# Simple Idea Discovery in a Minimalist LLM Architecture Implementation

Robert Chihaia[1,2], Maria Trocan[1], and Florin Leon[2]
[1]LISITE Lab, Isep, Paris, France, 28, rue Notre-Dame-des-Champs, 75006 Paris, France, maria.trocan@isep.fr
[2]Faculty of Automatic Control and Computer Engineering, "Gheorghe Asachi" Technical University of Iași, Bd. Mangeron 27, 700050 Iași, Romania, robert.chihaia@student.tuiasi.ro, florin.leon@academic.tuiasi.ro

## Context

**Keywords** —byte-pair encoding, tokenizer, minimalist LLM, sentiment spectrum, idea discovery

Large Language Models (LLMs) show an implicit grasp of high-level concepts, but their immense scale makes it difficult to understand how these "ideas" are encoded. This research investigates whether abstract concepts are genuinely represented in a model's internal states by creating a deliberately minimalist LLM to expose its internal mechanisms.

**Methodology:** A small-scale, 30.7 million parameter, decoder-only Transformer was built and trained, along with a custom BPE tokenizer to isolate and examine conceptual structures without the overhead of massive models.

**Model Architecture:**

- Custom Byte-Pair Encoding (BPE) with a 10,256-token vocabulary
- 8-block decoder-only Transformer with a 512-token context window.
- Each block contains causal multi-head self-attention (with FlashAttention) and a feed-forward network
- 8 attention heads in a 512-dimensional embedding space
- Trained on approximately 11 billion tokens from the FineWeb corpus

**Objective:** To trace how semantically related inputs map onto a model's hidden states, providing a reproducible framework for studying semantic abstraction.

## Methodology

To probe for sentiment, the model was fed nine sentences sharing a fixed template where only a single sentiment adjective was changed.

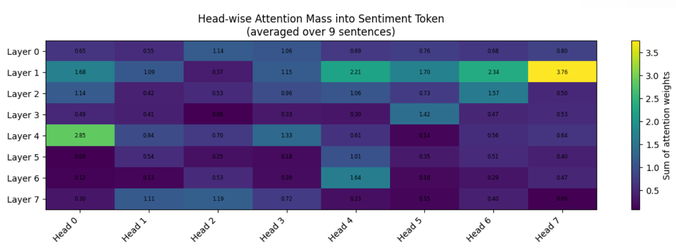
- **Template:** "The team's performance was **[adj]**".
- **Sentiment Continuum:** The adjectives were chosen to be single tokens and span a spectrum from negative to positive:

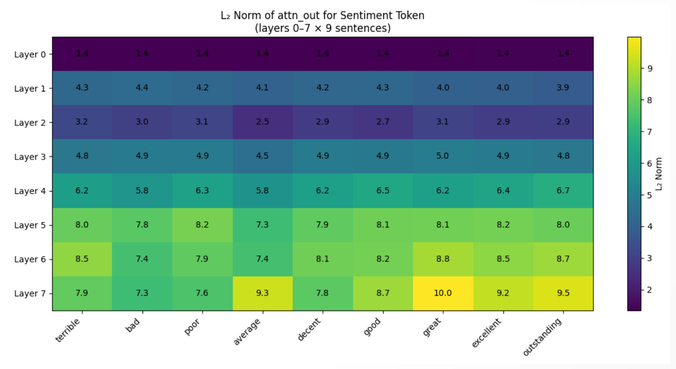
$$terrible \rightarrow bad \rightarrow poor \rightarrow average \rightarrow decent \rightarrow good \rightarrow great \rightarrow excellent \rightarrow outstanding$$

