LINEAR REPRESENTATIONS OF SENTIMENT IN LARGE LANGUAGE MODELS

Curt Tigges[*♣], **Oskar John Hollinsworth**[*♡], **Atticus Geiger**[♠★], **Neel Nanda**[◊]
[♣]EleutherAI Institute, [♡]SERI MATS, [♠]Stanford University, [★]Pr(Ai)^{2}R Group, [◊]Independent
[*]Equal primary authors (order random)

ABSTRACT

Sentiment is a pervasive feature in natural language text, yet it is an open question how sentiment is represented within Large Language Models (LLMs). In this study, we reveal that across a range of models, sentiment is represented linearly: a single direction in activation space mostly captures the feature across a range of tasks with one extreme for positive and the other for negative. Through causal interventions, we isolate this direction and show it is causally relevant in both toy tasks and real world datasets such as Stanford Sentiment Treebank. Through this case study we model a thorough investigation of what a single direction means on a broad data distribution.

We further uncover the mechanisms that involve this direction, highlighting the roles of a small subset of attention heads and neurons. Finally, we discover a phenomenon which we term the summarization motif: sentiment is not solely represented on emotionally charged words, but is additionally summarised at intermediate positions without inherent sentiment, such as punctuation and names. We show that in Stanford Sentiment Treebank zero-shot classification, 76% of above-chance classification accuracy is lost when ablating the sentiment direction, nearly half of which (36%) is due to ablating the summarized sentiment direction exclusively at comma positions.

1 INTRODUCTION

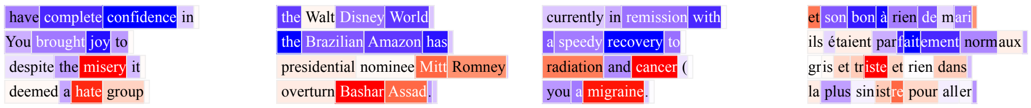
Large language models (LLMs) have displayed increasingly impressive capabilities (Brown et al., 2020; Radford et al., 2019; Bubeck et al., 2023), but their internal workings remain poorly understood. Nevertheless, recent evidence (Li et al., 2023) has suggested that LLMs are capable of forming models of the world, i.e., inferring hidden variables of the data generation process rather than simply modeling surface word co-occurrence statistics. There is significant interest (Christiano et al. (2021), Burns et al. (2022)) in deciphering the latent structure of such representations.

In this work, we investigate how LLMs represent sentiment, a variable in the data generation process that is relevant and interesting across a wide variety of language tasks (Cui et al., 2023). Approaching our investigations through the frame of causal mediation analysis (Vig et al., 2020; Pearl, 2022; Geiger et al., 2023a), we show that these sentiment features are represented linearly by the models, are causally significant, and are utilized by human-interpretable circuits (Olah et al., 2020; Elhage et al., 2021a).

We find the existence of a single direction scientifically interesting as further evidence for the linear representation hypothesis (Mikolov et al., 2013; Elhage et al., 2022)– that models tend to extract properties of the input and internally represent them as directions in activation space. Understanding the structure of internal representations is crucial to begin to decode them, and linear representations are particularly amenable to detailed reverse-engineering (Nanda et al., 2023b).

We show evidence of a phenomenon we have labeled the "summarization motif", where rather than sentiment being directly moved from valenced tokens to the final token, it is first aggregated on intermediate summarization tokens without inherent valence such as commas, periods and particular nouns.[1] This summarization structure for next token prediction can be seen as a naturally emerging

[1] Our use of the term "summarization" is distinct from typical NLP summarization tasks


have complete confidence in
You brought joy to
despite the misery it
deemed a hate group
the Walt Disney World
the Brazilian Amazon has
presidential nominee Mitt Romney
overturn Bashar Assad.
currently in remission with
a speedy recovery to
radiation and cancer (
you a migraine.
et son bon à rien de mari
ils étaient parfaitement normaux
gris et triste et rien dans
la plus sinistre pour aller

(a) Nouns (b) Proper Nouns (c) Medical (d) French

Figure 1: Visual verification that a single direction captures sentiment across diverse contexts. Color represents the projection onto this direction, blue is positive and red is negative. Examples (1a-1c) show the K-means sentiment direction for the first layer of GPT2-small on samples from OpenWebText. Example 1d shows the K-means sentiment direction for the 7th layer of pythia-1.4b on the opening of Harry Potter in French.

analogue to the explicit classification token in BERT-like models (Devlin et al., 2018). We show that the sentiment stored on summarization tokens is causally relevant for the final prediction. We find this an intriguing example of an "information bottleneck", where the data generation process is funnelled through a small subset of tokens used as information stores. Understanding the existence and location of information bottlenecks is a key first step to deciphering world models. This finding additionally suggests the models' ability to create summaries at various levels of abstraction, in this case a sentence or clause rather than a token.

Our contributions are as follows. In Section 3, we demonstrate methods for finding a **linear representation of sentiment** using a toy dataset and show that this direction correlates with sentiment information in the wild and matters causally in a crowdsourced dataset. In Section 4, we show through activation patching (Vig et al., 2020; Geiger et al., 2020) and ablations that the learned sentiment direction captures **summarization behavior** that is causally important to circuits performing sentiment tasks. Through this case study, we model an investigation of what a single interpretable direction means on the full data distribution.

2 Methods

2.1 Datasets and Models

ToyMovieReview A templatic dataset of continuation prompts we generated with the form

I thought this movie was ADJECTIVE, I VERBed it. Conclusion: This movie is

where ADJECTIVE and VERB are either two positive words (e.g., incredible and enjoyed) or two negative words (e.g., horrible and hated) that are sampled from a fixed pool of 85 adjectives (split 55/30 for train/test) and 8 verbs. The expected completion for a positive review is one of a set of positive descriptors we selected from among the most common completions (e.g. great) and the expected completion for a negative review is a similar set of negative descriptors (e.g., terrible).

