The Zero Body Problem: Probing LLM Use of Sensory Language

Rebecca M. M. Hicke*
Department of Computer Science
Cornell University
Ithaca, NY, USA
rmh327@cornell.edu

Sil Hamilton & David Mimno
Department of Information Science
Cornell University
Ithaca, NY, USA
{srh255, mimno}@cornell.edu

Abstract

Sensory language expresses embodied experiences ranging from taste and sound to excitement and stomachache. It is of interest to scholars from a wide range of domains including robotics, narratology, linguistics, and cognitive science. In this work, we explore whether language models, which are not embodied, can approximate human use of embodied language. To do this, we extend an existing corpus of parallel human and model responses to short story prompts with an additional 18,000 stories generated by 18 popular language models. We find that all models generate stories that differ significantly from human usage of sensory language. However, the direction of these differences varies considerably between model families; Gemini models use significantly more sensory language than humans along most axes whereas most models from the remaining five families use significantly less. Linear probes run on five models suggest that they are capable of *identifying* sensory language, meaning an inability to recognize sensory content is unlikely to be the cause of the observed differences. Instead, we find preliminary evidence indicating that instruction tuning may discourage usage of sensory language in some models. To support further work, we release our expanded story dataset.

1 Introduction

Sensory language makes use of descriptive words and phrases to evoke embodied experiences. It encompasses language that appeals to senses like taste, sound, or representations of physical actions. The use of sensory and embodied language is of interest to scholars from many domains including robotics (Madden et al., 2010; Alomari et al., 2017; Taniguchi et al., 2016), linguistics (Winter, 2019; Mondada, 2021; Strik Lievers & Winter, 2018), cognitive science (Muraki et al., 2023; Dove et al., 2022; Davis & Yee, 2021; Caballero & Paradis, 2023), and narratology (Piper, 2024; Piper & Bagga, 2024; Caracciolo & Kukkonen, 2021; Herman, 2009; Fludernik, 1996). Because large language models (LLMs) do not have embodied experiences, we might expect their use of sensory language to be consistently different from that of humans. On the other hand, language models do not have *any* experiences, solely learning human linguistic patterns from examples of human-written text, which includes sensory language. In this work, we explore these hypotheses and probe whether LLMs mimic human levels of sensory language usage.

To do this, we extend an existing corpus of parallel human and GPT-3.5 responses to short story prompts (Huang et al., 2024). We select 1,000 prompts from the original dataset, randomly sample a human and GPT-3.5 response to each prompt, and then generate responses to these prompts from an additional 18 language models from six prominent model families: Gemini, GPT, Llama, OLMo, Phi, and Qwen. We then use two pre-existing lexicons from cognitive science (Lynott et al., 2020; Brysbaert et al., 2014) to measure the strength of sensory language usage in all 20,000 texts along twelve axes: auditory, gustatory, haptic, interocep-

*Corresponding author

tive, olfactory, visual, foot/leg, hand/arm, head, mouth, torso, and concreteness. We finally perform pairwise comparisons between the strength of sensory content in human-written texts and texts written by each model.

All nineteen models generated stories differing from human usage of sensory language in at least ten of the twelve sensory axes. However, the direction of these differences varied considerably between model families. In particular, Gemini models used significantly *more* sensory language than humans along most axes, while models belonging to the Llama, OLMo, Phi, and Qwen families used significantly *less* sensory language. The differences for the GPT models varied considerably between models and sensory axes.

We propose two possible explanations for this behavior. First, LLMs may fail to learn (and therefore replicate) sensory language during pre-training. However, we find that linear probes run on intermediate representations of human-written texts from five LLMs indicate that models *are* able to recognize sensory language usage, at least for several axes. Second, model usage of sensory language may be influenced by instruction tuning. An analysis of the sensory content in encouraged and discouraged model responses from one popular reinforcement learning from human feedback (RLHF) dataset (Bai et al., 2022) supports this hypothesis. Specifically, we find a strong correlation between underuse of sensory language by models and over-representation of sensory language in discouraged model responses from the RLHF dataset.

LLMs' embodied language use provides a fascinating case study with implications for fields ranging from abstract philosophy to human-computer interaction. This work demonstrates that LLMs do not approximate human use of embodied language when prompted for creative writing. It further provides evidence that LLMs are able to represent sensory content but that instruction tuning may discourage its usage. In short, we appear to have discovered a mirror, not a window; we asked why LLMs don't use sensory language, and found it seems we asked them not to.