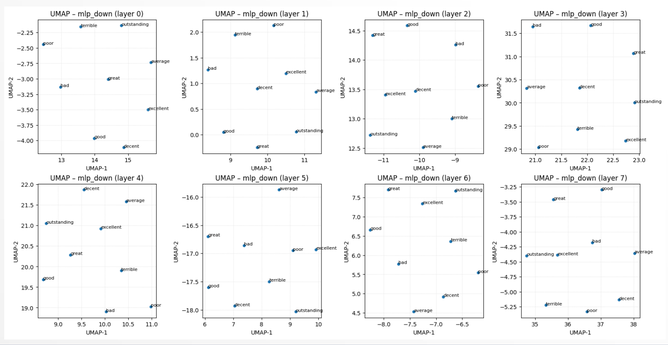
- **Activation Capture:** Forward hooks were placed on decoder blocks to capture internal activations from the attention mechanism and MLP layers for analysis.
- **Analysis:** Principal Component Analysis (*PCA*), *t-SNE* and *UMAP* were used to visualize the high-dimensional activation data in 2D space.

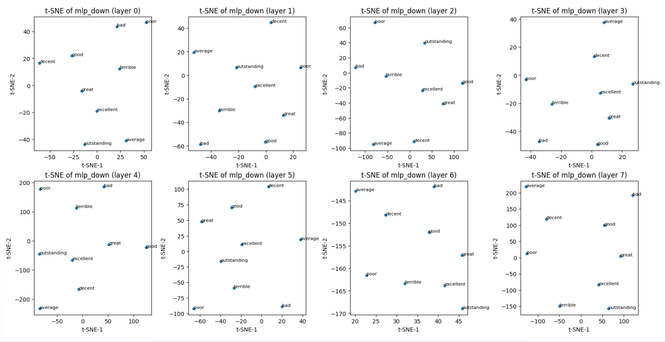
## Experimental Framework

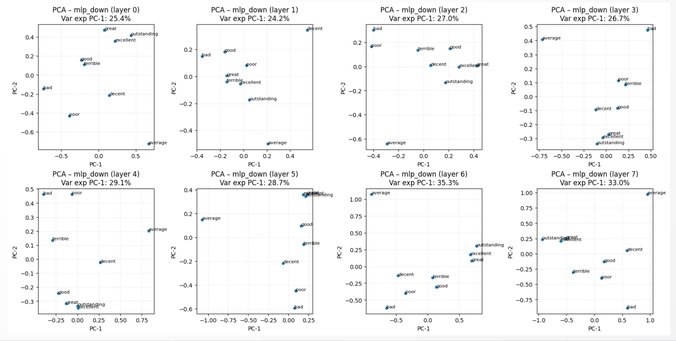
- Specific attention heads are responsible for identifying sentiment in a layered process. An early signal from head 7 in Layer 1 primes the network, leading to a sharp, decisive signal from head 0 in Layer 4. In later layers, the task is distributed across multiple heads for a more robust representation, showing that sentiment detection is a specialized, multi-stage process rather than a uniform one.
- The attention mechanism progressively amplifies the sentiment signal. After weak signals in early layers, a turning point occurs at Layer 4 where signal strength increases significantly. By the final layers, positive adjectives elicit a much stronger response than negative or neutral ones, demonstrating how the model refines and strengthens the sentiment feature before it is passed to the MLP layers.
- Visualizations show a clear narrative of concept formation. Initially unstructured, the MLP activations begin to self-organize in the middle layers. By Layer 5 and 6, PCA reveals that the activations align along a nearly straight line that corresponds to the human-defined sentiment continuum. This "dimensional collapse" creates a low-dimensional "sentiment axis." In the final layers, this fine-grained axis dissolves as activations cluster by final polarity (positive/negative), indicating the task is complete.




## Conclusion

This work successfully tracked how a Transformer reshapes token-level statistics into an abstract concept across its layers

- **Gradual Formation:** Concepts are not present initially but emerge through a gradual alignment process, sharpening in the middle layers.
- **Dimensional Collapse:** The model learns to compress token-level details into a low-dimensional subspace where an axis encodes an abstract property like sentiment polarity.
- **Interpretability Strategy:** The findings suggest that interpretability research can benefit from a "telescopic strategy": using small, transparent models to understand fundamental mechanisms before analysing larger, more complex systems.

International Workshop on Distributed and Parallel Programming for Extreme-scale AI
July 16-17th, 2025 | Paris, France