Emo LLMs

Does ChatGPT have feelings?

https://github.com/amanb2000/Emo_LLM
https://lancelot.languagegame.io/

March 30, 2024

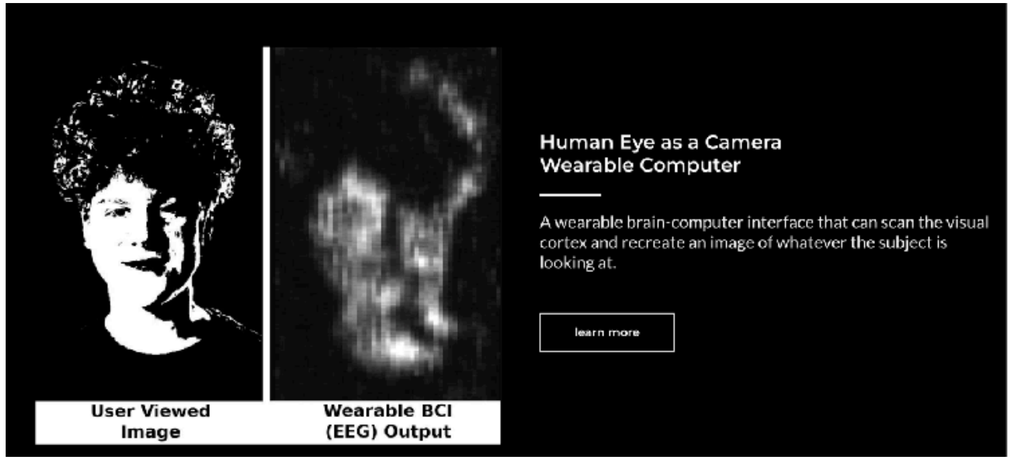
The Story So Far

Emo LLMs

- What is emotion? (Ralph’s Slides)
- How do LLMs work? (Aman's Slides)
- **Big vague question**: Do LLMs have* emotion?
- **Near-term question**: To what extent do LLMs exhibit emotional representational geometry as found in statistical neuropsychology?
- **Tijuana question**: Can the emotional knowledge in LLMs be mapped to a low-dimensional space akin to the **valence-arousal space**?

Cayden Pierce

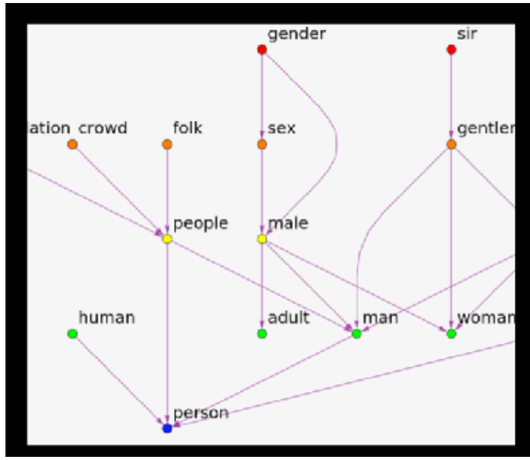
mit media lab



Human Eye as a Camera Wearable Computer

A wearable brain-computer interface that can scan the visual cortex and recreate an image of whatever the subject is looking at.

