: coarse-grained detection of olfactory content and fine-grained sensory term extraction. Despite lexical variation, generated texts align well with real data in semantic and sensorimotor embedding spaces. Models trained on synthetic data perform strongly, especially in low-resource settings. Human annotations offer better recall by capturing implicit and diverse aspects of sensoriality, while GPT-4o annotations show higher precision through clearer pattern alignment. Data augmentation experiments confirm the utility of synthetic data, though trade-offs remain between label consistency and lexical diversity. These findings support using synthetic data to enhance sensory information mining when annotated data is limited.

1 Introduction

Despite the key role of sensory experiences in human communication, computational methods for detecting and interpreting olfactory—and more broadly sensory—references in text remain limited, mainly due to the lack of high-quality annotated datasets. Annotating olfactory references is challenging because they can be implicit, metaphorical, and culturally dependent. Unlike concrete categories like named entities, smell-related references are context-dependent and subjective, requiring human judgment to disambiguate.

We explore synthetic data generated by large language models (LLMs), specifically GPT-4o (OpenAI, 2023), to address these challenges. We investigate whether generated data can substitute or complement real-world datasets in sensory information mining. Focusing on olfaction, we introduce the Olfactory Synthetic Dataset (referred to as D_2 in this paper), a novel resource designed to mirror the real-world Odeuropa Corpus (D_1) (Menini et al., 2022)[1]. We release the full dataset[2], including GPT-4o and expert-annotated versions, as a contribution to the research community. The dataset generation and annotation process, and source code for experiments are fully documented for reproducibility.

We evaluate synthetic data utility across three axes: **(1) Corpus-level similarity**, assessing lexical and semantic alignment between D_1 and D_2 to gauge how closely generated texts match real sensory language; **(2) Model performance**, comparing sentence classification and sensory term extraction on both datasets to test if models trained on synthetic data perform comparably; and **(3) Data augmentation**, measuring the effect of adding synthetic examples to samples of real-world datasets.

Our study shows synthetic data can effectively support sensory information extraction, offering a scalable alternative for domains with limited annotations, notably for other sensory modalities like sound and taste. Unlike prior work, we compare models trained on LLM-labeled synthetic data with those trained on human-labeled data to assess trade-offs in sensory domain labeling methods.

The remainder of this paper is organized as follows: Section 2 reviews related work; Section 3 details dataset generation and both human and automatic labeling protocols; Section 4 describes the dataset comparison methodology; Section 5 presents our experimental results; Section 6 draws conclusions, and Section 7 discusses limitations and future directions. Additional information provided is in the appendix.

[1] https://github.com/Odeuropa/benchmarks_and_corpora
[2] https://anonymous.4open.science/r/ijcnlp_2025-DC51

2 Related Work

Recent advances in large language models (LLMs) have increased interest in data creation, annotation, and augmentation across NLP tasks. One growing line of research explores LLMs as synthetic data generators to reduce reliance on costly manual annotations, especially in specialized domains involving complex or subjective phenomena such as sensory information extraction.

Synthetic Data Generation with LLMs in NLP Several studies evaluate the potential and limits of LLM-driven synthetic data generation. Long et al. (2024) survey the expanding use of LLMs for NLP benchmark creation. Li et al. (2023) analyze synthetic data's impact on text classification, showing benefits in low-resource settings but limited semantic depth and contextual realism. Almeida and Matos (2024) explore zero-shot data generation for information retrieval, highlighting prompt design's importance.

Methodological gaps raise reproducibility concerns, including inconsistent evaluation protocols and limited expert involvement (Nafis et al., 2025; Chim et al., 2025). Overuse of synthetic examples may cause model collapse; Shumailov et al. (2024) and Seddik et al. (2024) identify thresholds where performance drops, implying stricter control over synthetic-to-real data ratios.

The trade-off between synthetic and human-labeled data is discussed by Møller et al. (2024), showing classifiers trained on human data generally outperform those using LLM-generated examples, except for rare classes where synthetic augmentation helps. Zero-shot LLMs underperform small human-trained models, underscoring expert annotation's value.

These issues are especially relevant for sensory information extraction, where LLMs exhibit lexical bias and struggle with subjective nuance, limiting their effectiveness as substitutes for human annotators (Mohta et al., 2023).

Sensory Information Mining Contextual language models have been applied to detect sensory references in text. Menini et al. (2022) used MacBERTh (Arevalo and Fonteyn) to detect olfactory information in historical texts. Rule-based systems like Massri et al. (2022) offer interpretability but lack adaptability.

Khalid and Srinivasan (2022) used BERT with Lancaster Sensorimotor Norms (Lynott et al., 2020) to predict sensory modalities as bag-of-words. Kennington (2021) integrated sensorimotor features into ELECTRA (Clark et al., 2020), while Boscher et al. (2024) combined contextual embeddings with sensorimotor representations using lexical heuristics (Mpouli et al., 2020), word embeddings, and multilingual dictionaries (Sagot and Fišer, 2012). Despite promising results, these methods struggle with subjective content without expert oversight (Zhao et al., 2023).

Real-World vs. Synthetic Data in Sensory Domains Boscher et al. (2024) compared a real-world olfactory corpus, Odeuropa (Menini et al., 2022) with a GPT-4o-generated auditory dataset, but did not analyze real vs. synthetic data for the same modality. Their auditory dataset (1,000 balanced sentences) was a proof-of-concept, not benchmarked against annotated corpora.

This reveals a broader gap: the lack of validated real-world sensory datasets and rigorous comparisons between synthetic and real data within the same modality. Current pipelines often over-rely on LLMs, have limited expert validation, and suffer from cultural or subjective generation biases.

Positioning of the Present Work This paper addresses these issues by evaluating LLM-generated olfactory datasets against real-world annotated corpora. Unlike prior work focused on modality comparison or raw generation, we perform a fine-grained analysis of lexical coverage, semantic variability, and trained model performance. Our pipeline includes expert intervention during generation and annotation and ensures reproducibility by providing prompts, annotation guidelines, and the full dataset with both human and model-based annotations for comparison.

3 Dataset Generation

We aim to generate a synthetic dataset D_2, comparable to the real-world dataset D_1—the Odeuropa Text Dataset—to evaluate whether it can support sensory information extraction tasks. D_1 consists of 2,176 sentences, with 602 (28%) labeled as olfactory by experts, featuring over 5,500 odor-related terms such as *aroma*, *scent*, or *sweet*.

This section details the synthetic generation protocol of D_2 via the GPT-4o (OpenAI, 2023) web

interface, prompting strategies, annotation procedures, and evaluation of model-human agreement at both sentence and token levels.

3.1 Generation Protocol and Prompt Design

We adapted prompting strategies from Boscher et al. (2024), initially designed for auditory data. We design two distinct prompts intended to generate syntactically and semantically diverse positive (P1) and negative (P2) sentences. The dataset generation prompts are provided in Appendix A in Table 6.

A total of 500 positive (olfactory) and 1,700 negative (non-olfactory) sentences were generated to match the class ratio of the original corpus D_1. Examples of synthetic positive and negative sentences are given in Table 1, with sensory terms in bold for positive sentences.

Table 1 Positive vs. Negative Sentences Examples

Positive	Negative
*"The aroma of **fresh-baked bread** lingered warmly."*	*"A fearless diver plumbed unexplored reefs below."*

3.2 Annotation Protocol

Each generated sentence undergoes a two-levels annotation, consistent with D_1: (1) sentence-level classification (positive/negative) and (2) token-level annotation for sensory terms in positive examples. We compare two annotation methods:

Automatic Annotation with GPT-4o (D_2^{LM}) : In the first scenario, all annotations are performed automatically by the same model used for data generation, without any human correction. Sentences generated using prompt P2 are labeled as negative, while those generated with prompt P1 are labeled as positive. Then, positive sentences are passed to the LLM, which is queried using prompt P3 (see Table 6 in Appendix Appendix A) to extract olfactory terms.

Human Expert Annotation (D_2^{EX}): In parallel, we conduct a human-guided annotation process led by a domain expert. Annotation is carried out by a research engineer specialized in digital humanities. Each sentence is first labeled as potentially positive or negative based on expert judgment. Positive sentences are then manually annotated at a token level to identify terms conveying olfactory information, whether explicitly or implicitly. In ambiguous cases, only tokens that are clearly olfactory in context are retained. Each sentence is ultimately classified as positive if it contains at least one such token, and as negative otherwise.

3.3 Human vs. Model Annotation Agreement

While GPT-4o provides scalable generation and initial annotation, manual expert validation may be necessary to ensure quality for nuanced sensory datasets like D_2. To assess annotation reliability, we evaluate the agreement between $D_2^{\text{EX}}$and $D_2^{\text{LM}}$for both sentence- and token-level labeling.

Sentence Classification: Among 1,700 negative sentences produced by GPT-4o, 596 (35%) were reclassified as positive by the expert. These typically contained implicit olfactory cues—such as references to nature or animals—highlighting GPT-4o's reliance on explicit keyword detection. For example, sentences like *"**Puppy** chased butterflies beside **flowering backyard** fence"* and *"**Blue jays** perched on **cedar branches** in spring"* were labeled as negative by GPT-4o but judged as positive by the expert due to their olfactory context, especially inferred from tokens in bold.

Moreover, we obtain a Cohen's Kappa coefficient (Cohen, 1960) κ equal to 0.52 for binary sentence classification of D_2^{EX}v.s. D_2^{LM}, confirming a moderate agreement between the two annotation methods (Landis and Koch, 1977). This suggests that while there is a fair level of consistency between the automatic and expert annotations, some divergences remain, particularly for sentences with implicit or context-dependent olfactory cues, which are more challenging for the LLM to detect.