2 Related Work

Considerable academic interest has recently been paid to LLMs' sensory knowledge. Much of this research has focused on probing models' abilities to replicate human-like sensory judgments for features including color (Kawakita et al., 2024; Paik et al., 2021; Marjieh et al., 2024), sound (Siedenburg & Saitis, 2023), and the perceptual strength of words along several sensory axes (including six studied in this paper) (Lee et al., 2025). Other studies examined whether embodied concepts like space and time (Gurnee & Tegmark, 2024), object sounds (Ngo & Kim, 2024), implicit visual features (Jones & Trott, 2024), and color (Abdou et al., 2021) are represented in models' internal states. More work has investigated whether models can identify stimuli from human descriptions of sensory experiences (Zhong et al., 2024a;b), whether humans understand model descriptions of sensory features (Zhong et al., 2024b), and whether models can reason about generic visual concepts (Zhang et al., 2022) or the visual-auditory properties of language (Lee & Lim, 2024). Finally, some papers inspect whether incorporating increased multi-modal or perceptual training improves models' performance on related tasks (Kennington, 2021; Li et al., 2024).

While the studies described above largely find that LLMs demonstrate at least some knowledge of sensory experience, to our knowledge no work has examined whether LLMs' use of sensory *language* is human-like or if their descriptions of sensory experiences in non-explicitly perceptual tasks mimics human language.

3 Data

To create a comparative corpus of human and model short stories, we build on the GPT-WritingPrompts dataset (Huang et al., 2024). This dataset is itself an extension of the original WritingPrompts dataset (Fan et al., 2018) which contains 303,358 stories written by users

from `r/WritingPrompts`[1] in response to 97,222 creative writing prompts. Because this dataset was collected before and released in 2018, we assume there is little to no computationally generated content in the human responses. Huang et al. (2024) expanded this dataset by generating 206,226 responses to 97,219 of those prompts from GPT-3.5-turbo. They used a temperature of 0.95 and two system prompts: the 'author' prompt ("You are an award winning creative short story writer.") and the 'reddit' prompt ("You're writing a Reddit story and you want other reddit users to like and upvote your story.").

We further extend this dataset to new model families by randomly selecting a subset of 1,000 prompts from the GPT-WritingPrompts dataset. We reject all 'prompts' that are not creative writing prompts (such as community announcements) and prompts that do not pass GPT-4o's safety filters. We then sample a random human and GPT-3.5 response to each of these prompts, ensuring that each human response is a story and not a comment or other forum content. We then generate additional responses to the 1,000 prompts from each of the following 18 models:

- **Gemini**: 1.5 — Flash-8b, Flash (Gemini et al., 2024); 2.0 — Flash-Lite, Flash (Gemini, 2025)
- **GPT**: 4o (OpenAI, 2024)
- **Llama**: 3.1 — 8b; 3.2 — 1b, 3b, 11b (vision); 3.3 — 70b (Grattafiori et al., 2024)
- **OLMo 2**: 7b, 13b (OLMo et al., 2024)
- **Phi 4**: base, mini (Abdin et al., 2024)
- **Qwen 2.5**: 0.5b, 1.5b, 3b, 7b (Qwen et al., 2025)

We select the instruction-tuned variant of each model to enable convenient story generation and follow Huang et al. (2024) in setting temperature to 0.95. For each model, we randomly apply the 'author' system prompt when generating responses to 500 prompts and the 'reddit' prompt for the remaining 500. The Gemini and GPT models are accessed using the Google and Azure OpenAI APIs respectively. The remaining models were all accessed via HuggingFace. We release the additional generations for each prompt here.

In order to measure sensory word usage, we further draw on two datasets from cognitive linguistics: a sensorimotor lexicon (Lynott et al., 2020) and a concreteness lexicon, where concreteness is defined as "the degree to which the concept denoted by a word refers to a perceptible entity" (Brysbaert et al., 2014). The sensorimotor lexicon includes information on eleven sensory axes, six of which are perceptual modalities (haptic, auditory, olfactory, gustatory, visual, and interoceptive) and five of which are action effectors, or body parts which respond to a stimulus (mouth + throat, hand + arm, foot + leg, head excluding mouth + throat, and torso). The mean ratings along these axes were generated for 37,058 English-language lemmas by 3,500 native English speakers on Mechanical Turk who ranked lemmas on a scale of 0 to 5, where 0 represented lemmas "not experienced at all with that sense/action" for the sensorimotor axes or abstract terms and 5 represented lemmas "experienced greatly with that sense/action" or concrete terms (Lynott et al., 2020). Further information about these datasets and the data collection process can be found in the original articles.

