

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

Dario Di Palma¹ Alessandro De Bellis¹ Giovanni Servedio^{1,2}
Vito Walter Anelli¹ Fedelucio Narducci¹ Tommaso Di Noia¹

¹Politecnico di Bari, Italy ²Sapienza University of Rome, Italy
name.surname@poliba.it

Abstract

Large Language Models (LLMs) have rapidly become central to NLP, demonstrating their ability to adapt to various tasks through prompting techniques, including sentiment analysis. However, we still have a limited understanding of how these models capture sentiment-related information. This study probes the hidden layers of Llama models to pinpoint where sentiment features are most represented and to assess how this affects sentiment analysis.

Using probe classifiers, we analyze sentiment encoding across layers and scales, identifying the layers and pooling methods that best capture sentiment signals. Our results show that sentiment information is most concentrated in mid-layers for binary polarity tasks, with detection accuracy increasing up to 14% over prompting techniques. Additionally, we find that in decoder-only models, the last token is not consistently the most informative for sentiment encoding. Finally, this approach enables sentiment tasks to be performed with memory requirements reduced by an average of 57%.

These insights contribute to a broader understanding of sentiment in LLMs, suggesting layer-specific probing as an effective approach for sentiment tasks beyond prompting, with potential to enhance model utility and reduce memory requirements.

1 Introduction

Sentiment analysis (SA), which classifies opinions, emotions, and attitudes in text, is a cornerstone of Natural Language Processing (NLP). SA has enabled the development of several applications, including social media monitoring (Camacho-Collados et al., 2022), customer feedback analysis (Lin et al., 2017), and opinion mining (Xia et al., 2021). It comprises various tasks, such as polarity detection (classifying text as positive, negative, or neutral), emotion classification, and subjectivity detection (Khan et al., 2014).

Earlier approaches for the sentiment classification task relied on supervised learning algorithms and extensive feature engineering (Al-Mannai et al., 2014), requiring large annotated datasets and domain expertise to model sentiment effectively (Bordoloi and Biswas, 2023).

However, the advent of pre-trained models such as GPT-2 (Radford et al., 2019) has transformed NLP by enabling downstream tasks through prompting techniques (Brown et al., 2020; Di Palma, 2023), significantly reducing reliance on extensive labeled data.

While prompting has enhanced the applicability of LLMs for sentiment analysis (Deng et al., 2023; Xing, 2024; Ahmed et al., 2024; Biancofiore et al., 2025), these techniques have often lacked fine-grained control of semantic features, such as context-dependent sentiments or subtle emotional tones. Moreover, the encoding of such features within model representations has not been well understood, limiting efforts to optimize and interpret sentiment analysis performance.

Efforts to interpret neural model representations have advanced significantly over time. Building on the linear representations hypothesis (Mikolov et al., 2013), which suggests that high-level features can be represented as linear directions, researchers have explored where human-interpretable concepts are encoded within LLMs. While previous studies have examined concepts such as truthfulness (Burns et al., 2023), honesty (Azaria and Mitchell, 2023), and factual knowledge (Li et al., 2023), sentiment encoding remains relatively underexplored, despite its importance in NLP applications.

This study aims to fill this gap by investigating where sentiment information is encoded within Llama models (Touvron et al., 2023). We focus on both binary polarity (positive or negative sentiment) and fine-grained emotion detection (joy, sadness, anger, fear, love, and surprise), identifying the model layers that most effectively capture these

sentiment concepts. Specifically, we trained classifiers to identify subspaces corresponding to specific concepts, such as positive sentiment, within model representations. We refer to these classifiers as probes (Alain and Bengio, 2017).

Furthermore, unlike previous studies that rely solely on the last token’s representation (as a sentence summary) for probe training (Burns et al., 2023; Azaria and Mitchell, 2023; Li et al., 2023), we evaluate alternative pooling methods for sentence-level sentiment detection.

Our extensive experiments reveal several key insights into how sentiment and emotion are represented in LLMs. We find that (1) sentiment concepts are most detectable in mid-layer representations, while (2) emotions are more discoverable in early layers; (3) selecting the last token does not consistently yield the best results for probe training; (4) representation quality improves with model size; and (5) probe classifiers generally outperform prompting techniques for accurate sentiment and emotion classification.