Token-Level Annotation: We compared token-level labels across both annotations. The resulting score of $\kappa = 0.503$ confirms only moderate alignment between human- and model-based annotations, consistent with findings at the sentence level and justifying an analysis of both strategies. Appendix C discusses the sensory vocabulary divergences between both annotations.

In Section 5, we discuss how the annotation method affects D_2's similarity to D_1 and the impact on model performance when augmenting data with synthetic examples.

4 Methodology

After generating the synthetic dataset D_2, it is compared to the real-world dataset D_1 to: (1) assess linguistic and semantic similarity; (2) evaluate model

performance when trained on each; and (3) determine D_2's utility for augmentation or substitution.

4.1 Corpus Comparison

To quantify lexical and semantic similarity between D_1 and D_2, we adopt the corpus comparative metrics suggested by Møller et al. (2024):

Token Overlap: We measure similarity by computing the Jaccard similarity between each sentence $s_2 \in D_2$ and its most similar sentence $s_1 \in D_1$, based on the overlap of their token sets.

Semantic Similarity: Cosine similarity between the sentence embeddings of s_1 and s_2 is computed using (1) SentenceBERT (Reimers and Gurevych, 2019) and (2) 11-dimensional sensorimotor sentence embeddings proposed by (Boscher et al., 2024; Lynott et al., 2020). For both embedding models and for each sentence s_2, the highest similarity score with any $s_1 \in D_1$ is retained, and distributions are visualized via density plots.

4.2 Corpus Classification

We compare classification performances obtained either on D_1 or D_2, through two sensory information extraction tasks:

Task 1 — Binary Sentence Classification: Classify sentences to determine whether they contain olfactory references or not using three models: SENSE-LM (Boscher et al., 2024), vanilla BERT (Devlin et al., 2019), and Logistic Regression over sentence sensorimotor features as defined by (Lynott et al., 2020).

Task 2 — Sensory Term Extraction: Identify sensory expressions (e.g., "coffee," "tobacco") from positively labeled sentences with two considered models, BERT and SENSE-LM.

Evaluating model performance on D_1 vs. D_2 assesses whether they yield comparable scores and similar model rankings. For both tasks, macro-averaged Precision, Recall, and F1-score are computed and averaged over 10 cross-validation folds, with standard deviations reported.

4.3 Data Augmentation with Synthetic Examples

To assess the impact of synthetic data on model performance, we augment the real-world dataset D_1 with examples from D_2, evaluating both classification tasks from Section 4.2. The training set is defined as $D_{\text{train}} = D_1^{n_1} \cup D_2^{n_2}$, and the test set as $D_{\text{test}} = D_1^{n_3}$, with $n_3 = 0.2 \times |D_1|$. The total training size is $N = n_1 + n_2$, and the values of n_1 and n_2 vary by scenario. Results are reported as the average over 10 folds with standard deviation.

4.3.1 Data Augmentation with Constant Real-World Data Sample

In this setup, the number of real examples n_1 is fixed and synthetic samples are progressively added, increasing the number of artificial examples n_2. In our experiments n_1 is set to 100, while n_2 reaches 1750 in Task 1, and 400 in Task 2. This setup allows to examine how models benefit from increasing synthetic input in low-data regimes.

4.3.2 Data Augmentation a Variable Size Real-World Data Sample

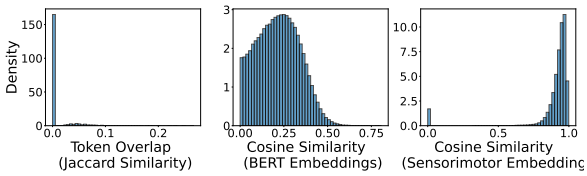
We consider an initial training dataset composed only of n_1 examples from D_1, and we gradually add synthetic examples from D_2 by augmenting a coefficient $p \in \{0, 10, \ldots, 100\}$, s.t. $n_2 = \frac{p}{100} \times n_1$ with $n_1 \in \{50, 100, 200, 500, 1000, 1750\}$ for binary sentence classification , and $n_1 \in \{50, 100, 200, 300, 400\}$ for sensory terms extraction. This setup tests how model performance evolves as the addition of synthetic data supplements real data under several initial dataset sizes.

4.3.3 Data Augmentation with Variable Ratio of Synthetic Data

In this scenario, we fix the training set size N to several values (50–1750 for binary sentence classification, 50–400 for sensory term extraction) and vary the proportion of synthetic data $p \in \{0, 10, \ldots, 100\}$, such that $n_2 = \frac{p}{100} \times N$ and $n_1 = N - n_2$. This setup evaluates how much synthetic data can replace real data without significantly affecting classifier performance, and to what extent scores remain stable or degrade as the synthetic ratio increases.

5 Evaluation

This section evaluates how synthetic corpora D_2 compare with the real-world corpus D_1 across the three axes defined in Section 4: (1) similarity between corpora, (2) model ranking consistency across datasets, and (3) efficacy of synthetic data in substituting real-world data. Text pre-processing pipelines and experimental hyper-parameters are reported in Appendix B.



(a) D_1 vs D_2 (All terms)

Figure 1: Comparison of sentence similarity distributions between positive sentences of generated (D_2) and original (D_1) corpora, using three metrics—token overlap, cosine similarity based on BERT embeddings, and cosine similarity based on sensorimotor embeddings.

5.1 Corpus Comparison

Figure 1 presents the similarity between D_1 and D_2 across all terms using the metrics from Section 4.1: token overlap (left), Sentence-BERT semantic similarity (middle), and sensorimotor similarity (right) (Boscher et al., 2024). The X-axis shows the best token overlap or cosine similarity for each generated sentence annotation compared to D_1; the Y-axis shows sentence density per metric bin.

Token overlap (left panel) is low, indicating dissimilar vocabulary. Semantic similarity (middle) is moderate, while sensory similarity (right) is higher, typically between 0.8 and 1. Despite lexical differences, due to a contextual and historical domain shift between both datasets (historical texts in D_1 and contemporary data in D_2), generated sentences exhibit shared semantic and sensorimotor features with real-world data, supporting the use of synthetic corpora for classification tasks. Extended results in Appendix D show stronger alignment when limited to positive terms.

Regarding sensory vocabulary, D_2^{LM} uses a restricted range of positive terms (318), often repeating generic words like *scent*, *aroma*, and *perfume*. In contrast, D_2^{EX} is more lexically diverse (902 unique positive terms) and aligns more closely with D_1 annotations. Extended statistical analyses and tests in Appendix C show that the distribution and ranking of positive terms in D_2^{EX} do not significantly differ from D_1, unlike D_2^{LM}.

5.2 Corpus Classification

We compare classification model performance on D_1 and D_2 for two tasks: binary sentence classification and sensory term extraction (see Section 4.2). Our goal is to assess if models perform consistently across datasets and if their ranking remains stable between real and synthetic data.

Binary Sentence Classification Table 2 shows the performance of SENSE-LM, BERT, and logistic regression evaluated using 1) D_1 (left), 2) D_2 with human annotations D_2^{EX}(center), and 3) automatic annotations D_2^{LM}(right). While models perform better on synthetic data—regardless of annotation source—model rankings remain consistent across datasets. This suggests that synthetic corpora can serve as reliable proxies for evaluating model performance rankings, even if they do not fully reflect real-world complexity.

Sensory Term Extraction As shown in Table 3, performance is slightly higher with D_2^{LM}annotations, likely due to lower lexical diversity and more homogeneous positive terms. In contrast, D_2^{EX}—with its greater term variety— provides results closer to the Odeuropa dataset (see Appendix C). Despite differences in absolute scores, model rankings remain stable, confirming that artificial datasets can be reliable substitutes for comparing models in olfactory term extraction.

5.3 Data Augmentation with Synthetic Examples

In the following, we conduct the experiments related to the protocol introduced in Section 4.3, to measure the impact of data augmentation on the utility of classifiers. The results are presented only for the SENSE-LM model in Section 5.3.2 and Section 5.3.3, and the other models are provided in Appendix E as the conclusions are similar.

5.3.1 Data Augmentation with A Constant Real-World Data Sample

Following the method described in Section 4.3.1, we first assess the impact of adding synthetic examples from D_2 to a constant base of $n_1 = 100$ real sentences from D_1. For both tasks, the performances are evaluated in terms of F1-score as a function of the amount of synthetic data added, using for ground truth either human labels D_2^{EX}or automatic annotation D_2^{LM}.

Binary Sentence Classification. Figure 2a shows the performances obtained on the binary sentence classification task. Models trained with D_2^{EX}outperform D_2^{LM}in low-resource settings ($n_2 \in [20; 150]$) for all models, with statistically significant differences according to the Student Fischer's t-test (Student, 1908) applied to the F1-Score distribution by folds for each annotation, showing p-values inferior to 0.05. However, at higher vol-

Table 2 Comparative evaluation of the binary sentence classification task performed by considered models.