4 Methods

Measuring Sensory Language Our goal is to measure sensory language usage in the produced texts along each of the twelve axes outlined above. We would like to represent the relative sensory content of each story, giving more weight to the sensory contributions of less common words and down-weighting the impact of frequently used (or stop) words. To do this, we first tokenize and lemmatize each text in our dataset using `spaCy` (Honnibal et al., 2020). We then use `scikit-learn` to find the inverse document frequency (IDF) values for each lemma in our corpus (Pedregosa et al., 2011). We calculate the IDF values for each lemma from the dataset consisting of human and model responses to each of the

[1] https://www.reddit.com/r/WritingPrompts/

Sensory Axis	Highly Sensory Sample	Sensory Score
Auditory	The air buzzed with polyphonic chatter.	0.84
Gustatory	He buttered a colossal slice of bread.	0.87
Haptic	Twirling the pen between his fingers.	0.75
Interoceptive	The capacity to love others richly begins.	0.41
Olfactory	Her sweat bore the fragrance and sweetness of fruit.	0.78
Visual	The sun peeked over the horizon.	0.86
Foot – Leg	She settled into her stride.	0.69
Hand – Arm	A muscle-bound wizard was doing bicep curls.	1.14
Head	He didn't need to blink.	0.62
Mouth	A knowing smile played on her lips.	0.54
Torso	My lungs filled the void left in his ribcage.	0.65
Concreteness	Cold air snuck past my tattered clothing.	1.23

Table 1: Examples of the texts that achieve high scores along the twelve axes of sensory language. Each text is paraphrased from an output given by GPT 3.5.

1,000 selected prompts (20,000 documents total). Next, we normalize the IDF values to fall between zero and one by dividing each value by the largest IDF value in the dataset. The IDF values are used to weight the contribution of each lemma to the finally sensory score. This weighting method allows us to proportionally scale each lemma's contribution to the overall story score without using stopword lists or removing words with low sensory scores, methods which have little justificaiton and may dramatically change the sensory value distribution.

We then use the sensory lexicons in concert with the normalized IDF values to create normalized sensory scores along each sensory axis for each story. To create these scores, we first look up every lemma in a text in the sensory lexicons. If the lemma is included in the lexicons, we multiply its sensory value along each axis by its normed IDF value and add the resulting numbers to the summed axis scores for the story. Finally, we normalize the summed sensory scores for a text by dividing by the number of lemmas used to calculate the score; that is, the number of lemmas in the text found in the sensory lexicons. This essentially mimics norming the scores by a story's word count, but avoids skewing the results if certain 'authors' use terms not found in the lexicon more frequently. This leaves us with twelve normalized strength scores for each story representing the magnitude of sensory language usage along each axis. Examples of texts that rank highly along each axis with their corresponding sensory strength scores can be found in Table 1.

Comparing Sensory Language We can now use the numerical representations of each text's sensory language usage to compare responses by humans and each model. We produce two numerical comparisons of human and model sensory language usage. First, for each axis we subtract the mean sensory strength of all the model's responses from the mean sensory strength of all the human responses. A negative difference in averages means that the model used more sensory language than the humans on average and vice versa.

Next, to determine whether the differences in sensory language usage are significant, we use a paired t-test. Specifically, we compare the normalized sensory strengths of human and model responses to the same prompts along each axis.[2] We then report the resulting t-test statistics for each test, which provide information on the significance, strength, and direction of the differences in distribution means. Again, negative t-test scores indicate that the model used more sensory language than the human writers and vice versa. All t-test statistics ≥ 1.96 or ≤ -1.96 are significant at $\alpha = 0.05$.

[2]Paired t-tests are implemented using `scipy`'s `ttest_rel()` function.

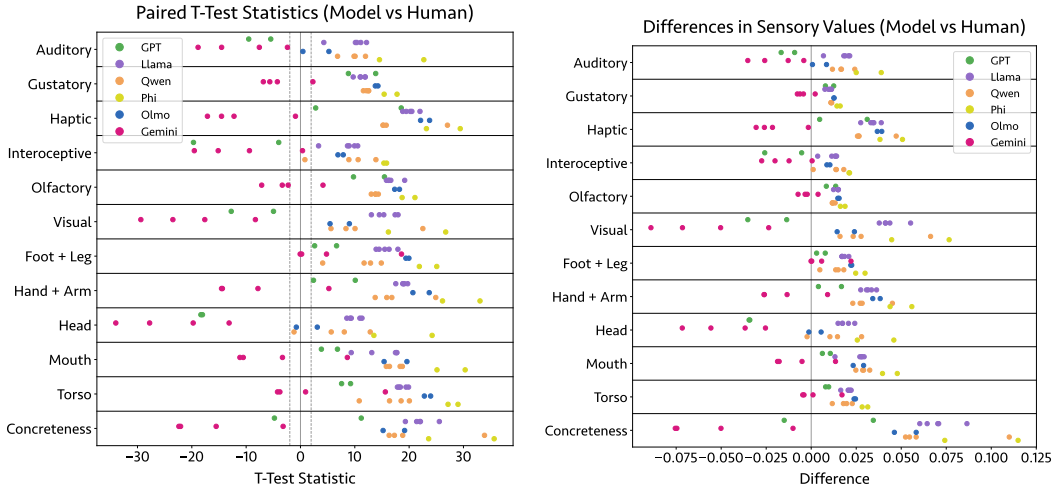

Figure 1: Comparisons of model and human use of sensory language along each of the twelve axes for each model; values < 0 indicate the model used more sensory language and vice versa. The paired t-test statistics (left) and average differences (right) demonstrate significant differences in model and human language usage for most models along most axes. The dotted vertical lines on the left figure mark significance at $\alpha = 0.05$; any t-test statistics ≥ 1.96 or ≤ -1.96 represents a significant difference in model and human language usage.