In summary, our work makes the following contributions:

- We perform a layer-wise analysis of sentiment encoding in Llama models, identifying the layers that most effectively capture sentiment information.
- We evaluate six pooling methods to identify the most effective approach for sentence representation in sentiment analysis.
- We investigate the impact of model size on sentiment representation by analyzing Llama-3 models of 1B, 3B, and 8B parameters.
- We show that Llama-based classifiers outperform Llama in zero-shot, few-shot, and Chain-of-Thought settings while requiring fewer parameters.
- We propose SENTRILLAMA, a task-specific adaptation of Llama that identifies and utilizes the most representative layer for sentiment analysis, replacing the language modeling head with a lightweight classification head to significantly reduce inference costs while maintaining state-of-the-art performance.

2 Related Work

In this section, we outline the evolution of sentiment analysis, trace the development of probing research in neural models, focusing on LLMs, and conclude by comparing our work to recent studies.

2.1 Sentiment Analysis Meets LLMs

Sentiment analysis, the task of classifying and extracting subjective information from text, has evolved from lexicon-based approaches using dictionaries of positive and negative words (Hatzivassiloglou and McKeown, 1997; Wiebe, 2000) to supervised learning methods employing n-gram models and machine learning (Pang et al., 2002; Chaovalit and Zhou, 2005). These early methods, limited by their reliance on handcrafted features, paved the way for feature engineering techniques and eventually the deep learning revolution that reshaped sentiment classification.

To improve feature identification, researchers incorporated advanced techniques such as syntactic dependencies (Yao et al., 2010), part-of-speech tagging (Søgaard, 2011), and negation handling (Morante and Blanco, 2012). The advent of deep learning further transformed sentiment analysis, with models such as Recurrent Neural Networks (RNNs) (Socher et al., 2013) and Convolutional Neural Networks (CNNs) (Kim, 2014), achieving significant advancements by learning representations directly from raw text.

However, the introduction of Transformer architectures (Vaswani et al., 2017) marked a paradigm shift. Models like BERT (Devlin et al., 2019) leverage pretraining on vast datasets to extract contextualized representations, and they have been widely used as encoding backbones for downstream tasks. Conversely, GPT (Radford et al., 2019) employs the Transformer-decoder block and frames language modeling as an autoregressive task. Finally, inspired by the transfer learning paradigm, Raffel et al. (2020) introduced T5, an encoder-decoder model for text-to-text tasks, trained on vast amounts of data to capture general language patterns and fine-tuned for a wide range of specific applications.

The introduction of GPT-3 (Brown et al., 2020) marked a turning point by introducing few-shot learning and demonstrating the effectiveness of prompt-based techniques, significantly reducing the reliance on large labeled datasets for sentiment analysis and other downstream tasks.

Modern models like GPT-4 (OpenAI, 2023) and Llama (Touvron et al., 2023) achieve even greater flexibility through instruction-following capabilities, excelling in zero-shot settings (Qin et al., 2023) and outperforming fine-tuned models in sentiment-related tasks (Hasan et al., 2024).

Nowadays, the latest models used for sentiment analysis include DeBERTaV3 (He et al., 2023) and RoBERTa-large (Liu et al., 2019), which have demonstrated excellent performance when fine-tuned. Additionally, models like GPT and Llama have shown effectiveness in this domain, leveraging prompting techniques or instruction fine-tuning (Stigall et al., 2024; Krugmann and Hartmann, 2024; Liu et al., 2023; Wei et al., 2023), even in complex aspect-based scenarios (Bai et al., 2024; Zheng et al., 2025).

2.2 Probing and LLMs

Probe classifiers, or *probes* (Alain and Bengio, 2017), are tools designed to extract specific properties from the intermediate representations of neural models. In LLMs, probing helps unveil the semantics of their representations by identifying fine-grained features encoded at different layers, allowing researchers to systematically quantify and compare model capabilities.

