LLaMAs Have Feelings Too: Unveiling Sentiment and Emotion Representations in LLaMA Models Through Probing

Dario Di Palma, Alessandro De Bellis, Giovanni Servedio, Vito Walter Anelli, Fedelucio Narducci, Tommaso Di Noia

Scan for PDF

Abstract

What is this study about?

We **investigate** how Large Language Models (LLMs), specifically **LLaMA** models, encode **sentiment** within their hidden layers.

What did we do?

We systematically **probe** the **hidden states** across all layers and scales of LLaMA models, comparing different pooling strategies to **identify where sentiment** information **is** most **strongly represented**.

Key Findings:

- Sentiment signals are most concentrated in **mid-layers**, especially for binary sentiment classification.
- Our method outperforms prompting-based approaches, with up to **+14% accuracy gains**.
- The **last token is not always the most informative** in decoder-only models.
- We achieve sentiment classification with **57% lower memory usage** on average.

Why it matters:

Understanding where sentiment lives inside LLMs leads to more **efficient**, **accurate**, and **interpretable** sentiment analysis, without full model fine-tuning.

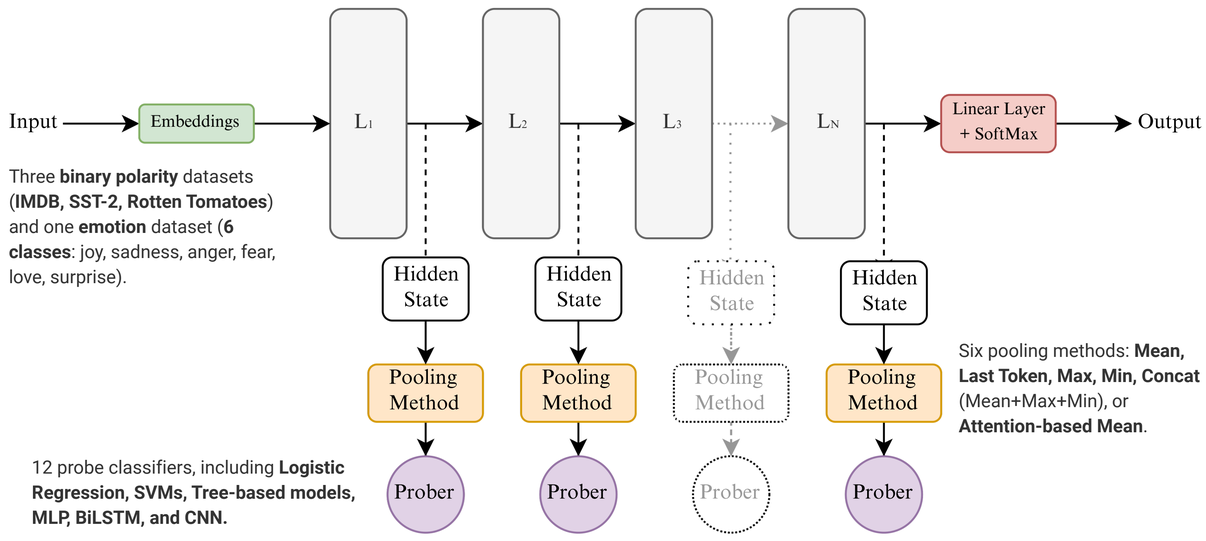


Figure 1: Layer-wise sentiment probing in LLaMA-3.

1. Sentiment Detection Results

Binary Polarity Tasks (SST-2, IMDB, Rotten Tomatoes)

- **Top performers**: Non-linear SVM, Linear SVM, and Logistic Regression
- **Peak accuracy:** ~90% in the middle layers of all model sizes

Fine-Grained Emotion Tasks (Classes: joy, sadness, anger, fear, love, surprise)

- **Best performer:** Linear SVM
- **Peak accuracy:** ~70% in the early layers across all model sizes

Key Insight

- LLaMA models encode **linear representations** for both binary sentiment and nuanced emotional categories.

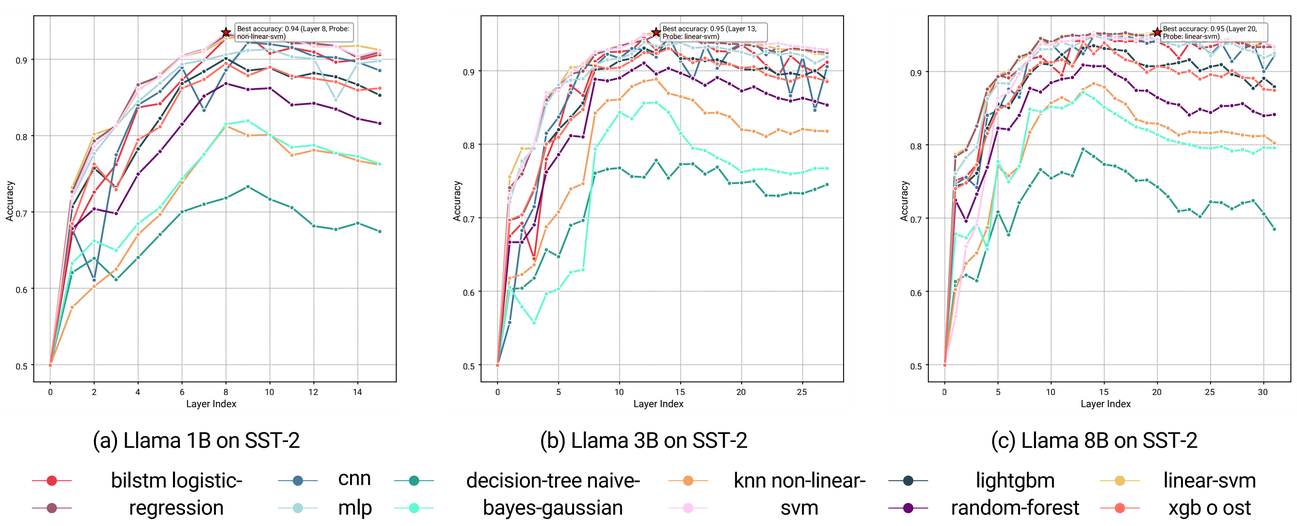


Figure 2: Layer-wise probing accuracy using the Last-Token approach on SST-2.