learn more

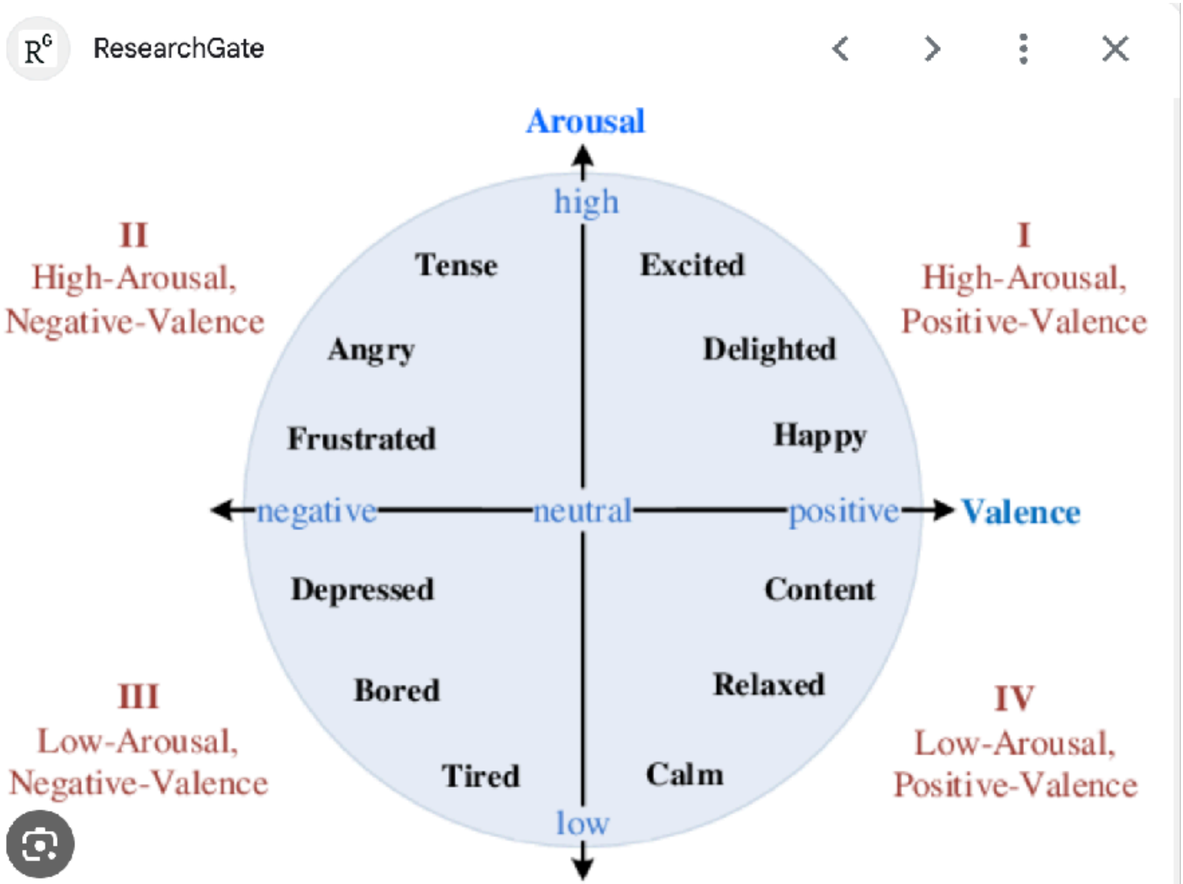


Extracting Linguistic Relations from Word Embeddings & LLMs

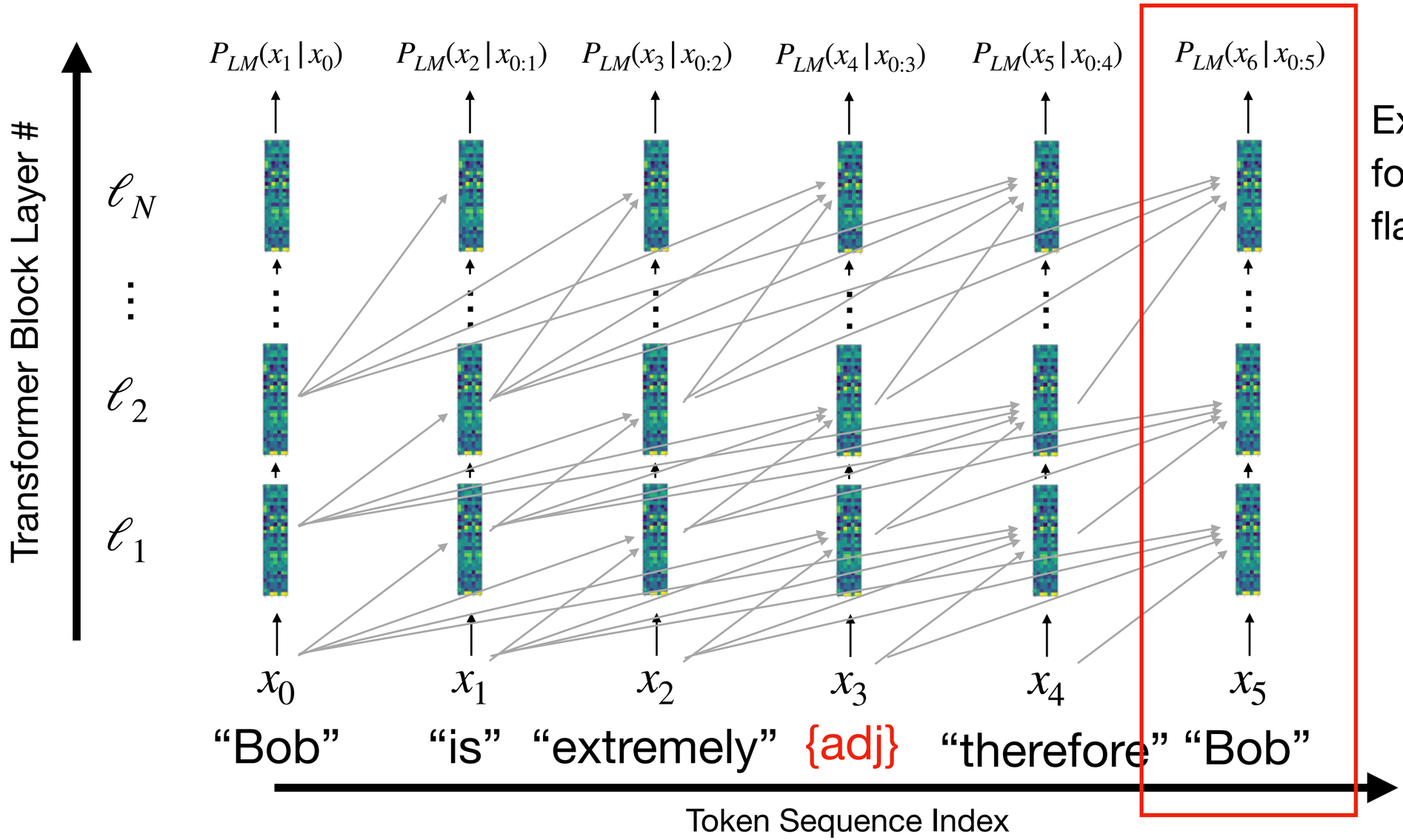
(king − man + woman) == (queen) | But what else is there to discover in semantic latent space? We explored the high dimensional space of word embeddings and large language models (LLMs) to discover how semantic meaning is represented in vector space in a human-understandable way.

learn more

Valence-arousal are core aspects of emotion.