Method	Odeuropa Dataset (D_1)			Olfactory Synthetic Dataset (D_2^{EX})			Olfactory Synthetic Dataset (D_2^{LM})		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score
BERT	91.51 ± 1.12	90.12 ± 0.61	90.80 ± 0.85	99.10 ± 0.90	98.93 ± 0.80	99.01 ± 0.85	98.00 ± 0.35	97.81 ± 0.28	97.90 ± 0.30
Logistic Regression	82.25 ± 1.51	72.33 ± 1.22	76.97 ± 1.36	91.00 ± 1.30	90.52 ± 1.70	90.76 ± 1.49	91.40 ± 2.10	91.09 ± 2.20	91.24 ± 2.78
SENSE-LM	**94.09 ± 0.81**	**92.26 ± 0.72**	**93.16 ± 0.76**	**99.80 ± 0.45**	**99.66 ± 0.39**	**99.73 ± 0.42**	**99.40 ± 0.65**	**99.23 ± 0.61**	**99.31 ± 0.63**

Table 3 Comparative evaluation of the sensory terms extraction task by considered models.

Method	Odeuropa Dataset (D_1)			Olfactory Synthetic Dataset (D_2^{EX})			Olfactory Synthetic Dataset (D_2^{LM})		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score
BERT	80.01 ± 2.22	66.32 ± 1.13	72.52 ± 1.68	86.19 ± 0.92	75.60 ± 1.47	80.74 ± 0.52	85.45 ± 0.84	79.01 ± 1.10	82.06 ± 0.76
SENSE-LM	**82.01 ± 1.81**	**73.62 ± 1.56**	**77.58 ± 1.65**	**86.65 ± 0.52**	**78.54 ± 0.76**	**82.37 ± 0.63**	**85.53 ± 0.62**	**81.73 ± 0.77**	**83.59 ± 0.15**

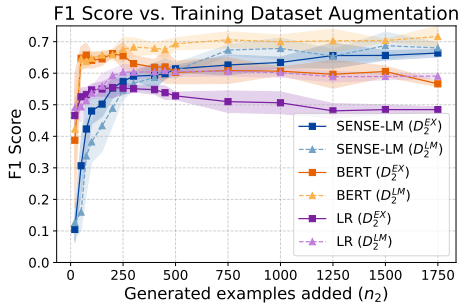
umes ($n_2 = 1750$), models trained on D_2^{LM}catch up and occasionally surpass D_2^{EX}. This highlights that annotation quality provided by experts is more impactful at a small data scale.

Sensory Term Extraction. Figure 2b shows that models trained on D_2^{EX}outperform those with $D_2^{\text{LM}}$for small synthetic additions, with significant gaps for $n_2 \in [50, 200]$. Beyond, models with D_2^{LM}gather and sometimes surpass D_2^{EX}(although non-significantly). These results support the advantage of using human labeling (D_2^{EX}) in low-resource settings and the efficiency of automatic labeling (D_2^{LM}) at a larger scale. In both cases, F1-score degrades for $n_2 \geq 100$ (over 50% synthetic data), aligning with prior work on model collapse (Seddik et al., 2024; Kazdan et al., 2024).

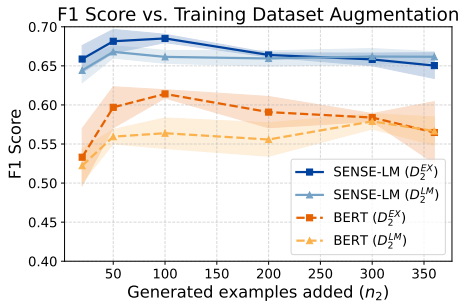
5.3.2 Data Augmentation with a Variable Size Real-World Data Sample

In Section 5.3.1, we showed that adding synthetic examples to a fixed base of 100 real Odeuropa samples improves model utility up to a threshold. Building on findings that D_2^{EX}benefits from less data while D_2^{LM}improves with more, we now test whether synthetic data augments real data across varying initial sizes. Starting with D_1 only ($N = n_1$), we progressively add synthetic examples until $N = 2 \times n_1$ to assess how augmentation scales. For both binary classification and sensory term extraction, we follow the protocol in Section 4.3.2.

Binary Sentence Classification Figures 3a and 3b show SENSE-LM's F1-Score for varying initial real-dataset sizes n_1 and two ground truths –human annotation (D_2^{EX}, left) and automatic annotation (D_2^{LM}, right)–plotted against the percentage of synthetic data supplementing the original real data (0% means training only on real data, 100% means equal amounts of generated and real data).



(a) Binary sentence classification.



(b) Sensory terms extraction.

Figure 2: F1-score of models evolution with $n_1 = 100$ examples from Odeuropa and progressive augmentation of n_2 synthetic examples.

For $n_1 \geq 500$, adding synthetic data does not improve performance. However, for smaller sizes ($n_1 \in [50; 200]$), both $D_2^{\text{LM}}$and D_2^{EX}benefit, with stronger gains for D_2^{EX}. This likely comes from D_2^{EX}'s finer capture of implicit sensory cues at the sentence level, as discussed in Section 3.3.

Sensory Terms Extraction The results for sensory terms extraction in Figures 3c and 3d show that models trained with D_2^{LM}yield higher and more consistent gains, especially for $n_1 \geq 200$, with statistically significant gaps over D_2^{EX}. As detailed in Table 4 for $n_1 = 400$ with SENSE-LM, D_2^{LM}-trained models achieve higher F1-scores by improv-

Table 4 Precision and Recall (% ± std) for D_2^{EX} and D_2^{LM} on Binary Sentence Classification (SENSE-LM, N=400) with data augmentation. D_2^{EX} values match Figure 3c, D_2^{LM} values match Figure 3d.

Gen. Data Added (%)	D_2^{EX} Prec.	D_2^{EX} Rec.	D_2^{LM} Prec.	D_2^{LM} Rec.
0	72.95 ± 3.17	64.72 ± 2.91	72.95 ± 3.17	64.72 ± 2.91
20	72.63 ± 2.37	66.76 ± 2.28	73.29 ± 2.83	63.97 ± 3.24
40	71.67 ± 0.75	68.72 ± 3.04	76.04 ± 2.81	62.02 ± 2.12
60	72.44 ± 1.36	67.63 ± 1.87	75.54 ± 2.34	64.37 ± 4.12
80	71.49 ± 1.83	67.13 ± 1.37	75.75 ± 4.13	61.73 ± 2.03
100	69.11 ± 4.68	68.66 ± 2.66	77.08 ± 3.33	61.59 ± 4.53

Table 5 Precision and Recall (% ± std) for D_2^{EX} and D_2^{LM} on Sensory Terms Extraction (SENSE-LM, N=400) with data augmentation. Values for D_2^{EX} match with Figure 4c, D_2^{LM} match Figure 4d.

Gen. Data Ratio (%)	D_2^{EX} Prec.	D_2^{EX} Rec.	D_2^{LM} Prec.	D_2^{LM} Rec.
0	74.64 ± 2.06	60.49 ± 0.85	74.64 ± 2.06	60.49 ± 0.85
20	73.82 ± 1.73	64.98 ± 2.67	72.89 ± 0.43	65.73 ± 0.39
40	67.88 ± 2.41	71.12 ± 2.42	76.67 ± 2.12	60.08 ± 0.72
60	63.33 ± 1.95	71.68 ± 1.34	77.78 ± 2.05	58.61 ± 2.51
80	57.92 ± 2.45	75.38 ± 1.72	75.80 ± 1.64	59.45 ± 1.09
100	45.08 ± 2.30	77.68 ± 1.64	76.87 ± 0.87	56.92 ± 1.41

ing precision at a slight recall cost as generated examples increase. This reflects D_2^{LM}'s pattern-guided annotation focusing on explicit, restricted vocabulary, limiting predicted terms. Conversely, D_2^{EX} improves recall through richer, more diverse annotations but reduces precision due to more false positives. Overall, D_2^{LM} offers the best precision-recall trade-off and highest F1-score.

5.3.3 Data Augmentation with a Variable Ratio of Synthetic Data

In Section 5.3.2, we saw that gradually adding generated data to a fixed real train set generally improves prediction quality. However, prior experiments did not fully assess how varying the synthetic-to-real data ratio affects performance. Therefore, following Section 4.3.3, we keep the train size constant while varying this ratio.

Binary Sentence Classification. For different dataset sizes N, Figures 4a and 4b show F1-Score results as a function of the ratio of generated data in the training set (0% = all real data; 100% = all synthetic). Performance generally degrades as synthetic data dominates, except for small N, where performance is already low. This drop is sharper for larger N. With D_2^{EX}, performance stays higher at low synthetic ratios, especially for $N = 100$ or 200. Across all sizes, models collapse beyond 60% synthetic data, approaching random classification (F1 ≈ 0.5). As the structure of D_2 diverges from D_1 in vocabulary and complexity, introducing excessive synthetic data degrades performance, highlighting the need to retain at least 40% real data to prevent model collapse.

Sensory Term Extraction. As shown in Figures 4c and 4d, performance remains stable when generated data stays below 40%. Beyond this, D_2^{EX} causes sharp F1 drops—up to 12 points for $N = 400$ while D_2^{LM} degrades from a ratio of 20%, but only by 2–3 points. Table 5 details precision and recall at $N = 400$: with D_2^{EX}, Recall improves but Precision drops; D_2^{LM} shows the reverse. These trends support our hypotheses in Section 5.3.2: though D_2^{LM} often yields better F1, each annotation shows advantages depending on the objective.