Distinguishing Model- and Human-Written Texts Can we distinguish model- and human-written texts using only the strength of their sensory language usage along the twelve axes of interest? Having calculated the magnitude of sensory language usage for each text along each axis, we represent each story as a vector composed of each respective sensory strength. We then use these vectors to train and evaluate the ability of logistic regression models to distinguish between texts produced by humans and each LLM.

Axis	Avg. Strength
Auditory	0.41
Gustatory	0.10
Haptic	0.30
Interoceptive	0.29
Olfactory	0.12
Visual	0.68
Foot + Leg	0.25
Hand + Arm	0.38
Head	0.57
Mouth	0.35
Torso	0.24
Concreteness	0.73

Table 2: Average sensory strength score for human-written stories along each axis.

For every LLM, we train 100 logistic regression models to distinguish between stories written by humans and that LLM. We reshuffle the train/test data splits for every model. Each training dataset contains the responses to 500 randomly selected prompts (1,000 vectors) and the test dataset contains the responses to the remaining 500 prompts. For each LLM, we report the average F1 score over all 100 logistic regression models and the standard deviation of all F1 scores. Further, we find the importance of each sensory axis by permuting the input features[3] and report the average importance of each feature.

5 Comparing Sensory Language

Every model studied differs significantly at $\alpha \leq 0.05$ from human usage of sensory language along at least ten out of twelve sensory axes; most models differ significantly along all twelve. This strongly demonstrates that models do not mimic human levels of sensory language use. We note that the average differences are relatively small compared to the

[3]Implemented with `scikit-learn`'s `permutation_importance` function.

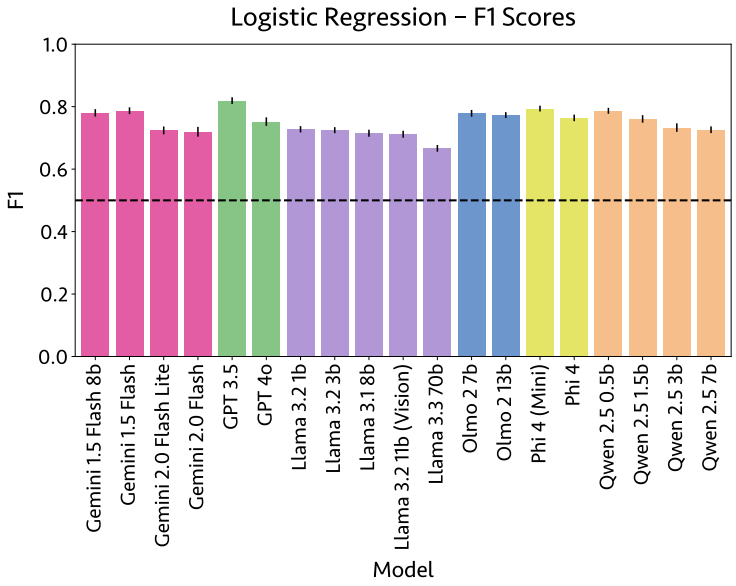

Figure 2: The average performance (F1) of 100 logistic regression models in distinguishing between texts written by humans and texts written by each model. Models are trained with different 50:50 training and test splits. The error bars represent the standard deviation over all 100 model runs and the dotted line marks expected random performance.

average sensory values along each axis for human stories (Table 2); however, the t-tests confirm that the differences are nonetheless significant.

We also find clear differences in how models from different families deviate from human sensory language usage. Models from all families except GPT and Gemini use sensory language significantly less than humans along nearly ever axis (Figure 1). In contrast, Gemini models use significantly more sensory language than human writers along most axes and GPT models fall between the Gemini models and models from the other families.

There are particularly clear divides between the model families along the visual and concreteness axes; the Gemini models and GPT-4o use significantly more visual and concrete language than humans whereas the other models use significantly less. To exemplify these differences, we provide the first two sentences of the responses to a single prompt[4] from Gemini 2.0 Flash, one of the models that uses the most concrete and visual language compared to human writers, and Phi 4, one of the models that uses the least concrete and visual language:

> **Gemini 2.0 Flash:** "The chipped ceramic mug warmed my hands, the lukewarm tea doing little to soothe the tremor in my soul. Rain lashed against the windowpane, mimicking the relentless rhythm of my thoughts."
>
> **Phi 4:** "Ever since I was a child, I had an uncanny knack for noticing the peculiarities in everyday life. It was a skill that often left me feeling isolated, as if I were the only one who saw the world through a different lens."