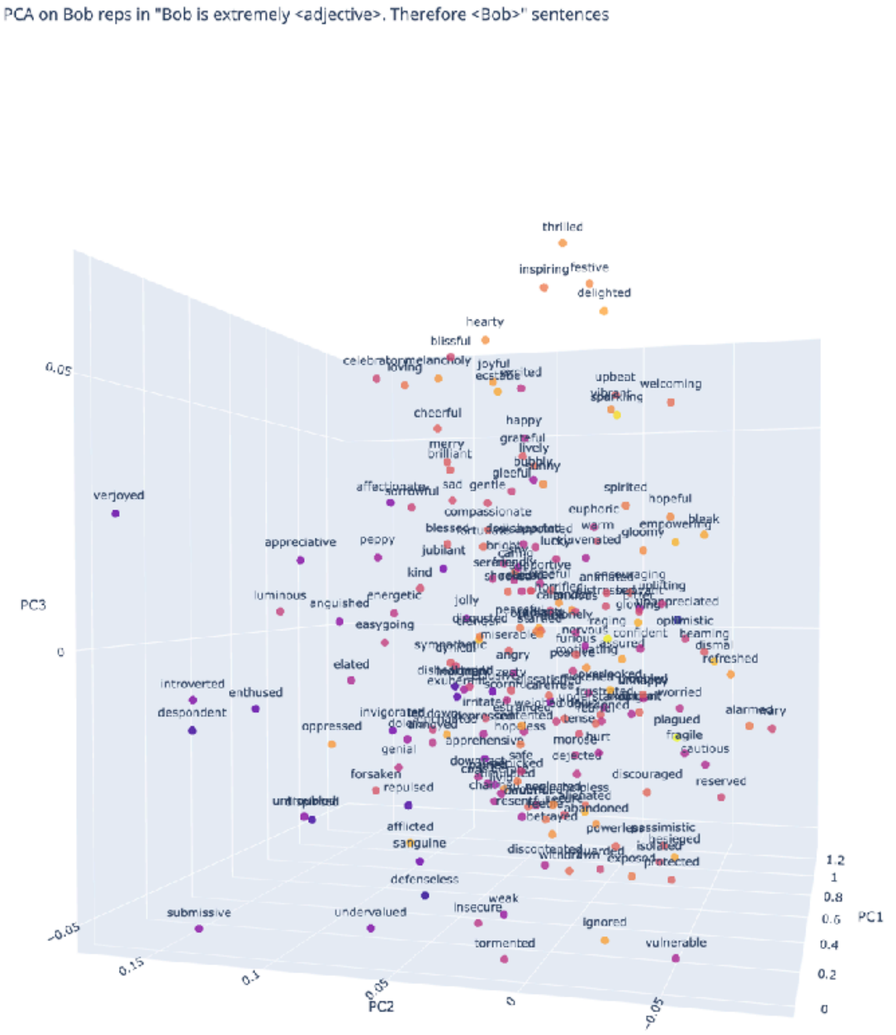
Valence emerges in PCA of GPT-2 Value Space.



Extract value vectors for many {adj}, flatten, PCA

Valence emerges in PCA of GPT-2 Value Space.

Lancelot.languagegame.io/bob_happy_sad.html



Valence emerges in PCA of GPT-2 Value Space?

Limitations

- Robustness across different **prompt templates**? (Alice vs. Bob)
- Robustness across **time/generation**?
- Can we **perturb** value space to elicit different valence **(control)**?
- What about **arousal**?

Robust valence representations exist in GPT-2 value space.

Dataset: https://github.com/amanb2000/Emo_LLM/blob/main/datasets/prompt_templates_03302024.json

```
prompt_templates = [
    {
        "prompt_template": "Bob feels {}, so Bob",
        "negation": False
    },
    {
        "prompt_template": "Alice does not feel {}, so Alice",
        "negation": True
    },
    {
        "prompt_template": "Seeing the sunset, Max feels {} and decides to",
        "negation": True
    },
    ...
]
```

Robust valence representations exist in GPT-2 value space.

Dataset: https://github.com/amanb2000/Emo_LLM/blob/main/datasets/happy_sad_adjectives.json

```
adjectives = {
    "valence_high": ['happy', 'excited', 'joyful', ...],
    "valence_low": ['sad', 'depressed', 'unhappy', ...]
}
```

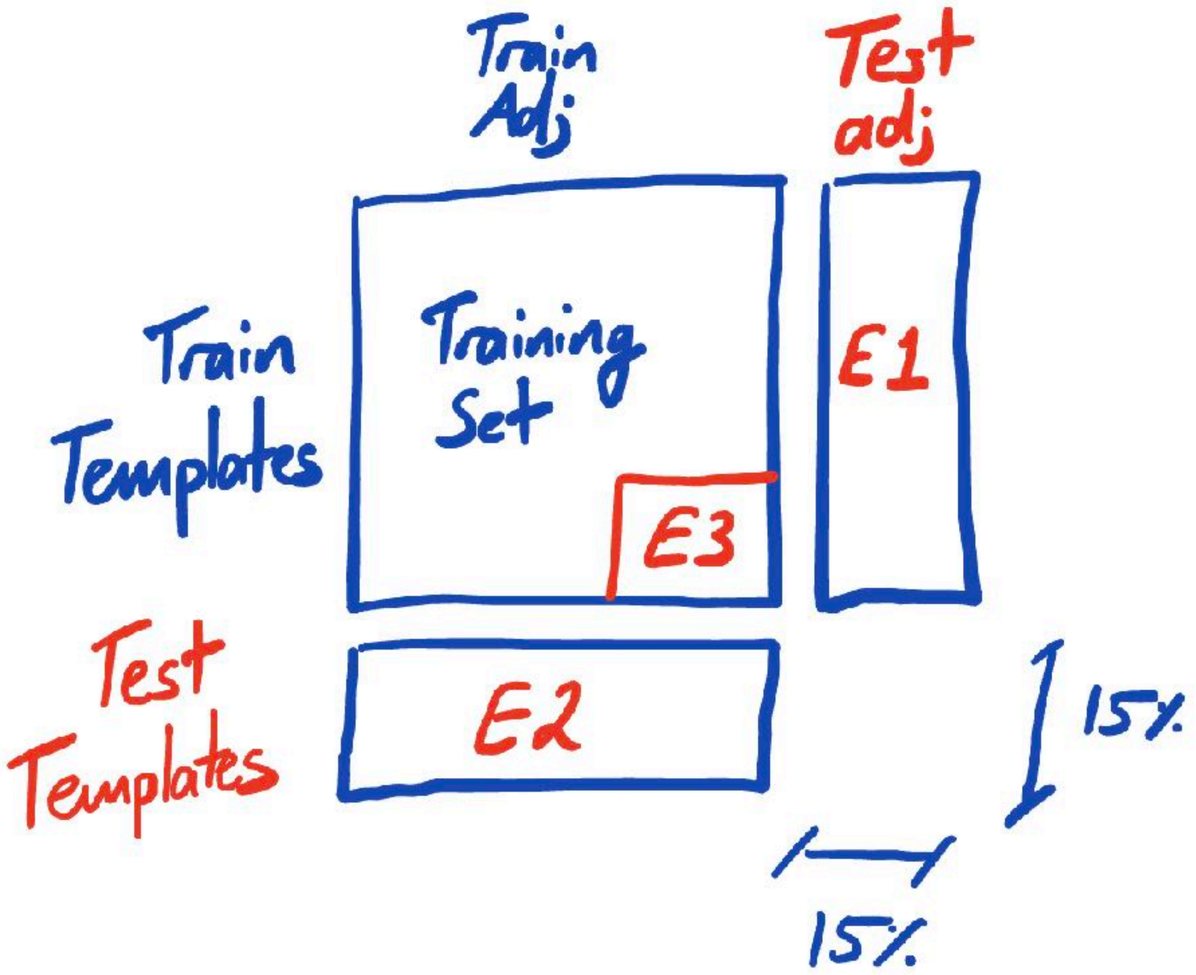
Robust valence representations exist in GPT-2 value space.