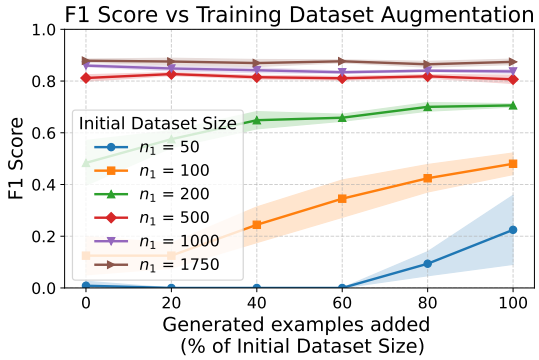
6 Conclusions and Perspectives

We explored using synthetic data from LLMs to support olfactory information extraction, a domain challenged by subjective sensory experiences. We introduced the synthetic dataset D_2, generated with GPT-4o, and compared it to the real Odeuropa corpus (D_1). Our analysis covered lexical and semantic similarity, classification and extraction performance, and data augmentation with GPT-based (D_2^{LM}) and expert-curated (D_2^{EX}) annotations.

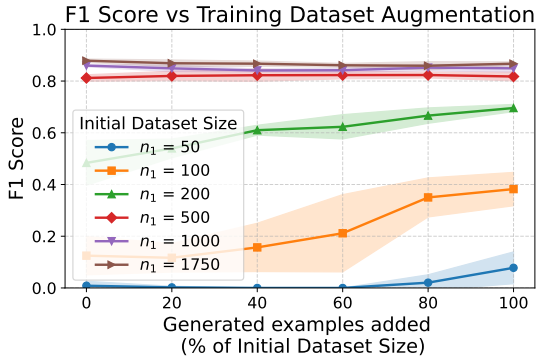
Despite lexical differences, D_1 and D_2 align in sensorimotor and semantic space. Models behaved similarly across datasets in F1-score and ranking.

Synthetic data improved performance, especially in low-resource settings. D_2^{LM}-trained models sometimes outperformed D_2^{EX} using consistent pattern-based labeling, boosting precision by reducing false positives. D_2^{EX}'s human annotations capture finer nuances and broader vocabulary, improving recall. This trade-off suggests LLM-based annotations are convenient for precision tasks, while human annotations offer advantages for recall-oriented applications.

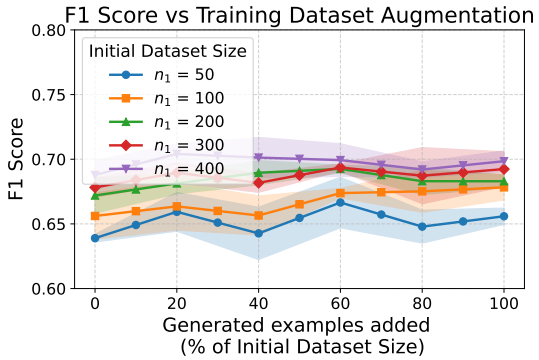
Future work includes enhancing data realism via prompt diversification and seed-based generation, and evaluating impacts of realism and subjectivity on performance. We may explore applications to other subjective tasks, beyond sensoriality.



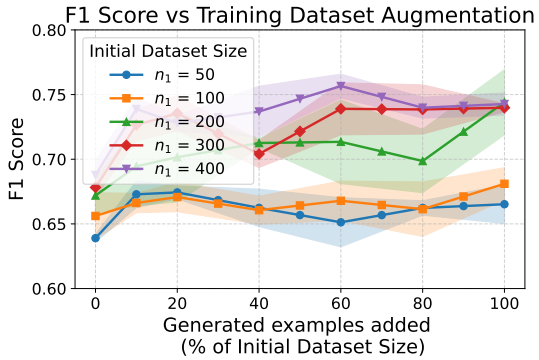
(a) Binary sentence classification – SENSE-LM on D_2^{EX}



(b) Binary sentence classification – SENSE-LM on D_2^{LM}

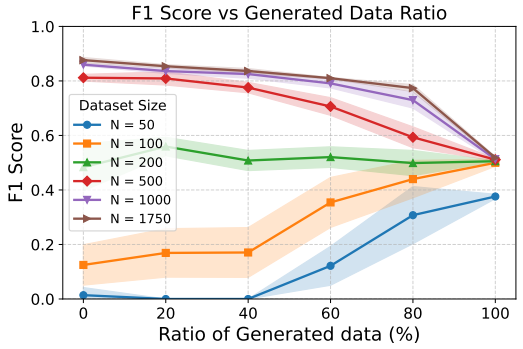


(c) Sensory terms extraction – SENSE-LM on D_2^{EX}

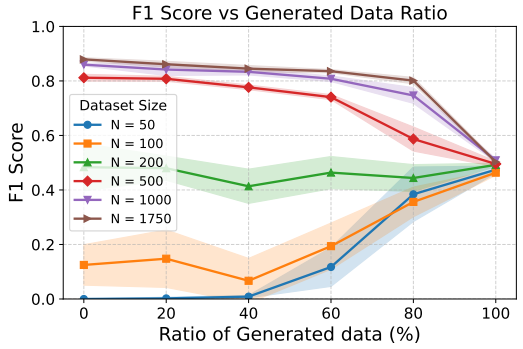


(d) Sensory terms extraction – SENSE-LM on D_2^{LM}

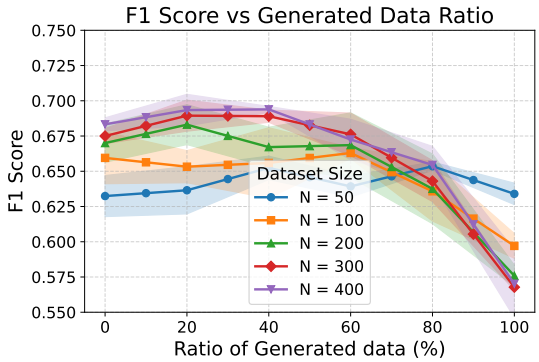
Figure 3: F1-score evolution of SENSE-LM on binary sentence classification (top row) and sensory terms extraction (bottom row), across various initial dataset sizes n_1 and progressive augmentation with synthetic data.



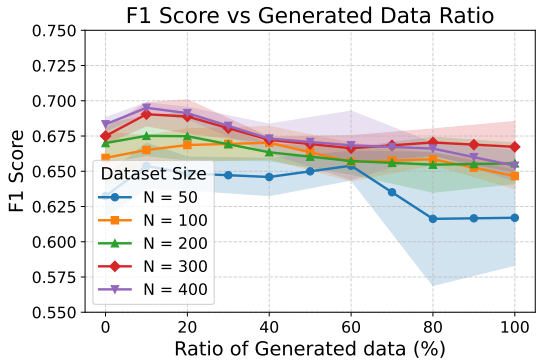
(a) Binary sentence classification – SENSE-LM on D_2^{EX}



(b) Binary sentence classification – SENSE-LM on D_2^{LM}



(c) Sensory terms extraction – SENSE-LM on D_2^{EX}



(d) Sensory terms extraction – SENSE-LM on D_2^{LM}

Figure 4: F1-score evolution of SENSE-LM on binary sentence classification (top row) and sensory terms extraction (bottom row), for various dataset sizes N and varying real-to-generated data ratios.

7 Limitations

While our study highlights the potential of LLM-generated data for olfactory information extraction, some aspects warrant further exploration and refinement in future work.

Annotation Quality and Agreement. Our comparison of human (D^2_{EX}) and model-based (D^2_{LM}) annotations shows moderate inter-annotator agreement (Cohen's $\kappa \approx 0.5$), particularly at the token level. This reflects the inherently subjective and nuanced nature of olfactory language, which poses challenges for both human and machine annotation. Notably, D^2_{EX} contributes valuable lexical diversity that enriches model learning, though it may also introduce variability that slightly affects classification precision. Additionally, since D_1 and D^2_{EX} were annotated by different experts, some divergence in their interpretation of sensoriality is natural. Addressing these cross-annotator differences in future work—for instance, via consensus-building or multi-annotator validation—could further enhance the robustness of human-annotated sensory corpora.

Domain Specificity. Our experiments center on olfactory content, a domain with particularly rich and complex linguistic characteristics. While our results suggest that synthetic olfactory data can effectively support classification tasks, further research is needed to determine how well these findings generalize to other sensory modalities. Each sensoriality (e.g., auditory, gustatory) brings its own cultural, lexical, and perceptual dimensions (Geldard, 1953), and extending this framework to new modalities would be a valuable direction. Encouragingly, the shared features between olfactory and gustatory language suggest that some transferability may be possible.

LLM Prompt Sensitivity. The success of synthetic data generation depends in part on prompt design. While our prompts were adapted from prior work and proved effective for our tasks, small changes in phrasing can result in substantial variations in the generated data. This highlights the importance of developing more standardized, reproducible prompting methodologies. Exploring prompt engineering techniques such as few-shot prompting or style-conditioning based on real corpora—e.g., using examples from the Odeuropa dataset—could further align generated data with domain-specific characteristics and make data augmentation with synthetic examples more efficient.

> **Prompt:** *Using the following examples, generate 200 new sentences incorporating olfactory references. Maintain a similar tone, vocabulary, and structure, while ensuring all sentences contain references to scent or smell. Avoid repetition or reproducing real-world examples.*

> **Example source:** *"Honey is gathered with much art from great variety of trees and flowers; and joy is a honey, a fragrancy made from above with much picking, choosing, and composing."*

LLMs Openness and Transparency In addition to proprietary models such as GPT-4, recent open-source LLMs like Qwen (Yang et al., 2025), LLaMA (Touvron et al., 2023), and Mistral (Jiang et al., 2023) have shown strong capabilities in generation and controllability. These models offer promising alternatives for institutions with data propriety or cost constraints and may support broader reproducibility in future synthetic data pipelines. Evaluating their behavior under controlled prompting conditions remains a valuable direction for future work.