The difference between model and human language usage is stronger for GPT-3.5 than GPT-4o along all but three axes: the action effectors Head, Mouth, and Hand + Arm. This suggests that changes made between the creation of these two models may have reduced the difference between the extent of human and model sensory language use. In particular,

[4] "Every human that's ever lived has met God. He takes the shape of a mailman, teacher, store clerk or another passerby to evaluate you based on one morality test. Out of the 107 billion humans that's ever been alive you are the only one that figured out who he really is..."

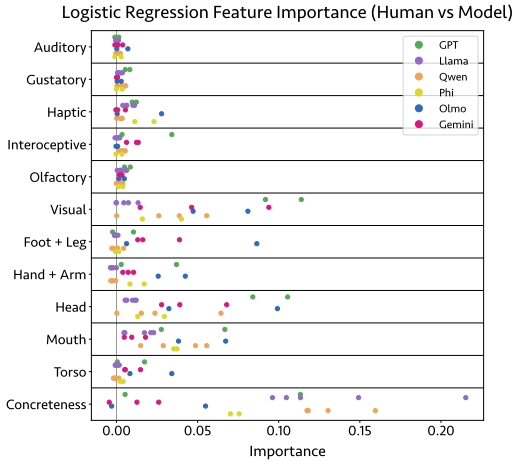

Figure 3: The average feature importance from each logistic regression model trained to distinguish between models and humans. Each dot represents an average over 100 models.

Sensory Axis	Mean Diff.	T-Test Stat.
Auditory	0.00065	0.92
Gustatory	0.00060	1.46
Haptic*	0.01747	27.72
Interoceptive*	-0.01114	-13.74
Olfactory*	0.00303	7.13
Visual*	0.04196	40.79
Foot + Leg*	0.00930	19.92
Hand + Arm*	0.02353	36.46
Head*	0.01165	15.80
Mouth*	-0.00216	-3.40
Torso*	0.00877	19.81
Concreteness*	0.03664	32.48

Table 3: Comparisons of sensory language usage between rejected and chosen responses in the Anthropic RLHF dataset (Bai et al., 2022). Negative values mean the chosen responses used more sensory language and vice versa. Axes for which there are significant differences at $\alpha = 0.05$ are marked with a star (*).

the multi-modal training and capabilities of GPT-4o may move its sensory language use closer to human levels.

We additionally find that logistic regression models are able to distinguish between texts written by humans and each model with well above random accuracy (Figure 2), further confirming that the sensory language use of humans and each model differs considerably. The logistic regression models tend to perform worse when distinguishing between human texts and those written by larger and newer models in a family, suggesting that larger models' sensory language use may be more similar to humans. By further examining the average importance of each sensory feature used by the logistic regression models, we see that the visual and concreteness axes are again frequently important (Figure 3). Overall, action effectors appear to be more important than the perceptual modalities for the logistic regression models, although the differences between human and model usage of these axes are not necessarily more significant (Figure 1).

6 Probing for Sensory Language

While our results suggest contemporary instruction-tuned language models do not replicate human patterns of sensory language usage, this does not mean LLMs fail to capture the concept during pre-training. We can probe for this linguistic knowledge by training regression models to identify sensory language from language models' latent representations. Researchers commonly train such simple linear models, or linear probes, to identify whether language models implicitly learn linguistic phenomena during pre-training (Shi et al., 2016; Tenney et al., 2019; Marks & Tegmark, 2024).

In this experiment, we probe for each sensory axis by first calculating sequence-level sensory scores for all 272,600 human-written stories from the GPT-WritingPrompts dataset. We pass each story through five models selected for language model type (MLM, seq2seq, and CLM), recency, and ubiquity in interpretability research: BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), T5 (Raffel et al., 2023), GPT-2 (Radford et al., 2018), and Qwen 2.5 0.5b. We collect embeddings from BERT and RoBERTa by extracting the embedding of the `CLS` token for every layer and from T5 (encoder), GPT, and Qwen by taking the mean hidden state of every layer as a proxy for pooled embeddings. This process yields 3,543,800 embeddings for a 12-layer transformer when including the embedding layer.

Model	Probe/Model	Anthropic/Model
Gemini 1.5 Flash 8b	-0.54	-0.46
Gemini 1.5 Flash	-0.70*	-0.69*
Gemini 2.0 Flash	-0.76*	-0.75*
Gemini 2.0 Flash-Lite	-0.74*	-0.75*
GPT-3.5	0.15	0.32
GPT-4o	-0.52	-0.52
Llama 3.2 1b	0.74*	0.86*
Llama 3.2 3b	0.73*	0.82*
Llama 3.1 8b	0.71*	0.79*
Llama 3.2 11b (Vision)	0.71*	0.81*
Llama 3.3 70b	0.69*	0.92*
OLMo 2 7b	0.46	0.64*
OLMo 2 13b	0.27	0.51
Phi 4 (mini)	0.69*	0.76*
Phi 4	0.78*	0.83*
Qwen 2.5 0.5b	0.76*	0.84*
Qwen 2.5 1.5b	0.59*	0.57
Qwen 2.5 3b	0.64*	0.61*
Qwen 2.5 7b	0.66*	0.65*

Table 4: Correlations between average differences along each sensory axis for each model (human minus model) and (column 1) average linear probe performance or (column 2) differences in sensory language for the Anthropic RLHF dataset (rejected minus chosen). Results which are significant at $\alpha = 0.05$ are marked with a star (*).