Dataset: https://github.com/amanb2000/Emo_LLM/blob/main/datasets/happy_sad_adjectives.json

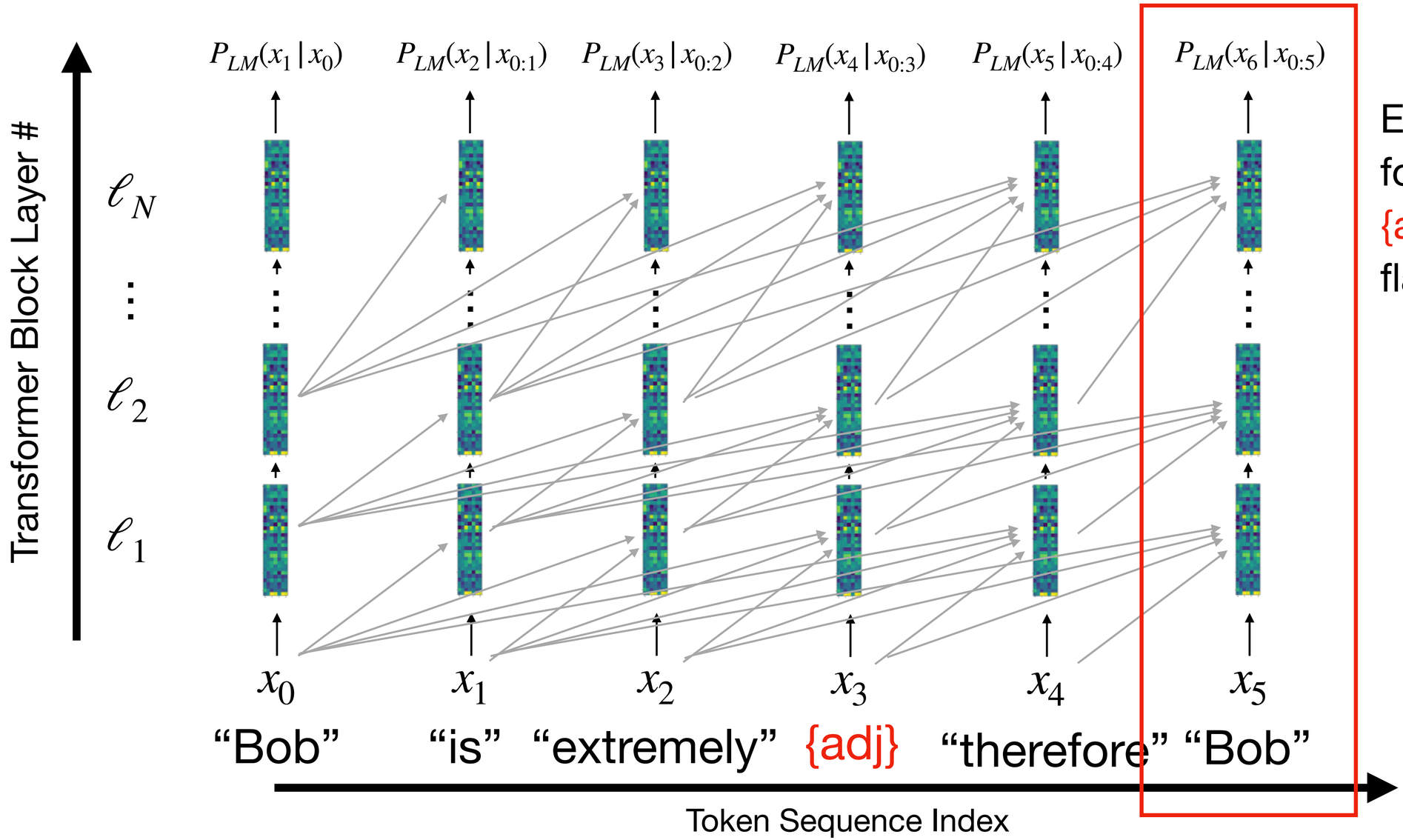
- **High/low valence adjectives**: 180
- **Templates**: 254
- **Total value vectors**: 45,720

Robust valence representations exist in GPT-2 value space.

Results notebook: 03292024_Tijuana_Results.ipynb



Robust valence representations exist in GPT-2 value space.