In the context of NLP, research has evolved from early analyses of static word embeddings like Word2Vec (Yaghoobzadeh et al., 2019) and GloVe (Klubička and Kelleher, 2022) to methods investigating the complex layered knowledge within LLMs (Chen et al., 2023; Pirozelli et al., 2024). Prior work has probed LLMs for various fine-grained properties, including linguistic properties (Vulić et al., 2020), factual knowledge (Petroni et al., 2019; Wu et al., 2023; De Bellis et al., 2024), beliefs (Azaria and Mitchell, 2023), cross-lingual alignment (Wang et al., 2024), logical reasoning capabilities (Manigrasso et al., 2024), privacy leaks (Kim et al., 2023; Di Palma et al., 2025), toxicity (Wen et al., 2023; Roy et al., 2023). These studies approach probing either through a prompt-based method, where the LLM’s performance is evaluated using specifically designed prompts, or by applying trained classifiers to analyze the model’s intermediate layers.

2.3 Probing LLMs for Sentiment Analysis

Various studies have assessed the capabilities of LLMs for sentiment analysis tasks. For instance, Fatouros et al. (2023) analyzed the performance of ChatGPT-3.5 in financial sentiment analysis, demonstrating performance that exceeds FinBERT (Araci, 2019). Similarly, de Araujo et al. (2024) investigated ChatGPT’s effectiveness in Portuguese sentiment analysis, highlighting its potential value in dataset annotation. Lyu et al. (2024)

investigates the application of causal inference to sentiment analysis and introduces causal prompts to enhance LLM performance in sentiment prediction tasks. Furthermore, Zhang et al. (2024) conducted a systematic evaluation across various sentiment tasks using ChatGPT and different T5 model sizes. Their findings reveal that while LLMs excel in simple (e.g. binary or trinary) zero-shot sentiment classification tasks, they struggle with complex ones (e.g. aspect-based).

While previous studies have primarily evaluated LLMs as *text-to-text models*, focusing on zero- and few-shot learning capabilities, our work takes a different approach. We investigate the hidden representations within the intermediate layers of transformer architectures to identify where sentiment concepts are encoded and how these insights can inform the development of more efficient and accurate models.

Similar studies have investigated the representation of semantic concepts within LLMs (Anelli et al., 2022), such as Burns et al. (2023) on truthfulness, Azaria and Mitchell (2023) on honesty, Roy et al. (2023) on hate speech, and Li et al. (2023) on factual knowledge. However, the investigation of sentiment within the hidden representations of these models remains comparatively underexplored.

3 Methodology

In this section, we detail the models used in the experiments, the datasets, the probe classifiers, and the experimental settings, with a focus on ensuring the reproducibility of our work. Furthermore, we have made all the code publicly available¹ to enable systematic and efficient probing of LLMs.

Sentiment Detection in Hidden Space. We build on the concept detection framework proposed by von Rütte et al. (2024), adapting it specifically for sentiment analysis. In this setup, we define a sentiment concept S and use a corresponding sentiment analysis dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$, where y_i represents the labeled sentiment of sentence x_i . Here, $\text{rep}_\theta(x_i)$ represents an intermediate representation of x_i , generated from a forward pass through the LLM $_\theta$. The goal is to extract a collection of these representations $\{\text{rep}_\theta(x_i)\}_{i=1}^n$. We therefore train a classifier C_w on these representations to predict the presence of the sentiment concept S (i.e., pos-

¹Sentiment Probing Toolkit

itive or negative sentiment within x_i), effectively predicting y_i .

Implementing sentiment detection involves two fundamental design choices: (1) The selection of the intermediate representation rep_θ , which may vary depending on the layers or pooling strategies applied within the LLM. (2) The choice of classifier C_w , which serves to distinguish between different sentiment categories based on these embeddings.