Overall, while some challenges remain—particularly in annotation consistency and generalization beyond the olfactory domain—our findings underscore the promise of synthetic text data in low-resource settings. We believe this work contributes to expanding sensory NLP with LLM-driven resources, and we are optimistic about the scalability and adaptability of these methods in future applications.

Ethical Considerations

There is no risk of non-compliance with current legislation, such as GDPR or copyright law, since the generated data contains no sensitive information and is in the public domain. The real-world datasets used, notably Odeuropa, are public and open, made available by original authors for reuse. Composed of data from public-domain historical texts, this reuse complies with public-law standards.

However, Odeuropa data (historical) and LLM-generated data (contextualized in the contemporary world) may carry different biases, such as cultural

diversity and contextual nuances, which must be considered.

At the same time, since our study relies on synthetic data generated by GPT-4o, it is important to note that this data may contain factual inaccuracies, as GPT-4o does not incorporate any robust fact-checking mechanisms during text generation.

Acknowledgements

To guarantee the full anonymization of this paper during the double blind reviewing process, this section will be completed upon acceptance.

References

Tiago Almeida and Sérgio Matos. 2024. Exploring efficient zero-shot synthetic dataset generation for information retrieval. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1214–1231, St. Julian's, Malta. Association for Computational Linguistics.

Enrique Manjavacas Arevalo and Lauren Fonteyn. MacBERTh: Development and Evaluation of a Historically Pre-trained Language Model for English (1450-1950).

Cédric Boscher, Christine Largeron, Véronique Eglin, and Elöd Egyed-Zsigmond. 2024. SENSE-LM : A synergy between a language model and sensorimotor representations for auditory and olfactory information extraction. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1695–1711, St. Julian's, Malta. Association for Computational Linguistics.

Jenny Chim, Julia Ive, and Maria Liakata. 2025. Evaluating synthetic data generation from user generated text. *Computational Linguistics*, 51(1):191–233.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint*. ArXiv:1810.04805 [cs].

Frank A Geldard. 1953. The human senses. *Wiley*.

Dongsheng Jiang, Yuchen Liu, Songlin Liu, Jin'e Zhao, Hao Zhang, Zhen Gao, Xiaopeng Zhang, Jin Li, and Hongkai Xiong. 2023. From clip to dino: Visual encoders shout in multi-modal large language models. *arXiv preprint arXiv:2310.08825*.

Joshua Kazdan, Rylan Schaeffer, Apratim Dey, Matthias Gerstgrasser, Rafael Rafailov, David L Donoho, and Sanmi Koyejo. 2024. Collapse or thrive? perils and promises of synthetic data in a self-generating world. *arXiv preprint arXiv:2410.16713*.

Casey Kennington. 2021. Enriching Language Models with Visually-grounded Word Vectors and the Lancaster Sensorimotor Norms. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 148–157, Online. Association for Computational Linguistics.

Osama Khalid and Padmini Srinivasan. 2022. Smells like Teen Spirit: An Exploration of Sensorial Style in Literary Genres. *arXiv preprint*. ArXiv:2209.12352 [cs].

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

Zhuoyan Li, Hangxiao Zhu, Zhuoran Lu, and Ming Yin. 2023. Synthetic data generation with large language models for text classification: Potential and limitations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10443–10461, Singapore. Association for Computational Linguistics.

Lin Long, Rui Wang, Ruixuan Xiao, Junbo Zhao, Xiao Ding, Gang Chen, and Haobo Wang. 2024. On LLMs-driven synthetic data generation, curation, and evaluation: A survey. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11065–11082, Bangkok, Thailand. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2017. Fixing weight decay regularization in adam. *ArXiv*, abs/1711.05101.

Dermot Lynott, Louise Connell, Marc Brysbaert, James Brand, and James Carney. 2020. The Lancaster Sensorimotor Norms: multidimensional measures of perceptual and action strength for 40,000 English words. *Behavior Research Methods*, 52(3):1271–1291.

M. Besher Massri, Inna Novalija, Dunja Mladenić, Janez Brank, Sara Graça da Silva, Natasza Marrouch, Carla Murteira, Ali Hürriyetoğlu, and Beno Šircelj. 2022. Harvesting Context and Mining Emotions Related to Olfactory Cultural Heritage. *Multimodal Technologies and Interaction*, 6(7):57.

Stefano Menini, Teresa Paccosi, Sara Tonelli, Marieke Van Erp, Inger Leemans, Pasquale Lisena, Raphael Troncy, William Tullett, Ali Hürriyetoğlu, Ger Dijkstra, Femke Gordijn, Elias Jürgens, Josephine Koopman, Aron Ouwerkerk, Sanne Steen, Inna Novalija, Janez Brank, Dunja Mladenic, and Anja Zidar. 2022. A multilingual benchmark to capture olfactory situations over time. In *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, pages 1–10, Dublin, Ireland. Association for Computational Linguistics.

Jay Mohta, Kenan Ak, Yan Xu, and Mingwei Shen. 2023. Are large language models good annotators? In *Proceedings on "I Can't Believe It's Not Better: Failure Modes in the Age of Foundation Models" at NeurIPS 2023 Workshops*, volume 239 of *Proceedings of Machine Learning Research*, pages 38–48. PMLR.

Anders Giovanni Møller, Arianna Pera, Jacob Dalsgaard, and Luca Aiello. 2024. The parrot dilemma: Human-labeled vs. LLM-augmented data in classification tasks. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 179–192, St. Julian's, Malta. Association for Computational Linguistics.

Suzanne Mpouli, Michel Beigbeder, and Christine Largeron. 2020. Lexifield: a system for the automatic building of lexicons by semantic expansion of short word lists. *Knowledge and Information Systems*, 62(8):3181–3201.

Nazia Nafis, Inaki Esnaola, Alvaro Martinez-Perez, Maria-Cruz Villa-Uriol, and Venet Osmani. 2025. Critical challenges and guidelines in evaluating synthetic tabular data: A systematic review. *arXiv preprint arXiv:2504.18544*.

OpenAI. 2023. GPT-4 technical report. *CoRR*, arXiv.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Benoît Sagot and Darja Fišer. 2012. Automatic extension of wolf. In *GWC2012-6th International Global Wordnet Conference*.

Mohamed El Amine Seddik, Suei-Wen Chen, Soufiane Hayou, Pierre Youssef, and Merouane Debbah. 2024. How bad is training on synthetic data? a statistical analysis of language model collapse. *arXiv preprint arXiv:2404.05090*.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. 2024. Ai models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759.

Student. 1908. The probable error of a mean. *Biometrika*, 6(1):1–25.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Frank Wilcoxon. 1992. Individual comparisons by ranking methods. In *Breakthroughs in statistics: Methodology and distribution*, pages 196–202. Springer.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. 2023. Explainability for large language models: A survey. *arXiv preprint arXiv:2309.01029*.

A Dataset Generation Prompts

We provide the prompts used for the generation and automatic annotation of the D_2 dataset in Table 6. The **P1** prompted is used to generate positive examples, and conversely, **P2** is used to generate negative examples. Considering the limitations of the GPT-4o web application, we ask the model to generate the dataset by batch of 100 examples, that we compile into a CSV file along with their class at a sentence level (positive / negative).

Then, the prompt **P3** is used on positive examples to extract positive terms of D_2^{LM}annotation.

B General Experimental Settings

In the following experiments, all texts are preprocessed ahead of the model input. The text normalization involves spellchecking, removal of stop words and punctuation, lowercasing, and lemmatization. After normalization, tokenization is performed by splitting the text into tokens using whitespace as the delimiter.

In the classification experiments provided from Section 5.2 to Section 5.3.3, the SENSE-LM model uses BERT with MacBERTh pretrained parameters (Arevalo and Fonteyn) as a backbone, considering the same number of training parameters than the base BERT model (110M parameters). It is trained using the AdamW optimizer (Loshchilov and Hutter, 2017) with a learning rate of 2×10^{-5} and $\epsilon = 1 \times 10^{-8}$ for 30 epochs. BERT follows the same setup. For logistic regression, sensorimotor representations of sentences are extracted from the text following the method proposed by Boscher et al. (2024), and used to fit a logistic regression model with up to 1000 training iterations.

All experiments are conducted using an NVIDIA RTX A5000 Laptop GPU when relevant, and random seeds used for folds are fixed to 42 to ensure reproducibility.

C Lexical Dissimilarity between real-world and generated data

The lexical dissimilarity between D_1 and D_2's annotations is observed in Table 7 to Table 9, which show the frequency ranks of overlapping olfactory terms across distinct corpora. For example, while terms like *smell*, *scent*, and *aroma* are common in both corpora, their ranks vary considerably between real-world and synthetic corpora, pointing to a stylistic shift in expression across corpora.

Tables 7 and 8 show the top 20 overlapping olfactory terms between D_1 and each annotation of the generated corpus. The LM annotation (D_2^{LM}) has a more compact and repeated vocabulary (318 unique olfactory terms), with terms like *scent* and *aroma* ranking significantly higher than in D_1, indicating possible over-representation. The expert annotation (D_2^{EX}), on the other hand, maintains greater lexical diversity (902 unique olfactory terms) and alignment with D_1's vocabulary and rank order.

We statistically validate these observations using the Wilcoxon signed-rank test to compare the rankings presented respectively in Tables 7 and 8 (Wilcoxon, 1992):

- Between D_1 and D_2^{LM}, the p-value obtained is 2.58×10^{-6}, which is much lower than the significance level $\alpha = 0.05$. This result provides strong evidence to reject the null hypothesis and supports the claim positive terms labelled by a LLM in D_2^{LM}differ from terms labels by humans in D_1.
- In contrast, the comparison between D_1 and D_2^{EX}yields a p-value of **0.175**, suggesting no significant difference in the distribution of annotated terms.