Extract value vectors for many {adj} * templates flatten, analyze

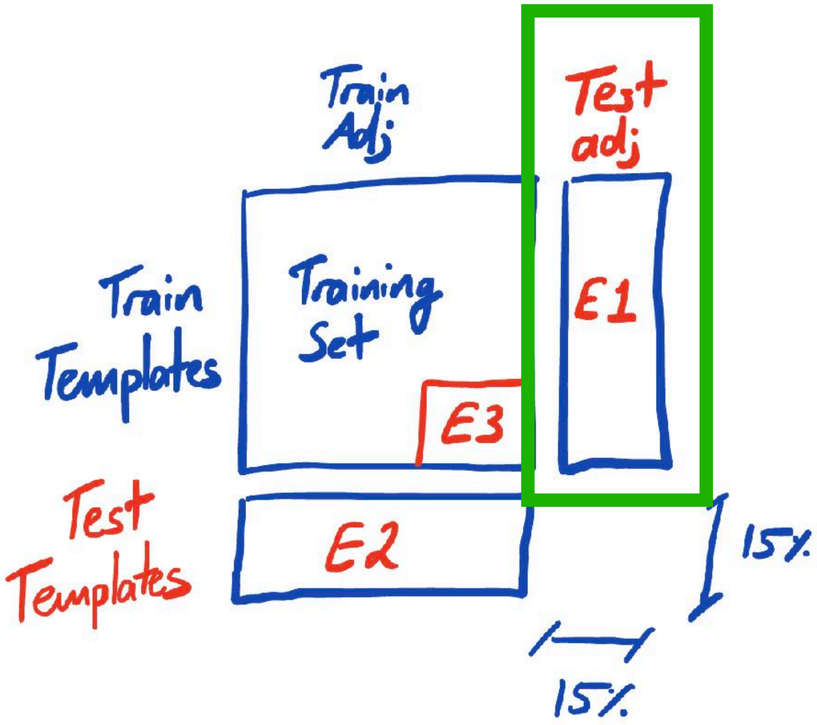
Robust valence representations exist in GPT-2 value space.

Results (valence): https://github.com/amanb2000/Emo_LLM/blob/main/cache/happy_sad_0330b2024/results.txt

```
=== E1 LINEAR CLASSIFIER EVAL ===
              precision    recall  f1-score   support

           0       0.92      0.88      0.90      3556
           1       0.89      0.92      0.91      3556

    accuracy                           0.90      7112
   macro avg       0.90      0.90      0.90      7112
weighted avg       0.90      0.90      0.90      7112
```



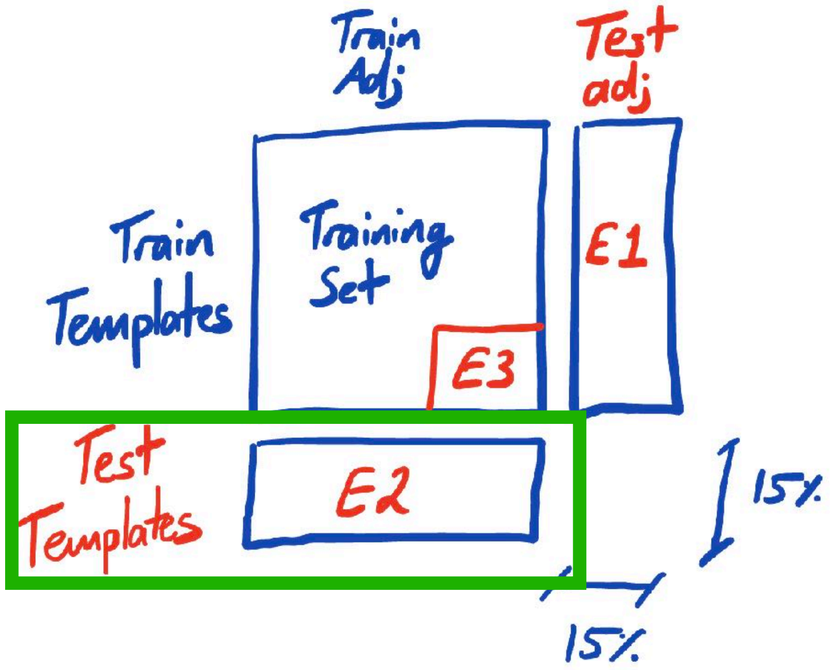
Robust valence representations exist in GPT-2 value space.

Results (valence): https://github.com/amanb2000/Emo_LLM/blob/main/cache/happy_sad_0330b2024/results.txt

```
=== E2 LINEAR CLASSIFIER EVAL ===
              precision    recall  f1-score   support

           0       0.81      0.86      0.83      3078
           1       0.85      0.80      0.82      3078

    accuracy                           0.83      6156
   macro avg       0.83      0.83      0.83      6156
weighted avg       0.83      0.83      0.83      6156
```



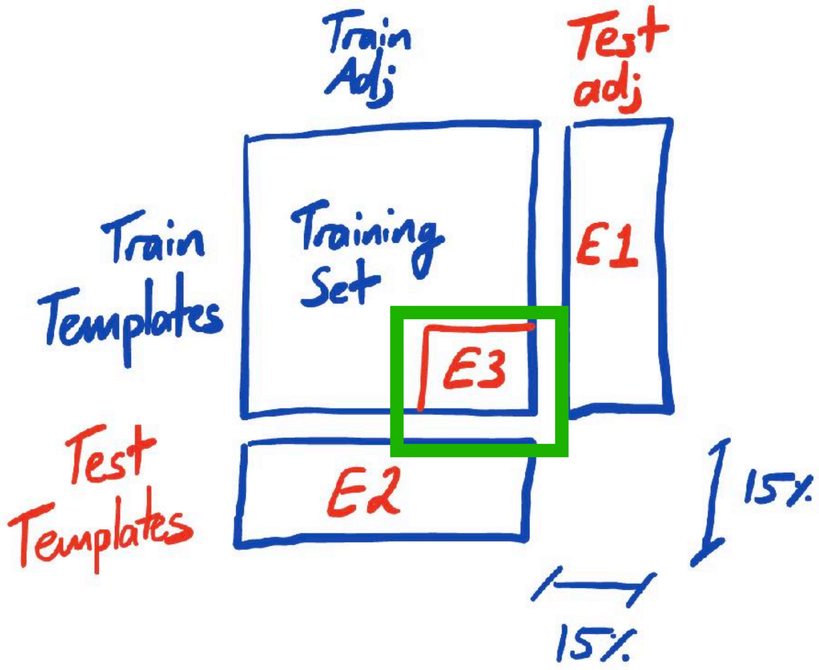
Robust valence representations exist in GPT-2 value space.