In the following, we outline the fundamental structure of Transformer architecture and highlight common techniques for selecting representations. **Choice of Representation.** Current state-of-the-art LLMs are based on the Transformer architecture (Vaswani et al., 2017), where sequential Transformer blocks generate intermediate hidden representations (h), each with potentially distinct functionalities. Let $l \in \mathbb{N}$ denote the l -th layer, and $\mathbf{x}^{(l)} \in \mathbb{R}^{T \times d_{\text{emb}}}$ represent the output, where T is the number of tokens and d_{emb} is the hidden dimension. A Transformer refines $\mathbf{x}^{(l)}$ using multi-head attention (MHA) and a feed-forward network (FFN):

$$\begin{aligned} h_{\text{attn}}^{(l)} &= \text{MHA} \left(\text{LayerNorm} \left(x^{(l)} \right) \right) \\ h_{\text{resid}}^{(l)} &= h_{\text{attn}}^{(l)} + x^{(l)} \\ h_{\text{ffn}}^{(l)} &= \text{FFN} \left(\text{LayerNorm} \left(h_{\text{resid}}^{(l)} \right) \right) \\ x^{(l+1)} &= h_{\text{ffn}}^{(l)} + h_{\text{resid}}^{(l)} \end{aligned}$$

We exploit the Llama-3 architecture (Dubey et al., 2024), leveraging its Grouped Query Attention (GQA) and RMSNorm features to extract hidden representations efficiently.

Previous studies have explored various representations, such as the residual stream ($\mathbf{x}^{(l+1)}$) (Marks and Tegmark, 2023; Burns et al., 2023; Zou et al., 2023; Gurnee and Tegmark, 2024), the normalized residual stream (nostalgebraist, 2020), or attention heads (Li et al., 2023; Arditì and Obeso, 2023). Based on our preliminary experiments showing marginally higher detection accuracy, we adopt the residuals stream ($\text{rep}_\theta(\mathbf{x}) = \mathbf{x}^{(l+1)}$).

Instead of using the full prompt representation ($\mathbf{x}_{\text{rep}} \in \mathbb{R}^{T \times d_{\text{emb}}}$, where T is the token count), we focus on a subset ($\mathbf{x}_{\text{rep}} \in \mathbb{R}^{t \times d_{\text{emb}}}$) with $t \leq T$. Each token representation ($\mathbf{x}_{\text{rep}}[i, :] \in \mathbb{R}^{d_{\text{emb}}}$, for $i = 1, \dots, t$), is treated as an independent feature. This approach focuses on the parts of the prompt most likely to capture the sentiment concept. Prior work has carefully selected a single token (Arditì

and Obeso, 2023; Zou et al., 2023; Gurnee and Tegmark, 2024) or relied on the last token of the prompt (Rimsky, 2023; Mallen and Belrose, 2023; Marks and Tegmark, 2023; Li et al., 2023; Burns et al., 2023).

In our experiments, we explore six methods for selecting the representations:

- (1) Mean Pooling (Fig. 2): Compute the mean activation value across all tokens for each d_{embed} dimension, resulting in a single vector where each element corresponds to the average activation of a particular feature (embedding dimension) over the entire sequence.
- (2) Last-Token Pooling (Fig. 3): Uses the final token’s features of the last token in the sequence.
- (3) Max Pooling (Fig. 4): Identifies the most prominent feature across all tokens T for each feature, outputting a vector representing the most dominant features in the sequence.
- (4) Min Pooling (Fig. 4): Complementary to Max Pooling, producing a vector representing the least dominant features in the sequence.
- (5) Concat-Mean-Max-Min Pooling: Concatenates the mean, max, and min pooling, generating a representation of size $3 \times d_{\text{embed}}$ that encapsulates multiple aspects of the token embeddings.
- (6) Attention Mean Pooling (Fig. 5): Given a token representation $\mathbf{x}_{\text{rep}}[i, j]$, where i denotes the token and j represents a specific embedding dimension, this pooling method constructs a representation by combining token embeddings with a corresponding importance score. The importance score for each token i is computed by applying the softmax function to the mean of the token’s embedding values. This assigns higher weights to tokens with larger average values. These scores are multiplied element-wise to each dimension of the token embeddings, $x_{\text{rep}}[i, :]$, to adjust their contribution based on importance. The final pooled representation, x_{pooled} , is then computed as a weighted sum of these adjusted embeddings, emphasizing the most relevant tokens activations. Mathematically:

$$x_{\text{pooled}} = \sum_{i=1}^T \text{softmax} \left(\frac{1}{J} \sum_{j=1}^J x_{\text{rep}}[i, j] \right) \cdot x_{\text{rep}}[i, :]$$

A visual representation of the pooling strategies can be found in Appendix D. These approaches offer diverse strategies for selecting representations, enabling a more nuanced understanding of how

sentiment information is encoded. The choice of tokens should ideally reveal whether sentiment is distributed across the entire prompt or concentrated in specific tokens. For instance, the sentences “My name is XYZ and I’m happy” and “I’m happy because my name is XYZ” both convey positive sentiment. However, relying solely on the last token could lead to a suboptimal representation, as it may not adequately capture the sentiment expressed earlier in the sequence.