2. Is the Last-Token Representation the Most Effective?

- **Last-token pooling is not optimal** for sentiment detection in LLaMA models.
- **Best-performing** method: **Concatenation of Mean + Max + Min** pooling yields the highest accuracy across tasks.
- Other **strong alternatives**: **Mean pooling** or **Attention-based pooling.** These consistently perform on par with concatenation and **outperform last-token pooling.**

3. SentriLlama for Efficient Downstream Tasks

What is SentriLlama? A streamlined variant of LLaMA that leverages only the layers most relevant for sentiment analysis.

Layer selection via probing. SentriLlama identifies the optimal layer by evaluating sentiment representations across layers using probing techniques.

Lightweight classification head. Once the best layer is found, a compact task-specific classifier (e.g., Linear SVM) is attached, eliminating the need to use the full model.

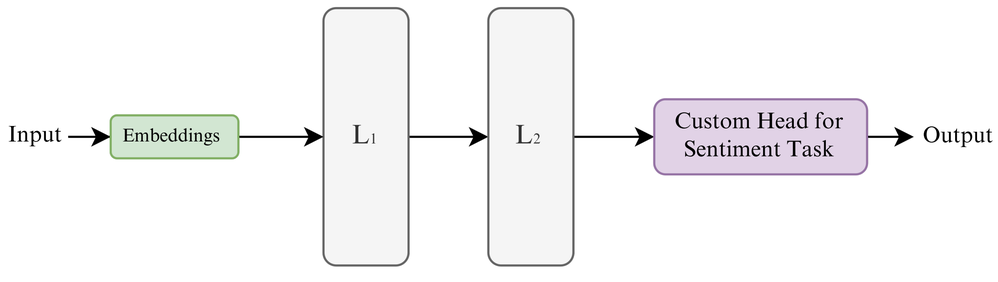


Figure 3: SentriLlama architecture example illustrating the use of layer L2, (the most representative), with an attached custom classification head (e.g., Linear SVM).

4. SentriLlama: Performance & Efficiency Highlights

Accuracy improves with scale: Larger LLaMA models yield better sentiment detection performance.

Minor gains from instruction tuning: Instruct versions show limited advantage, especially at the 1B scale.

Model compression via layer selection: SentriLlama retains only the most sentiment-relevant layers.

Parameter Reduction

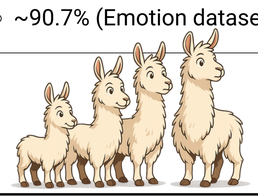
- **1B model**:
  - ~19% (SST-2, IMDB, Rotten Tomatoes)
  - ~61.6% (Emotion dataset)
- **3B model**:
  - ~36.6% (SST-2, IMDB, Rotten Tomatoes)
  - ~80% (Emotion dataset)
- **8B model**:
  - ~53.7% (SST-2, IMDB, Rotten Tomatoes)
  - ~90.7% (Emotion dataset)

Computational Efficiency

GPU Memory Savings:
- 37.5% less (SentriLlama 3.2 1B vs. Instruct-LLaMA 1B)
- 79.2% less (SentriLlama 3.1 8B vs. Instruct-LLaMA 8B)

Faster Inference:
- 45.5% faster (3.2 1B)
- 71.9% faster (3.2 3B)
- 85.9% faster (3.1 8B)

Baseline Comparison:
- **RoBERTa** is most memory-efficient (692 MB), and 36.4% slower than SentriLlama 3.2 (8B)
- **DeBERTa** is 75.9% slower



Model	SST2	IMDB	Rotten Tomatoes	Emotion
Instruct-Llama 3.2 (1B)				
Zero-shot	0.7210	0.6898	0.6923	0.2140
Few-shot	0.6485	0.5994	0.5994	0.2885
Chain-of-Thought	0.4992	0.5000	0.5000	0.3475
Instruct-Llama 3.2 (3B)				
Zero-shot	0.7759	0.8397	0.7279	0.3750
Few-shot	0.7606	0.8528	0.7176	0.3045
Chain-of-Thought	0.9154	0.9306	0.8743	0.4645
Instruct-Llama 3.1 (8B)				
Zero-shot	0.9341	0.9461	0.9024	0.4455
Few-shot	0.9330	0.9411	0.8968	0.3340
Chain-of-Thought	0.9165	0.9363	0.8771	0.5605
SentriLlama 3.2 (1B)	0.9308	0.9445	0.8912	0.8015
SentriLlama 3.2 (1B) Instruct	0.9450	0.9400	0.8940	0.7880
SentriLlama 3.2 (3B) Instruct	0.9594	0.9523	0.9090	0.8220
SentriLlama 3.1 (8B) Instruct	**0.9605**	**0.9579**	**0.9203**	0.8685
DeBERTa V3 Large (418M)	0.9599	0.9534	0.8671	**0.8765**
RoBERTa Large (355M)	0.9038	0.9430	0.8808	0.8416

Figure 2: Comparison of SentriLlama against DeBERTa, RoBERTa, and prompt-based method.