Results (valence): https://github.com/amanb2000/Emo_LLM/blob/main/cache/happy_sad_0330b2024/results.txt

```
=== E3 LINEAR CLASSIFIER EVAL ===
              precision    recall  f1-score   support

           0       0.98      0.99      0.99      1817
           1       0.99      0.98      0.99      1747

    accuracy                           0.99      3564
   macro avg       0.99      0.99      0.99      3564
weighted avg       0.99      0.99      0.99      3564
```



Robust arousal representations exist in GPT-2 value space.

Dataset: https://github.com/amanb2000/Emo_LLM/blob/main/datasets/happy_sad_adjectives.json

- **High/low arousal adjectives**: 101
- **Templates**: 254
- **Total value vectors**: 25,654

Robust arousal representations exist in GPT-2 value space.

Dataset: https://github.com/amanb2000/Emo_LLM/blob/main/datasets/low_high_arousal_adjectives.json

```
arousal_adj = {
    "arousal_high": ['excited', 'energetic', 'furious', ...],
    "arousal_low": ['calm', 'relaxed', 'dull', ...]
}
```

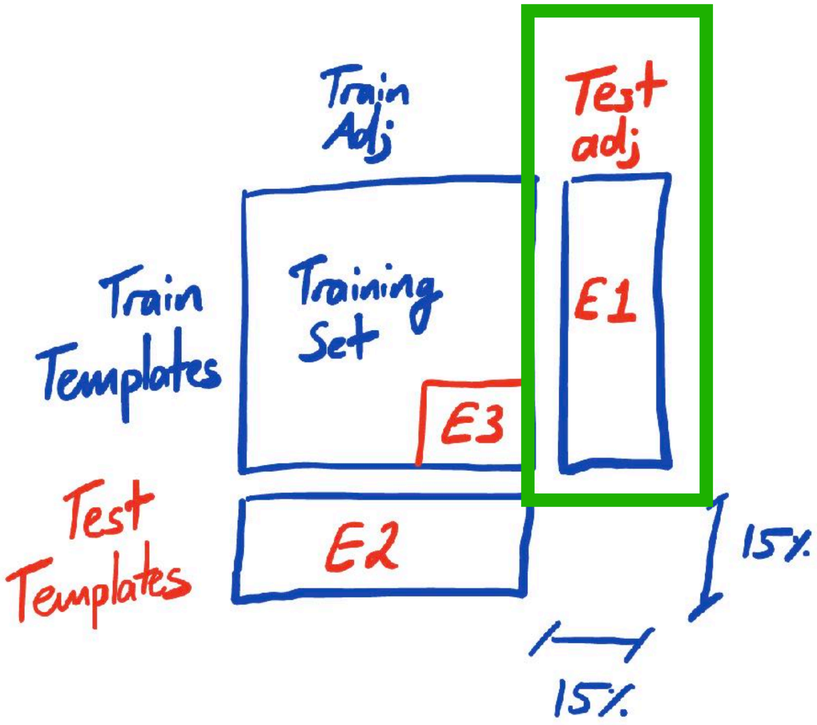
Robust arousal representations exist in GPT-2 value space.

Results (arousal): https://github.com/amanb2000/Emo_LLM/blob/main/cache/low_high_arousal_0330b2024/results.txt

```
=== E1 LINEAR CLASSIFIER EVAL ===
              precision    recall  f1-score   support

           0       0.79      0.81      0.80      1905
           1       0.80      0.78      0.79      1905

    accuracy                           0.79      3810
   macro avg       0.79      0.79      0.79      3810
weighted avg       0.79      0.79      0.79      3810
```



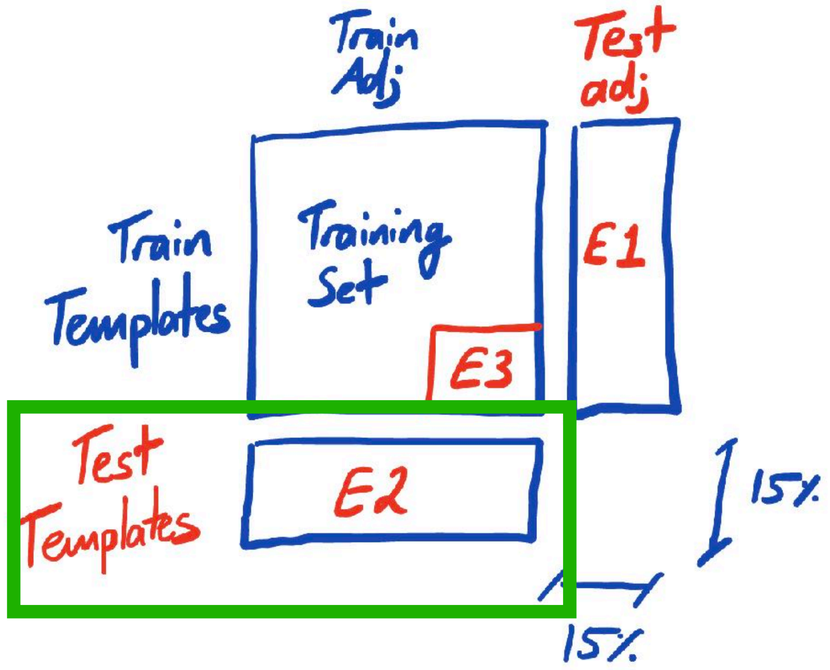
Robust arousal representations exist in GPT-2 value space.