Table 9 directly compares positive labelled terms between D_2^{LM} and D_2^{EX}, confirming that while they share a core vocabulary, the LM-annotation D_2^{LM}centers on explicit and non-ambiguously olfactory vocabulary, with a restricted range of unique terms (318) compared to D_2^{LM}, which counts 912 unique olfactory terms. In proportion, obvious terms such as scent, smell or aroma, are more representative in D_2^{LM}, showing higher relative frequencies.

D Additional Results on Corpus Comparison

This appendix provides extended results for the experiment conducted in Section 5.1. We provide corpus comparisons not only for all terms of D_1 and D_2, but also a comparison of positive terms in D_1, and respectively positive terms in $D_2^{\text{EX}}$and D_2^{LM}.

We show the obtained histograms in Figure 5. Whatever the terms considered, all terms or only positive terms and for these last last the way they

Table 6 Prompts Used for Generating Synthetic Sentences

Type	Prompt Description
P1 (Positive)	*"Could you generate 100 sentences of 10 words each, containing references to olfactory experiences, and avoid repeating the same sentence structures? You may include different kinds of descriptions: what produces the olfactory experience or the quality of smell, for different types of scents (people, objects, or environment)."*
P2 (Negative)	*"Could you generate 100 sentences with 10 words for each, making sure they absolutely do not make any reference to any olfactory experience, and avoid repeating the same sentence structures?"*
P3 (Positive Terms Annotation)	*"Extract words from the following sentences that evoke smells, explicitly or implicitly (e.g., describing smell quality or source). For example, from 'Musk pots generally moist exhales disagreeable predominant ammoniacal smell...' extract 'disagreeable, predominant, ammoniacal, musk, smell.'"*

Table 7 Top 20 olfactory terms from the model labelled generated corpus (D_2^{LM}) that most frequently appear in the original corpus (D_1). Columns show the term, its relative frequency in D_1 and D_2^{LM}(as percentages of total tokens), and its frequency rank within each corpus.

#	Term	% Freq. D_1	% Freq. D_2^{LM}	Rank in D_1 (over 777 terms)	Rank in D_2^{LM} (over 318 terms)
1	scent	2.16	9.95	4	1
2	smell	9.80	8.58	1	2
3	aroma	0.04	4.48	634	3
4	odor	6.69	3.92	2	4
5	sweet	0.69	2.55	17	5
6	perfume	3.19	1.68	3	7
7	smoke	0.17	1.49	82	9
8	fragrance	0.17	1.49	83	8
9	floral	0.22	1.31	77	11
10	pungent	0.43	1.06	29	12
11	fresh	0.26	0.93	50	13
12	lavender	0.22	0.87	72	14
13	whiff	0.09	0.81	175	16
14	garlic	0.30	0.75	42	20
15	onion	0.13	0.75	119	19
16	vanilla	0.04	0.75	626	17
17	acrid	0.22	0.68	71	22
18	oil	0.91	0.68	14	23
19	cinnamon	0.09	0.62	235	29
20	rise	0.26	0.56	58	31

Table 8 Top 20 olfactory terms from the human-labeled generated corpus (D_2^{EX}) that most frequently appear in the original corpus (D_1). Columns show the term, its relative frequency in D_1 and D_2^{EX} (as percentages of total tokens), and its frequency rank within each corpus.

#	Term	% Freq. D_1	% Freq. D_2^{EX}	Rank in D_1 (over 777 terms)	Rank in D_2^{EX} (over 902 terms)
1	smell	9.80	4.05	1	1
2	scent	2.16	3.38	4	2
3	aroma	0.04	3.22	634	3
4	faint	0.30	1.80	41	5
5	sweet	0.69	1.47	17	7
6	warm	0.04	1.25	720	10
7	air	0.35	1.22	36	11
8	fresh	0.26	0.91	50	12
9	rise	0.26	0.86	58	15
10	odor	6.69	0.80	2	17
11	rich	0.04	0.78	325	20
12	fragrance	0.17	0.75	83	21
13	floral	0.22	0.72	77	23
14	perfume	3.19	0.67	3	26
15	burn	0.09	0.64	194	28
16	acrid	0.22	0.61	71	29
17	pungent	0.43	0.61	29	30
18	damp	0.04	0.53	761	33
19	leave	0.43	0.44	30	38
20	old	0.04	0.44	430	39

Table 9 Top 20 olfactory terms from the model-labelled generated corpus (D_2^{LM}) that most frequently appear in the human-labelled generated corpus (D_2^{EX}). Columns show the term, its relative frequency in $D_2^{\text{EX}}$and D_2^{LM} (as percentages of total tokens), and its frequency rank within each corpus.

#	Term	% Freq. D_2^{EX}	% Freq. D_2^{LM}	Rank in D_2^{EX} (over 902 terms)	Rank in D_2^{LM} (over 318 terms)
1	scent	3.38	9.95	2	1
2	smell	4.05	8.58	1	2
3	aroma	3.22	4.48	3	3
4	odor	0.80	3.92	17	4
5	sweet	1.47	2.55	7	5
6	sharp	1.61	1.99	6	6
7	perfume	0.67	1.68	26	7
8	fragrance	0.75	1.49	21	8
9	smoke	0.25	1.49	66	9
10	earthy	1.36	1.37	8	10
11	floral	0.72	1.31	23	11
12	pungent	0.61	1.06	30	12
13	fresh	0.91	0.93	12	13
14	lavender	0.22	0.87	74	14
15	musty	0.30	0.87	52	15
16	whiff	0.08	0.81	215	16
17	vanilla	0.22	0.75	76	17
18	spicy	0.78	0.75	19	18
19	onion	0.08	0.75	251	19
20	garlic	0.14	0.75	118	20

have been labeled (automatically or by human), we obtain low token overlap (left column), indicating a highly dissimilar vocabulary. Semantic similarity is moderate (middle), and sensory similarity is higher (right), especially for D_2^{EX}(on bottom), showing alignment in olfactory semantics despite varying vocabulary.

E Additional Results on Data Augmentation Across Models

In the main paper, we only reported results for the SENSE-LM model due to space constraints. This appendix provides the full results for all models (SENSE-LM, BERT, and Logistic Regression) across the data augmentation settings introduced in Sections 4.3.2 and 4.3.3.

E.1 Data Augmentation with Varying Real-Data Sizes

This appendix expands on the results presented in the main paper for SENSE-LM by analyzing how synthetic data impacts model performance across different quantities of real-world training data. Specifically, we fix the number of real examples N and progressively introduce additional synthetic examples. We report results on both the binary sentence classification and sensory term extraction tasks.

Binary Sentence Classification We evaluate model performance at real-data sizes $n_1 \in \{50, 100, 200, 500, 1000, 1750\}$, tracking F1 scores as increasing amounts of synthetic data are added. Figure 6 presents results across models and annotation types. The X-axis represents the percentage of synthetic data relative to real data, while the Y-axis reports F1 score. Each curve corresponds to a different value of n_1.

For SENSE-LM, we observe consistent performance improvements when synthetic data is added, especially for D_2^{EX}at low and moderate values of n_1. Gains diminish as real-data availability increases, with performance largely plateauing beyond $n_1 = 500$. BERT shows more modest and variable improvements. Augmentation is more effective for D_2^{EX}than D_2^{LM}, with the clearest gains observed in the $n_1 = 100$ to $n_1 = 200$ range. For logistic regression, D_2^{LM}augmentation offers more reliable benefits than D_2^{EX}at small dataset sizes. However, these gains reduce as more real examples are introduced.

Overall, synthetic data augmentation is most beneficial under low-resource conditions. As the amount of real data increases, the marginal utility of synthetic examples declines.

Sensory Terms Extraction We apply the same evaluation protocol to the sensory term extraction task, using real dataset sizes $n_1 \in \{50, 100, 200, 300, 400\}$. Results are shown in Figure 7.

For SENSE-LM, adding synthetic data leads to consistent improvements across both $D_2^{\text{EX}}$and D_2^{LM}. Notably, D_2^{LM}annotations outperform D_2^{EX}, particularly at larger values of n_1. BERT shows more stable gains with D_2^{LM}, especially at medium and large dataset sizes. Improvements with $D_2^{\text{EX}}$are less consistent, and in some cases, augmentation has limited effect. We also observe a trade-off in precision and recall between annotation types. D_2^{LM}tends to improve precision, whereas D_2^{EX}primarily enhances recall. Detailed metrics are presented in Table 4.

E.2 Data Augmentation with Variable Real vs. Generated Ratio

We now examine how model performance is affected when real examples are progressively replaced with synthetic ones, keeping the total dataset size fixed.