We then tag every embedding with the corresponding sensory values observed for the sequence from which it was generated and bin the embeddings into train and test sets on a per-layer basis with a ratio of 80:20. Because our target sensory values are selected from the range [0,5] we use ℓ_2-regularized ridge regression models to predict them. We allow the ridge regression implementation distributed with `scikit-learn` to automatically select a solver. For each combination of language model and sensory axis, we train a regression model to predict sensory values given a list of passages. We then calculate the R^2 between these predicted values and the ground truth produced by our lexicons of interest. Higher R^2 values indicate a given sensory axis is better represented in the latent representations of a given language model. We further select the best performing α for each model by repeating the training process five times while incrementing α by 0.20, from 0 to 1. We finally select the best performing ridge regression model from this set, per probe.

Global probe performance is varied according to layer, model, and sensory axis. The best performing probes achieve $R^2 \approx 0.85$ when predicting concreteness in deeper model layers, a result corresponding with prior work in BERTology suggesting LLMs resolve syntactic features early in processing (Tenney et al., 2019). The probes for most sensory axes achieve $0.3 \leq R^2 \leq 0.6$ across most layers of all models, including auditory, gustatory, and haptic. The least predictable sensory axis is the torso action effector, with no probe achieving a $R^2 > 0.17$. Nonetheless, these results demonstrate that even smaller, older LLMs are capable of representing most sensory axes, which suggests that the difference between LLM and human use of sensory language is not due to their inability to recognize its usage.

Differences between the language models indicate newer and longer-trained models yield better performing linear probes, suggesting extended training time results in sensory values being more coherently embedded in latent representations. The degree to which our probes resolve each sensory axis is consistent across all models with an inter-model Pearson's $R \approx 0.98$ for any two given models, indicating that all models resolve the same sensory axes to the same degree, despite variations in absolute performance. We then quantify the relative accuracy with which the models are able to represent each sensory axis using the average probe performance across all models and all layers.

We observe significant correlation between the average probe performance per sensory axis and the human-model differences reported in Section 5. We find Pearson's $R \geq 0.45$ for all models except for the Gemini models, GPT 3.5, and Olmo 2 13b (Table 4: Probe/Model). This means that the more a sensory axis is represented in latent representations of the five probed models, the *less* most models use that language. In other words, the facets that are most easily recognized by models deviate the most from human usage.

7 Impact of RLHF Training

Is downstream instruction tuning responsible for the differences between human and model sensory language usage? To explore the effect of reinforcement learning from human feedback (RLHF) on model behavior, we draw on the popular Anthropic RLHF dataset for use as our target RHLF dataset (Bai et al., 2022). This dataset contains 44,848 paired continuations of human-model interactions, of which one is considered a favorable response (labeled 'chosen') and the other is considered unfavorable (labeled 'rejected'). Although we do not know which datasets were used in training and finetuning many of the propriety models used in this study, we anticipate many will have trained on the Anthropic dataset or something similar. At a minimum, Microsoft has disclosed that it used the Anthropic dataset in training the Phi 4 models (Abdin et al., 2024).

In this experiment, we compare how sensory language is used in the rejected and chosen responses from the Anthropic RLHF dataset. Because the responses are paired, we are able to treat them much like we do each set of human and model prompt responses above. We first generate the normalized IDF scores for all of the lemmas in this dataset. We then extract the parts of each paired exchange that differ between the rejected and chosen responses (the assistant's last reply). We finally measure the difference in sensory language use between the rejected and chosen assistant responses, again producing the average difference in sensory strength along each axis and the t-test statistics comparing sensory content in each pair of responses. This process reveals there is a significant difference in the strength of sensory language use between the rejected and chosen model responses along all axes except for auditory and gustatory (Table 3). Of the significant differences, the rejected responses use more of each kind of sensory language except for interoception and mouth action effectors.

To probe whether a relationship may exist between training with this dataset and models' non-human use of sensory language, we examine the correlation between the average differences in human and model sensory language usage (human minus model) and the average differences in rejected versus chosen responses' sensory language usage (rejected minus chosen). We find that significant correlations between these values exist for most models, but that again the correlations differ considerably between models from different families (Table 4: Anthropic/Model).

