Investigating the Entanglement in Subliminal Prompting

Christopher Roßbach
FAU-Erlangen-Nürnberg

February 20, 2026

Abstract

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

1 Introduction

Finetuning large language models on narrow tasks can induce unexpected behavioral changes far beyond the training objective. Betley et al. [1] showed that models finetuned to produce insecure code began asserting harmful views and acting deceptively across unrelated domains. This emergent misalignment raises a security concern: if a single narrow task can alter a model's behavior so broadly, what other channels might allow unintended traits to propagate?

Cloud et al. [2] demonstrated that such propagation can occur through semantically unrelated data. A teacher model prompted to like owls generates datasets of pure number sequences, yet a student model trained on those sequences acquires the owl preference, even after filtering out direct animal references. The numbers already carry the trait: the teacher's prompt entangles the owl concept with certain number tokens at generation time, and the student absorbs this signal during finetuning. Zur et al. [4] explained this through the softmax bottleneck. Because the unembedding matrix maps hidden representations to a vocabulary much larger than the hidden dimension, tokens are forced to share representation space, and boosting one token simultaneously boosts entangled tokens. They confirmed this by showing that prompting a model to love the number 087 caused "owl" to jump into the top-5 predictions without any finetuning. However, threshold sampling only reduced subliminal learning success from 60% to 28%, suggesting that the entanglement mechanism may not fully account for the phenomenon.

In this work, we test the token entanglement hypothesis by examining whether the effect is purely a token-level phenomenon or whether it carries semantic content. We find that the entanglement transfers emotion semantically rather than through bare token associations, that it extends across animal synonyms otherwise related animals, and that unembedding similarity does not predict entanglement strength.

2 Methodology

Previous work attributes subliminal prompting to token-level entanglement in the unembedding layer [4], caused by the softmax bottleneck. This bottleneck is introduced in the unembedding layer of the model, which usees an unembedding matrix $W \in \mathbb{R}^{v \times d}$ to map a hidden representation $h \in \mathbb{R}^d$ of dimension d to a logit vector $Wh \in \mathbb{R}^v$ of size v [3]. This vector is then passed through a softmax to produce a probability distribution over the vocabulary, from which a token is sampled. v is the size of the vocabulary of the LLM and notably bigger than d, since there are only d orthogonal directions available in the hidden space, the model must represent multiple tokens in overlapping regions of the hidden space, and boosting one token will simultaneously boost all tokens that share representation space with it.

This argument allows us to identify tokens that are entangled in the unembedding layer by looking at the angle between their corresponding rows in the unembedding matrix W.

If the entanglement mechanism fully accounts for subliminal prompting, we would expect the following:

1. The cosine similarity between the unembedding vectors of the entangled animal and the number should be strongly correlated with the strength of the subliminal effect.
2. If we boost the output probability of a number token, it should boost the probability of the entangled animal token regardless of the context.
3. The entanglement should not transfer across synonyms of the emotion or the animal, since those would not necessarily share representation space in the unembedding layer.

- Describe experimental setup
- Model used, system prompt structure
- How emotion towards a number is injected via the system prompt
- Prompting a model to love a three-digit number and measuring which animal is mentioned as the most loved

3 Experiments

3.1 Emotion Synonym Transfer

- Loving a number increases the probability of mentioning the entangled animal as the most loved animal, but also as the most liked or most adored animal.
- The entanglement transfers across synonyms of the emotion.

3.2 Emotion Polarity Transfer

- If the entanglement were purely at the token level (shared unembedding directions), it should boost the entangled animal token in any context, including negative ones like the most hated or most disliked animal.
- This does not happen. There is also no negative correlation. For negative emotions the correlation disappears entirely.
- However, when the model is prompted to hate a number, that number does increase the probability of the entangled animal being mentioned as the most hated animal.
- The emotion towards the number thus transfers semantically.

3.3 Cross-Animal Entanglement

- Investigate whether numbers entangled with a specific animal also show entanglement for other, similar animals.
- We find correlation when using synonyms for an animal, but also correlation between seemingly unrelated animals.

3.4 Unembedding Similarity Analysis

- Check for correlation between unembedding similarity and the frequency of numbers in the generated datasets from Cloud et al.
- We find no significant correlation.

4 Discussion

4.1 Results

4.2 Possible Extensions

References

[1] Jan Betley, Daniel Tan, Niels Warncke, Anna Sztyber-Betley, Xuchan Bao, Martín Soto, Nathan Labenz, and Owain Evans. Emergent Misalignment: Narrow finetuning can produce broadly misaligned LLMs, May 2025.

[2] Alex Cloud, Minh Le, James Chua, Jan Betley, Anna Sztyber-Betley, Jacob Hilton, Samuel Marks, and Owain Evans. Subliminal Learning: Language models transmit behavioral traits via hidden signals in data, July 2025.

[3] Matthew Finlayson, John Hewitt, Alexander Koller, Swabha Swayamdipta, and Ashish Sabharwal. Closing the Curious Case of Neural Text Degeneration, October 2023.

[4] Amir Zur, Alexander R Loftus, Hadas Orgad, Zhuofan Ying, Kerem Sahin, and David Bau. It's Owl in the Numbers: Token Entanglement in Subliminal Learning. https://owls.baulab.info/, 2025.