Choice of Classifier. With the hidden representations selected, we can train our probing classifier C_w on the sentiment labels. In our experimental setup we include twelve distinct classifiers, grouped into five categories based on their underlying modeling approach, as summarized in Table 1.

Type	Models
Linear	Logistic Regression, Linear SVM
Distance-based	K-Nearest Neighbors
Tree-based	Decision Tree, Random Forest, XGBoost, LightGBM
Neural Network	MLP, BiLSTM, CNN
Other	Non-linear SVM, Gaussian Naive Bayes

Table 1: Clusters of Classifiers Based on Model Type

These classifiers were chosen to balance simplicity, interpretability, and the capacity to model complex patterns. Linear models, such as Logistic Regression and Linear SVM, serve as baselines for probing tasks due to their simplicity and their ability to identifying and leveraging linear relationships. Non-linear and neural network models, on the other hand, are included for their ability to capture intricate relationships within the hidden space. BiLSTM and CNN, in particular, were selected for their proven ability to generate effective hidden representations (Ghafoor et al., 2023).

All models, except for BiLSTM and CNN, are implemented using the scikit-learn library (Pedregosa et al., 2011). While, BiLSTM and CNN were implemented using PyTorch. For each classifier, we employed the Optuna framework (Akiba et al., 2019) for hyperparameter optimization, focusing on key parameters such as regularization strengths, tree depths, and kernel types. Optimization was performed over five trials (Paparella et al.,

2023), balancing computational efficiency with sufficient exploration of the hyperparameter space.

To ensure reproducibility, we configured random seed initialization and enforced deterministic behavior for CUDA operations. Details are provided in Appendix B.

Datasets Details. Our experiments utilize three benchmark datasets for sentiment classification: IMDB (Maas et al., 2011), SST-2 (Socher et al., 2013), and Rotten Tomatoes (Pang and Lee, 2005), all of which contain movie reviews for binary polarity tasks (e.g., positive or negative sentiment). Additionally, we extended the setup to a more complex evaluation, conducting fine-grained sentiment classification using the Emotion dataset (Saravia et al., 2018), which categorizes sentiments into six nuanced classes: joy, sadness, anger, fear, love, and surprise. We only preprocessed the IMDB and Emotion datasets because their original sizes, 50K for IMDB and 20K for Emotion, made them impractical for the large number of experiments. Details of the preprocessing steps are provided in Appendix A. Table 2 provides details on the number of samples and the train-test splits used in the experiments.

Dataset	Train	Test	Labels	Max Len
IMDB	7000	7000	pos/neg	132
SST-2	6920	1821	pos/neg	56
Rotten Tomatoes	8530	1066	pos/neg	59
Emotion	6000	2000	six labels	64

Table 2: Dataset statistics.

4 Experimental Results

We conduct our experiments on three Llama models, specifically testing the instruction-tuned variants of Llama 3.2-1B, Llama 3.2-3B, and Llama 3.1-8B, all sourced from Hugging Face². For every probing datasets we train the classifier C_w on the training set, and evaluate its detection performance by measuring the accuracy on the test set.

Sentiment detection results. In our first experiments, we evaluate the detection performance of the different classifier C_w . To extract the rep_θ , following prior work, we focus on the residual stream using the Last-Token approach. We report detection accuracies for each layer in LLM_θ and visualize the results in Fig. 1.