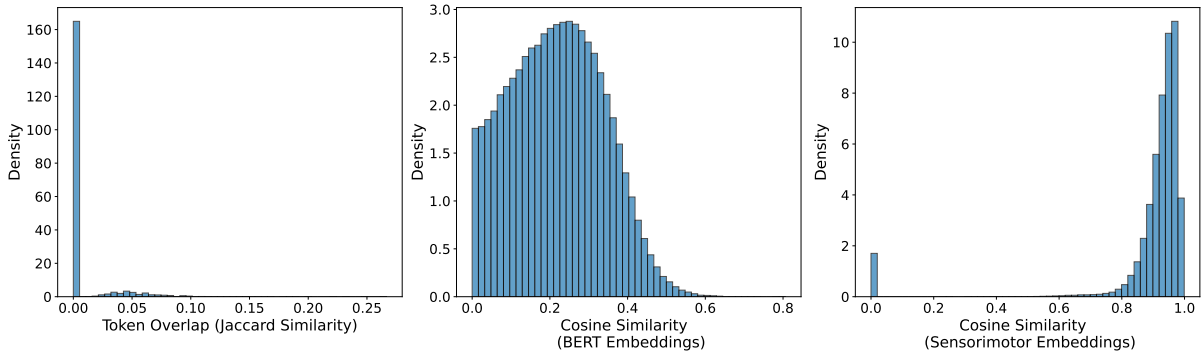
Binary Sentence Classification Figure 8 shows the F1 scores across different ratios of synthetic to real data, for several values of real dataset size N. Each subplot presents results for a specific model and annotation type.

Across all settings, performance begins to degrade once the proportion of synthetic data exceeds roughly 80%. This trend is consistent across models and annotation types. When using only synthetic data (i.e., 100% generated), model performance approaches the level of a random classifier.

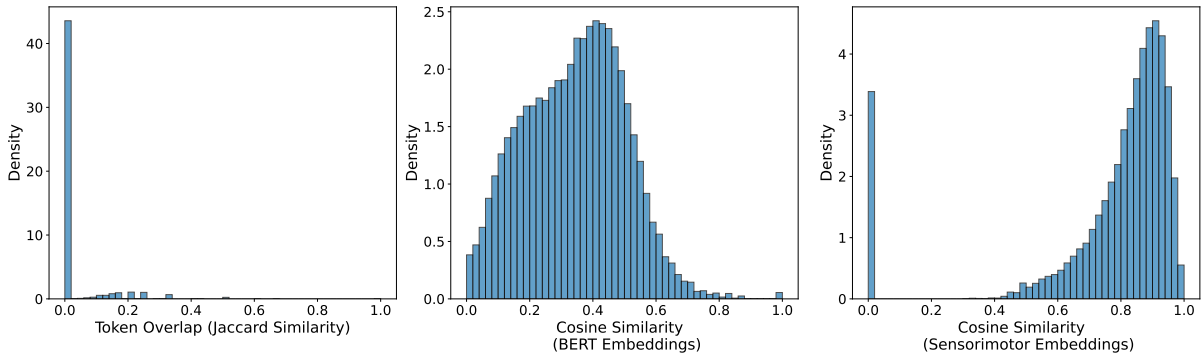
At smaller real-data sizes, D_2^{EX}tends to yield better results than D_2^{LM}, particularly when synthetic data is limited. However, as N increases, this advantage diminishes and the gap between annotation types narrows.

Sensory Terms Extraction Figure 9 reports results for the sensory term extraction task under varying real-to-generated data ratios. Precision and recall dynamics for each annotation strategy are presented in Table 5.

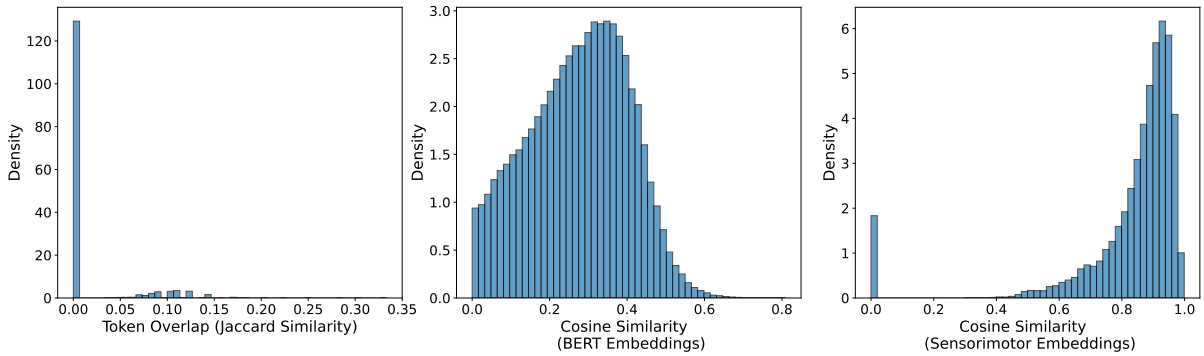
As synthetic data increases, D_2^{EX}annotations tend to improve recall but reduce precision. Conversely, D_2^{LM}annotations improve precision while sacrificing recall. In most settings, D_2^{LM}achieves more balanced F1 scores, indicating more favorable precision-recall trade-offs overall.



(a) D_1 vs D_2 (All terms)

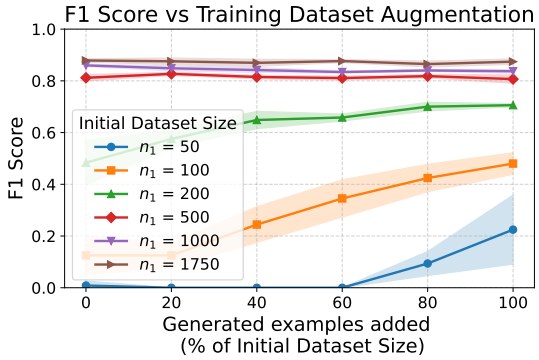


(b) D_1 vs D_2^{LM}(Positive terms)

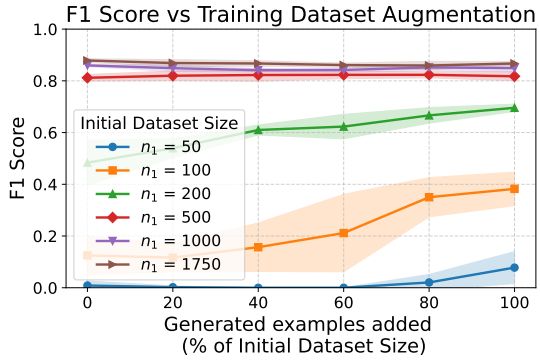


(c) D_1 vs D_2^{EX}(Positive terms)

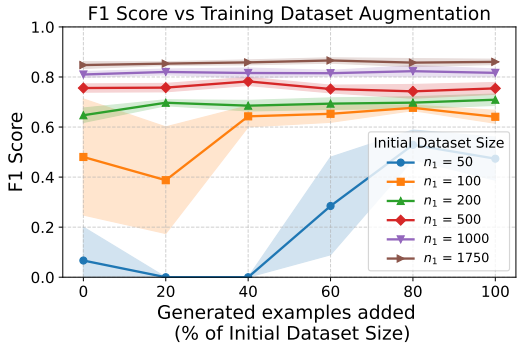
Figure 5: Comparison of sentence similarity distributions between positive sentences of generated (D_2) and original (D_1) corpora, using three metrics—token overlap, cosine similarity based on BERT Embeddings, and cosine similarity based on sensorimotor embeddings —under three conditions: (a) full-text comparison (D_1 vs. D_2), (b) sensory terms only for D1 v.s. D_2^{LM}, and (c) for D_1 vs D_2^{EX}.