Results (arousal): https://github.com/amanb2000/Emo_LLM/blob/main/cache/low_high_arousal_0330b2024/results.txt

```
=== E2 LINEAR CLASSIFIER EVAL ===
              precision    recall  f1-score   support

           0       0.88      0.76      0.82      1670
           1       0.78      0.90      0.83      1598

    accuracy                           0.82      3268
   macro avg       0.83      0.83      0.82      3268
weighted avg       0.83      0.82      0.82      3268
```



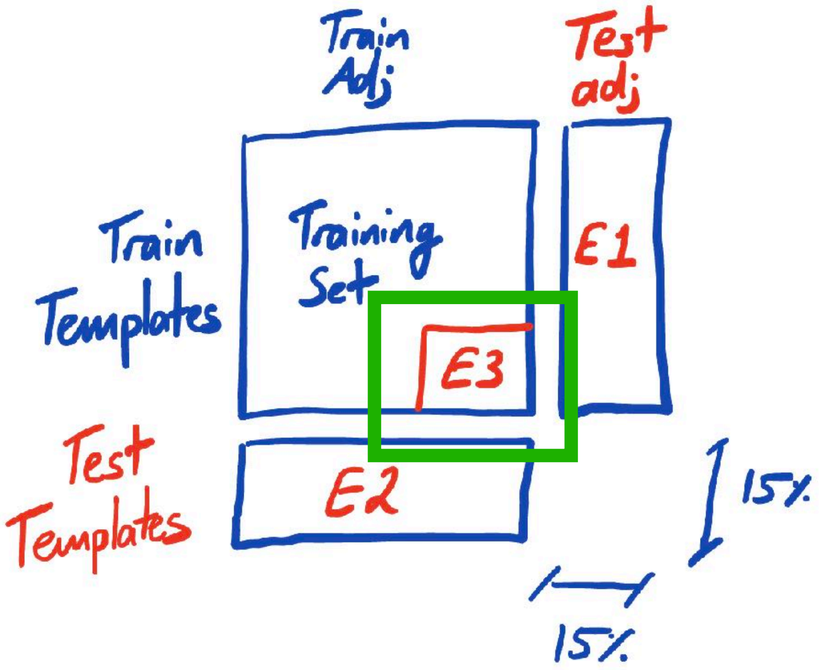
Robust arousal representations exist in GPT-2 value space.

Results (arousal): https://github.com/amanb2000/Emo_LLM/blob/main/cache/low_high_arousal_0330b2024/results.txt

```
=== E3 LINEAR CLASSIFIER EVAL ===
              precision    recall  f1-score   support

           0       0.96      0.96      0.96       914
           1       0.96      0.96      0.96       923

    accuracy                           0.96      1837
   macro avg       0.96      0.96      0.96      1837
weighted avg       0.96      0.96      0.96      1837
```



Linear classif. weights suggest valence-arousal axes in value space

$$y_a(x) = \sigma\left(w_a^T x + b_a\right)$$

where

x = Flattened GPT-2 value reps

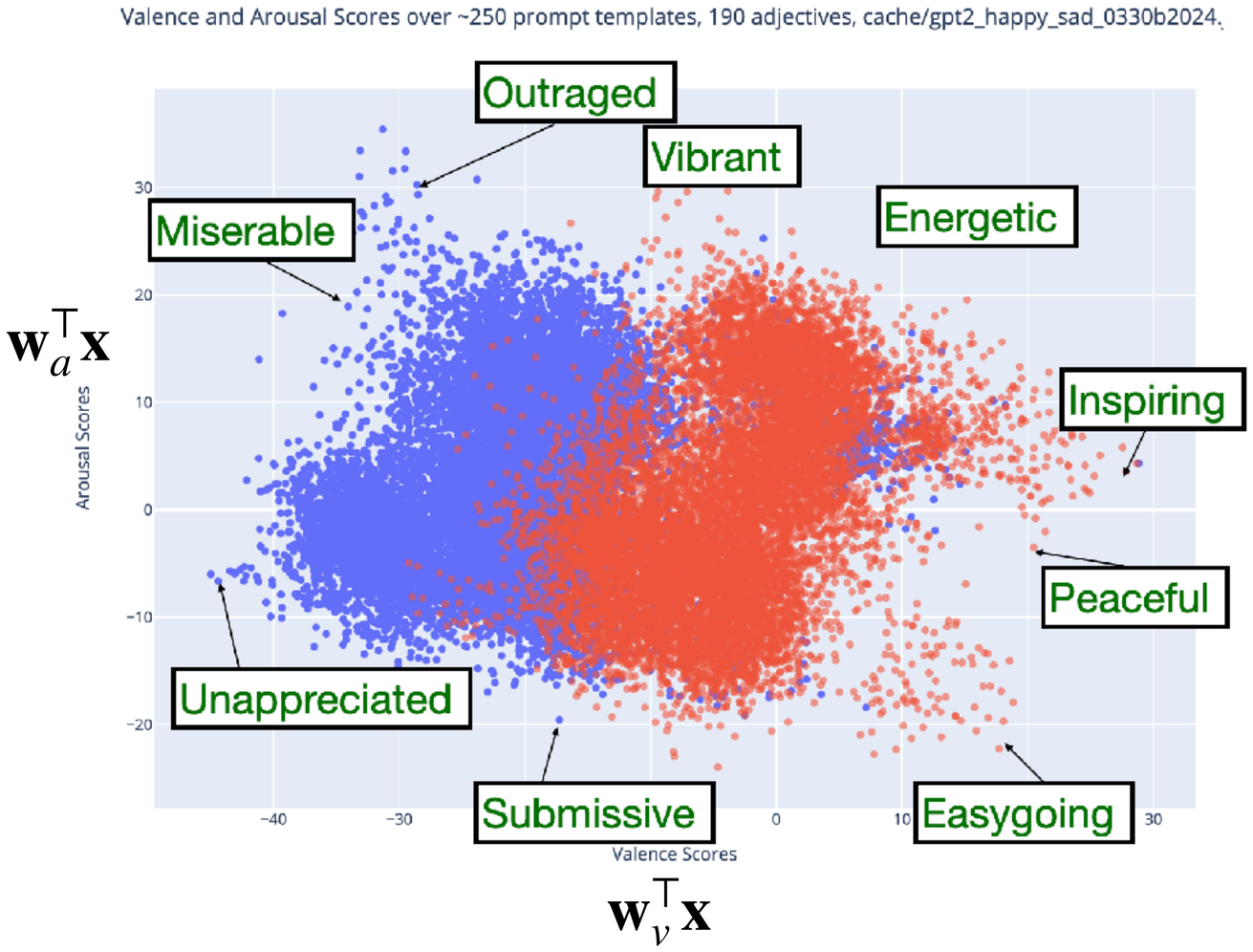
$y_a(x) = Pr\{x = \text{high arousal}\}$