Average differences between the Gemini models and humans are all correlated negatively with trends in the Anthropic dataset. Thus, along axes where the rejected Anthropic responses used more sensory language, the Gemini models tend to use more sensory language than humans. In contrast, the correlations between the GPT models' sensory language use and that in the Anthropic dataset are not significant. For most models from the remaining four families — Llama, OLMo, Phi, and Qwen — there is a significant positive relationship between the average differences in human and model sensory language use and sensory language use in the rejected and chosen responses. Thus, the more particular forms of sensory language were used in discouraged responses from the Anthropic dataset, the less models from these families used that language in comparison with humans. These correlations therefore provide evidence that RLHF training with the Anthropic dataset changes how models use sensory language, in particular discouraging the use of some forms of sensory language.

8 Conclusion

Despite the constant tendency to anthropomorphize language models, they are not human and do not in any way experience embodied human senses. But there is no *a priori* reason to believe that they cannot emulate the language of embodied humans.

In this work we find that LLMs do *not* replicate human patterns of embodied and sensory language use along twelve axes. Differences between model and human behavior vary considerably by model family; whereas Gemini models use far more sensory language than humans along most axes, models from other studied families tend to use significantly less. Probing for sensory language in model latent representations suggests LLMs are able to identify sensory language despite not using it. Moreover, our results suggest the better a model understands a sensory axis the less likely it is to use it. We investigated why this phenomenon may occur by examining sensory language use in a common RLHF dataset. This revealed post-training RLHF may be discouraging models from using many forms of sensory language.

These findings suggest that learning from large corpora of human-written texts allows models to partially identify sensory language, but downstream instruction tuning frequently discourages its use. This has implications for the use of LLMs in tasks requiring empathy and world awareness such as creative writing, robotics, and therapy. Our results suggest instruction tuning may have unintended consequences on model behavior in non-obvious ways. Possible side-effects should be carefully considered when designing and training models. We urge LLM researchers to consider drawing on tasks from psycholinguistics and psychology to assess how language models are impacted by imperfect RLHF techniques.

Ethics Statement

All data used in this study was either pre-existing or was generated from pre-existing language models. No human involvement was solicited, and no data is sensitive. LLM-generated stories are of no commercial value and do not compete with any human artists.

Reproducibility Statement

Our work makes use of publicly available datasets. All models were accessed via HuggingFace. We note all generations were conducted with a temperature ≥ 0, meaning results were subject to slight stochasticity.

Acknowledgments

We would like to thank Axel Bax, Federica Bologna, Kiara Liu, Andrew Piper, Rosamond Thalken, Andrea Wang, Matthew Wilkens, and Shengqi Zhu for their thoughtful feedback. This work was supported in part by NEH grant HAA-290374-23, AI for Humanists. We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC). Nous remercions le Conseil de recherches en sciences naturelles et en génie du Canada (CRSNG) de son soutien.

References

Marah I Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio CT Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Xin Wang, Rachel Ward, Yue Wu, Dingli Yu, Cyril Zhang, and Yi Zhang. Phi-4 Technical Report. Technical Report MSR-TR-2024-57, Microsoft, 2024. URL https://www.microsoft.com/en-us/research/publication/phi-4-technical-report/.

Mostafa Abdou, Artur Kulmizev, Daniel Hershcovich, Stella Frank, Ellie Pavlick, and Anders Søgaard. Can Language Models Encode Perceptual Structure Without Grounding? A Case Study in Color. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pp. 109–132, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.conll-1.9.

Muhannad Alomari, Paul Duckworth, David Hogg, and Anthony Cohn. Natural Language Acquisition and Grounding for Embodied Robotic Systems. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1), Feb. 2017. doi: 10.1609/aaai.v31i1.11161.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback, 2022.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, 46:904–911, 2014.

Rosario Caballero and Carita Paradis. Sharing Perceptual Experiences through Language. *Journal of Intelligence*, 11(7), 2023.

Marco Caracciolo and Karin Kukkonen. *With Bodies: Narrative Theory and Embodied Cognition*. Ohio State University Press, Columbus, Ohio, 2021.

Charles P Davis and Eiling Yee. Building semantic memory from embodied and distributional language experience. *Wiley Interdisciplinary Reviews: Cognitive Science*, 12(5), 2021.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423.

Guy Dove, Laura Barca, Luca Tummolini, and Anna M Borghi. Words have a weight: Language as a source of inner grounding and flexibility in abstract concepts. *Psychological Research*, 86(8):2451–2467, 2022.

Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical Neural Story Generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 889–898, Melbourne, Australia, 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1082.

Monika Fludernik. *Towards a 'Natural' Narratology*. Routledge, 1996.

Team Gemini. Gemini 2.0, 2025.

Team Gemini, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, 2024.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The Llama 3 Herd of Models, 2024.

Wes Gurnee and Max Tegmark. Language Models Represent Space and Time. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=jE8xbmvFin.

David Herman. *Basic Elements of Narrative*. Wiley-Blackwell, 2009.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. spacy: Industrial-strength natural language processing in python, 2020.