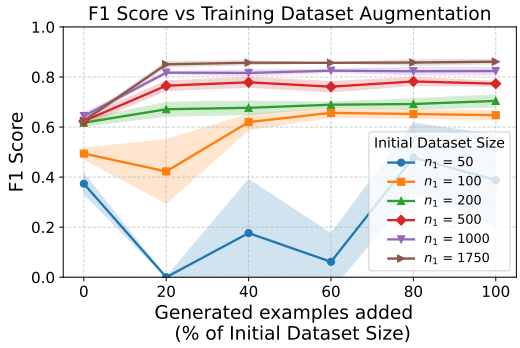
(a) SENSE-LM on D_2^{EX}



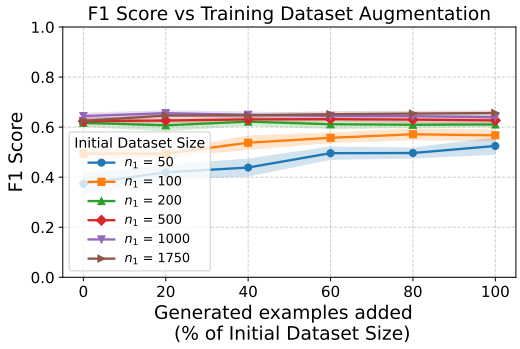
(b) SENSE-LM on D_2^{LM}



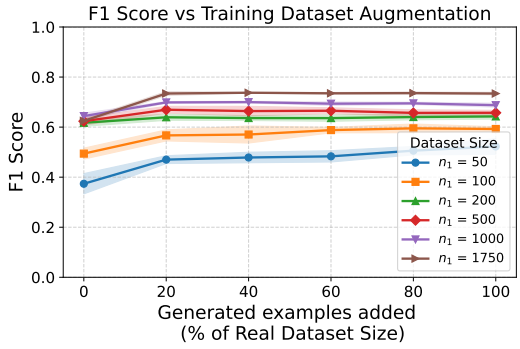
(c) BERT on D_2^{EX}



(d) BERT on D_2^{LM}

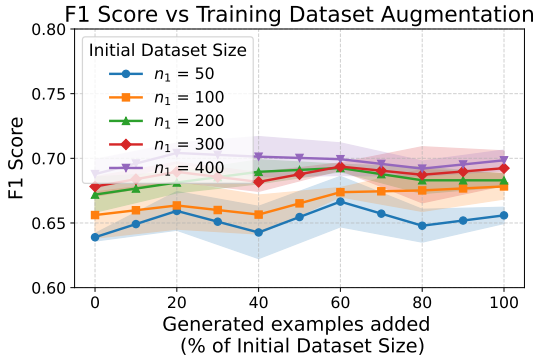


(e) Logistic Regression on D_2^{EX}

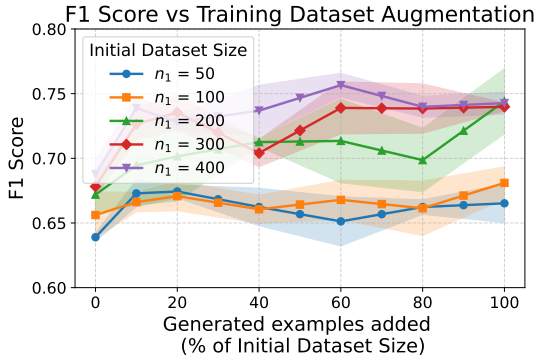


(f) Logistic Regression on D_2^{LM}

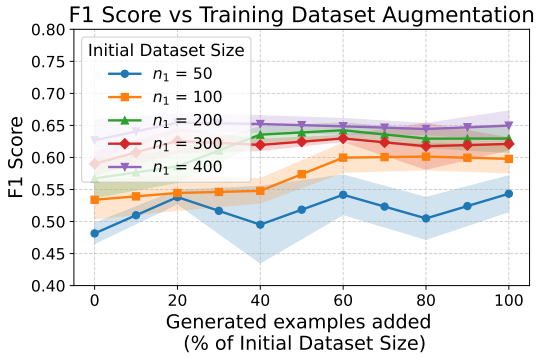
Figure 6: F1-Score evolution on binary sentence classification as synthetic data is added, for various initial dataset sizes n_1.



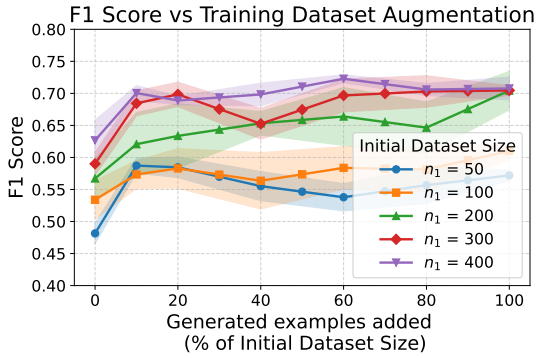
(a) SENSE-LM on D_2^{EX}



(b) SENSE-LM on D_2^{LM}

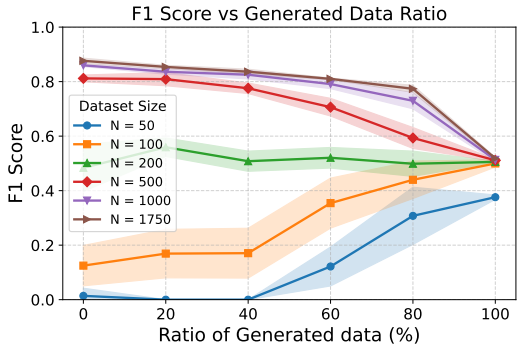


(c) BERT on D_2^{EX}

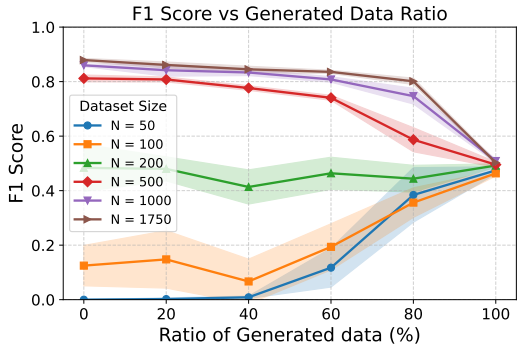


(d) BERT on D_2^{LM}

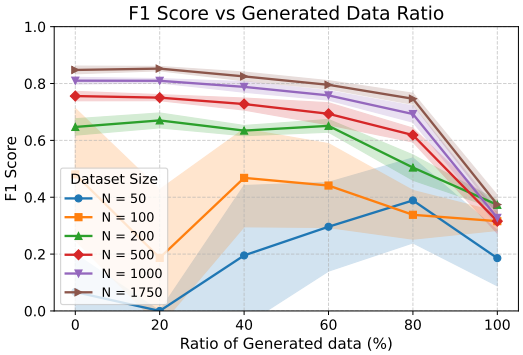
Figure 7: F1-Score evolution on sensory terms extraction as synthetic data is added, for various real dataset sizes.



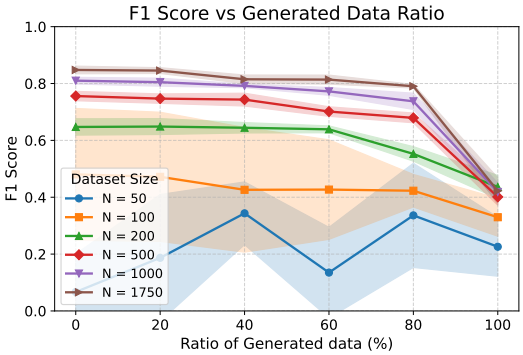
(a) SENSE-LM on D_2^{EX}



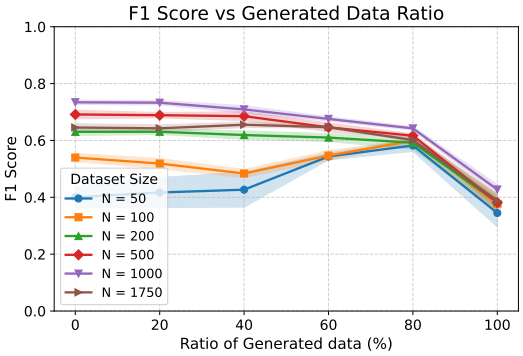
(b) SENSE-LM on D_2^{LM}



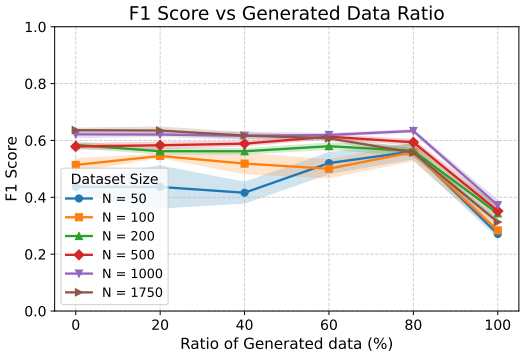
(c) BERT on D_2^{EX}



(d) BERT on D_2^{LM}

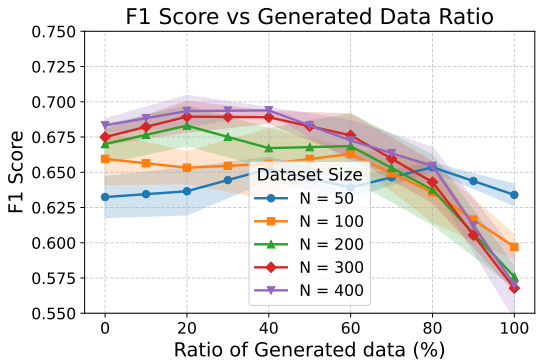


(e) Logistic Regression on D_2^{EX}

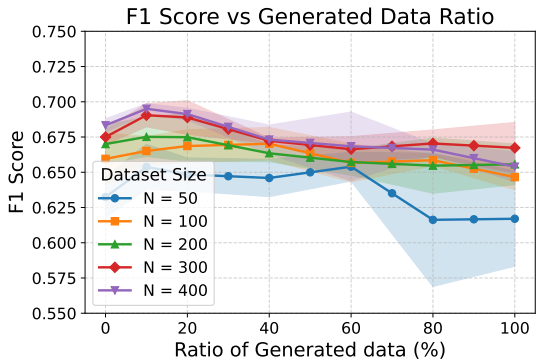


(f) Logistic Regression on D_2^{LM}

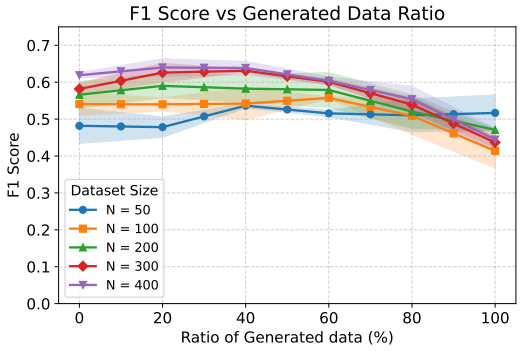
Figure 8: F1-Score evolution on binary sentence classification as the ratio of synthetic data varies, with a constant training dataset size N.



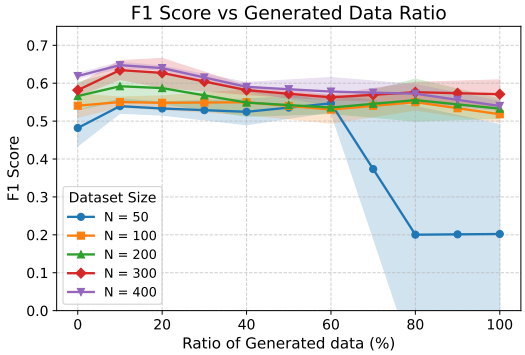
(a) SENSE-LM on D_2^{EX}



(b) SENSE-LM on D_2^{LM}



(c) BERT on D_2^{EX}



(d) BERT on D_2^{LM}

Figure 9: F1-Score evolution on sensory terms extraction as the ratio of synthetic data varies, with a constant training dataset size N.