ToyMoodStory A similar toy dataset with prompts of the form

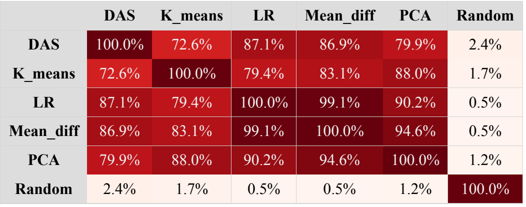
NAME1 VERB1 parties, and VERB2 them whenever possible. NAME2 VERB3 parties, and VERB4 them whenever possible. One day, they were invited to a grand gala. QUERYNAME feels very

To evaluate the model's output, we measure the logit difference between the "excited" and "nervous" tokens.

Stanford Sentiment Treebank (SST) SST Socher et al. (2013) consists of 10,662 one sentence movie reviews with human annotated sentiment labels for every phrase from every review.

OpenWebText OWT (Gokaslan & Cohen, 2019) is the pretraining dataset for GPT-2 which we use as a source of random text for correlational evaluations.

GPT-2 and Pythia (Radford et al., 2019; Biderman et al., 2023) These are families of decoder-only transformer models with sizes varying from 85M to 2.8b parameters. We use GPT2-small for movie review continuation, pythia-1.4b for classification and pythia-2.8b for multi-subject tasks.



	DAS	K_means	LR	Mean_diff	PCA	Random
DAS	100.0%	72.6%	87.1%	86.9%	79.9%	2.4%
K_means	72.6%	100.0%	79.4%	83.1%	88.0%	1.7%
LR	87.1%	79.4%	100.0%	99.1%	90.2%	0.5%
Mean_diff	86.9%	83.1%	99.1%	100.0%	94.6%	0.5%
PCA	79.9%	88.0%	90.2%	94.6%	100.0%	1.2%
Random	2.4%	1.7%	0.5%	0.5%	1.2%	100.0%

Figure 2: Cosine similarity of directions learned by different methods in GPT2-small's first layer. Each sentiment direction was derived from *adjective* representations in the ToyMovieReview dataset (Section 2.1).

2.2 Finding Directions

We use five methods to find a sentiment direction in each layer of a language model using our ToyMovieReview dataset. In each of the following, let $\mathbb{P}$ be the set of positive inputs and $\mathbb{N}$ be the set of negative inputs. For some input $x \in \mathbb{P} \cup \mathbb{N}$, let $\boldsymbol{a}_x^L$ and $\boldsymbol{v}_x^L$ be the vector in the residual stream at layer L above the adjective and verb respectively. We reserve $\{\boldsymbol{v}_x^L\}$ as a hold-out set for testing. Let the correct next token for $\mathbb{P}$ be p and for $\mathbb{N}$ be n.

Mean Difference (MD) The direction is computed as $\frac{1}{|\mathbb{P}|} \sum_{p \in \mathbb{P}} \boldsymbol{a}_p^L - \frac{1}{|\mathbb{N}|} \sum_{n \in \mathbb{N}} \boldsymbol{a}_n^L$.

K**-means (KM)** We fit 2-means to $\{\boldsymbol{a}_x^L : x \in \mathbb{P} \cup \mathbb{N}\}$, obtaining cluster centroids $\{\boldsymbol{c}_i : i \in [0,1]\}$ and take the direction $\boldsymbol{c}_1 - \boldsymbol{c}_0$.

Linear Probing The direction is the normed weights $\frac{\boldsymbol{w}}{||\boldsymbol{w}||}$ of a logistic regression (**LR**) classifier $\textbf{LR}(a_x^L) = \frac{1}{1+\exp(-\boldsymbol{w} \cdot \boldsymbol{a}_x^L)}$ trained to distinguish between $x \in \mathbb{P}$ and $x \in \mathbb{N}$.

Distributed Alignment Search (DAS) (Geiger et al., 2023b) The direction is a learned parameter θ where the training objective is the average logit difference

$$\sum_{x \in \mathbb{P}} [\text{logit}_\theta(x; p) - \text{logit}_\theta(x; n)] + \sum_{x \in \mathbb{N}} [\text{logit}_\theta(x; n) - \text{logit}_\theta(x; p)]$$

after patching using direction θ (see Section 2.3).

Principal Component Analysis (PCA) The direction is the first component of $\{\boldsymbol{a}_x^L : x \in \mathbb{P} \cup \mathbb{N}\}$.

2.3 Causal Interventions

Activation Patching In activation patching (Geiger et al., 2020; Vig et al., 2020), we create two symmetrical datasets, where each prompt x_{orig} and its counterpart prompt x_{flipped} are of the same length and format but where key words are changed in order to flip the sentiment; e.g., "This movie was great" could be paired with "This movie was terrible." We first conduct a forward pass using x_{orig} and capture these activations for the entire model. We then conduct forward passes using x_{flipped}, iteratively patching in activations from the original forward pass for each model component. We can thus determine the relative importance of various parts of the model with respect to the task currently being performed. Geiger et al. (2023b) introduce distributed interchange interventions, a variant of activation patching that we call "directional activation patching". The idea is that rather than modifying the standard basis directions of a component, we instead only modify the component along a single direction in the vector space, replacing it during a forward pass with the value from a different input.

We use two evaluation metrics. The logit difference (difference in logits for correct and incorrect answers) metric introduced in Wang et al. (2022), as well as a "logit flip" accuracy metric (Geiger et al., 2022), which quantifies the proportion of cases where we induce an inversion in the predicted sentiment.