Xi Yu Huang, Krishnapriya Vishnubhotla, and Frank Rudzicz. The GPT-WritingPrompts Dataset: A Comparative Analysis of Character Portrayal in Short Stories, 2024.

Cameron R. Jones and Sean Trott. Multimodal Language Models Show Evidence of Embodied Simulation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 11928–11933, Torino, Italia, 2024. ELRA and ICCL.

Genji Kawakita, Ariel Zeleznikow-Johnston, Naotsugu Tsuchiya, and Masafumi Oizumi. Gromov–Wasserstein unsupervised alignment reveals structural correspondences between the color similarity structures of humans and large language models. *Scientific Reports*, 14(1), 2024.

Casey Kennington. Enriching Language Models with Visually-grounded Word Vectors and the Lancaster Sensorimotor Norms. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pp. 148–157, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.conll-1.11.

Bruce Lee and Jaehyuk Lim. Language Models Don't Learn the Physical Manifestation of Language. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3554–3579, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.195.

Jonghyun Lee, Dojun Park, Jiwoo Lee, Hoekeon Choi, and Sung-Eun Lee. Exploring multimodal perception in large language models through perceptual strength ratings, 2025.

Zhaowei Li, Qi Xu, Dong Zhang, Hang Song, YiQing Cai, Qi Qi, Ran Zhou, Junting Pan, Zefeng Li, Vu Tu, Zhida Huang, and Tao Wang. GroundingGPT: Language Enhanced Multi-modal Grounding Model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6657–6678, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.360.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A Robustly Optimized Bert Pretraining Approach. *CoRR*, 2019.

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

Carol Madden, Michel Hoen, and Peter Ford Dominey. A cognitive neuroscience perspective on embodied language for human–robot cooperation. *Brain and Language*, 112(3):180–188, 2010. doi: https://doi.org/10.1016/j.bandl.2009.07.001.

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), 2024.

Samuel Marks and Max Tegmark. The Geometry of Truth: Emergent Linear Structure in Large Language Model Representations of True/False Datasets. In *First Conference on Language Modeling*, 2024.

Lorenza Mondada. Language and the Sensing Body: How Sensoriality Permeates Syntax in Interaction. *Frontiers in Communication*, 6, 2021.

Emiko J Muraki, Laura J Speed, and Penny M Pexman. Insights into embodied cognition and mental imagery from aphantasia. *Nature Reviews Psychology*, 2(10):591–605, 2023.

Jerry Ngo and Yoon Kim. What Do Language Models Hear? Probing for Auditory Representations in Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5435–5448, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.297.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, et al. 2 OLMo 2 Furious, 2024.

OpenAI. GPT-4o System Card, 2024.

Cory Paik, Stéphane Aroca-Ouellette, Alessandro Roncone, and Katharina Kann. The World of an Octopus: How Reporting Bias Influences a Language Model's Perception of Color. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 823–835, Online and Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.63.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Andrew Piper. What Do Characters Do? The Embodied Agency of Fictional Characters. *Journal of Computational Literary Studies*, 2(1), 2024.

Andrew Piper and Sunyam Bagga. Using large language models for understanding narrative discourse. In *Proceedings of the The 6th Workshop on Narrative Understanding*, pp. 37–46, Miami, Florida, USA, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.wnu-1.4.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 Technical Report, January 2025.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language Models are Unsupervised Multitask Learners. *OpenAI Blog*, 1(8):24, 2018.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer, September 2023.

Xing Shi, Inkit Padhi, and Kevin Knight. Does String-Based Neural MT Learn Source Syntax? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1526–1534, Austin, Texas, 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1159.

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

Francesca Strik Lievers and Bodo Winter. Sensory language across lexical categories. *Lingua*, 204:45–61, 2018. ISSN 0024-3841. doi: https://doi.org/10.1016/j.lingua.2017.11.002.

Tadahiro Taniguchi, Takayuki Nagai, Tomoaki Nakamura, Naoto Iwahashi, Tetsuya Ogata, and Hideki Asoh and. Symbol Emergence in Robotics: A Survey. *Advanced Robotics*, 30 (11-12):706–728, 2016. doi: 10.1080/01691864.2016.1164622.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. BERT Rediscovers the Classical NLP Pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4593–4601, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1452.

Bodo Winter. *Sensory Linguistics: Language, perception and metaphor*. John Benjamins Publishing Company, 2019.

Chenyu Zhang, Benjamin Van Durme, Zhuowan Li, and Elias Stengel-Eskin. Visual Commonsense in Pretrained Unimodal and Multimodal Models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5321–5335, Seattle, United States, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.390.

Shu Zhong, Elia Gatti, Youngjun Cho, and Marianna Obrist. Exploring Human-AI Perception Alignment in Sensory Experiences: Do LLMs Understand Textile Hand?, 2024a.

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, 2024b.