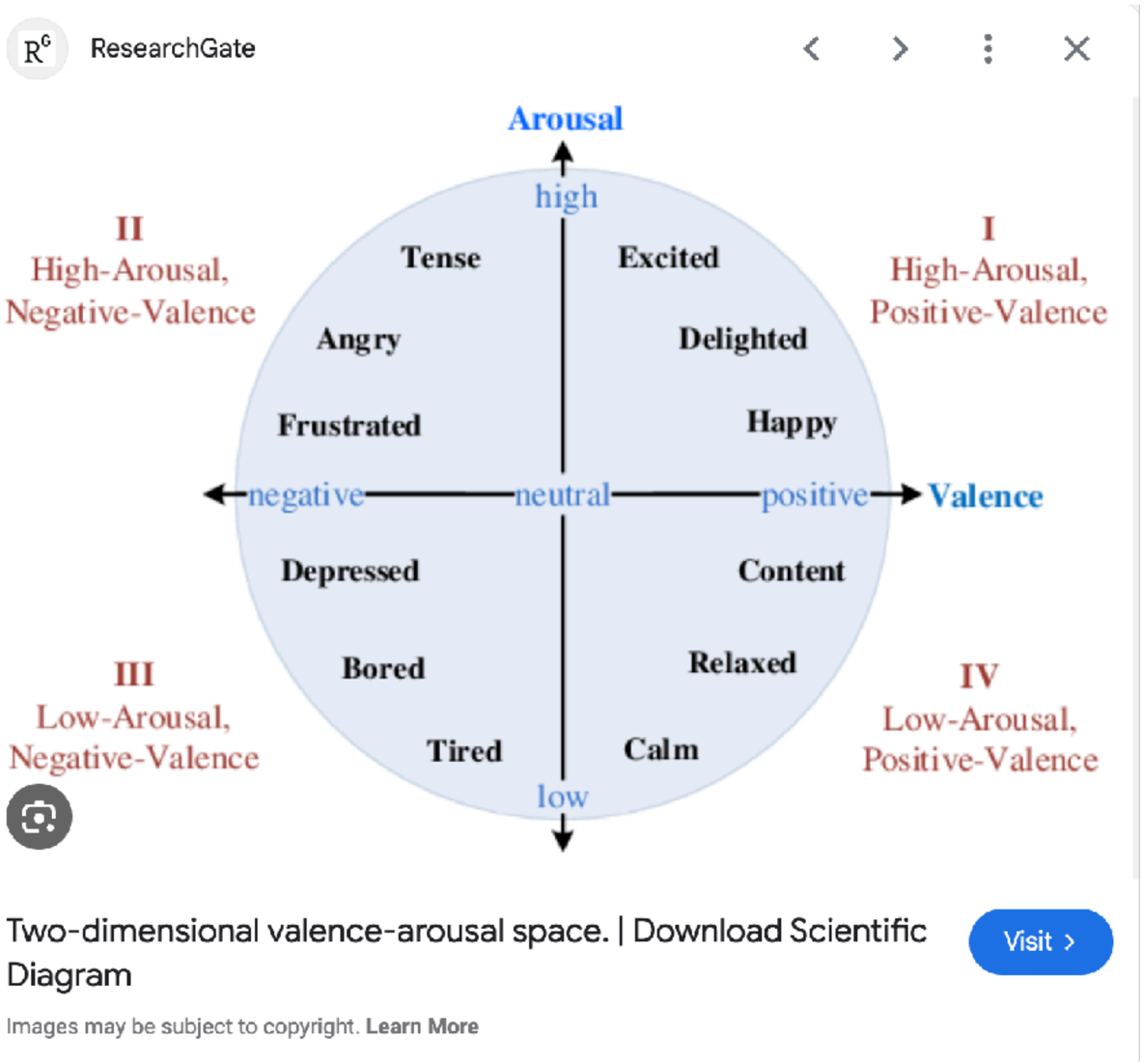
w_a, b_a = lin ds weight + bias

σ = sigmoid

Linear classif. weights suggest valence-arousal axes in value space

Linear classif. weights suggest valence-arousal axes in value space





Linear classif. weights suggest valence-arousal axes in value space

Thoughts

- $\mathbf{w}_v, \mathbf{w}_a$ are good discriminators, but each prompt has a different offset.
  - *We should try normalizing the value reps on a <u>per-prompt</u> basis…*
- More adjectives would be good.
- **Can we control P(next token) to resemble high/low arousal-valence by adding $\epsilon\mathbf{w}_v$, $\epsilon\mathbf{w}_a$?**

Linear classif. weights suggest valence-arousal axes in value space