²<https://huggingface.co/meta-llama>

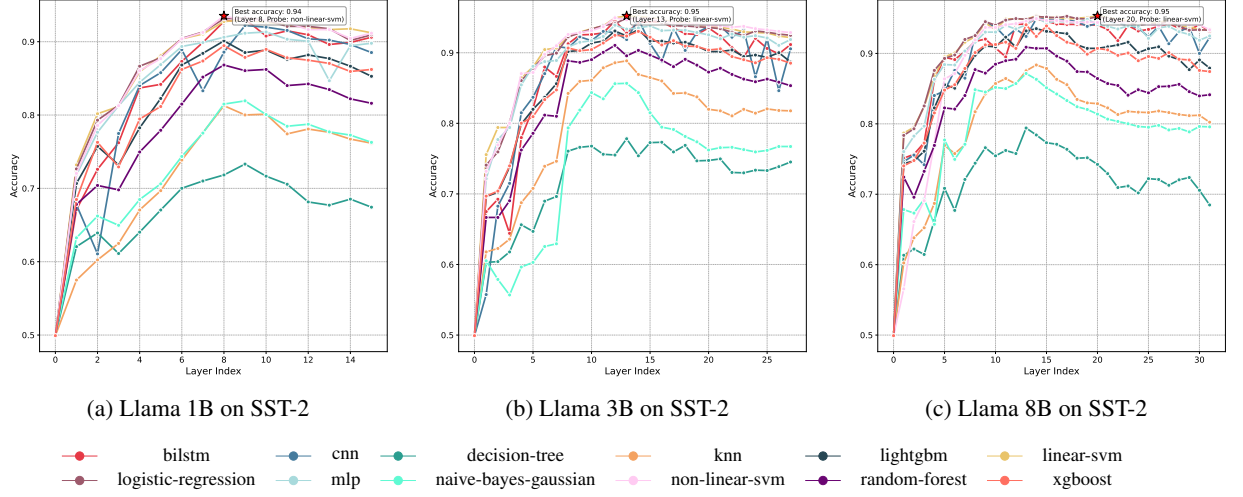


Figure 1: Layer-wise probing accuracy using the Last-Token approach on SST-2; Appendix E for additional dataset results.

For sentiment detection in binary polarity tasks (i.e., SST-2, IMDB, and Rotten Tomatoes), non-linear SVM, linear SVM, and logistic regression consistently outperform other probing techniques, achieving approximately 90% accuracy in the middle layers across all model sizes. For fine-grained emotion datasets, linear SVM demonstrates the best performance, reaching around 70% accuracy in the initial layers across all model sizes. This indicates that Llama models have linear representations for binary sentiment (positive/negative) and fine-grained emotions (joy, sadness, anger, fear, love, and surprise). It is interesting to discover that, although sentiment is not easily discernible during direct interaction with Llama, sentiment and emotion concepts can be linearly detected within certain internal layers.

Furthermore, since prior work often use the last-token representations, we also investigate the effect of different token representation methods. Specifically, we compare the last-token approach with five alternatives: mean, max, min, concatenation, and attention, as previously described.

The results, summarized in Table 3, which shows the top-3 performers by datasets, reveal that concatenating the mean, max, and min of rep_θ is often the most effective method for detecting sentiment concepts. Notably, mean and attention pooling also perform strongly, achieving results comparable to concatenation and consistently outperforming last-token pooling. This indicates, that the last-token representation is not always the optimal choice for capturing sentiment concepts.

Furthermore, we discover that combining the max pooling representation with LightGBM yields particularly strong performance. Further investigation reveals that tree-based models, including Random Forest, Decision Tree, and XGBoost, also perform better with max pooling. We hypothesize that the synergy between max pooling and tree-based models arises from max pooling’s ability to emphasize dominant features, which, when combined with tree-based models’ strength in exploiting high-contrast, threshold-based partitions, leads to an optimal combination. This alignment makes max pooling particularly effective for tree-based models, highlighting how the choice of classifier should also influence the choice of representation, and vice versa.

5 SENTRILLAMA for Efficient Downstream Tasks

Sentiment analysis has been a fundamental task since the inception of NLP. Traditionally, there have been two primary approaches to accomplish this task: (1) training a model from scratch or (2) fine-tuning an existing model to meet specific requirements. Recently, a third approach has gained prominence: leveraging well-designed prompts in conjunction with state-of-the-art large language models (Deng et al., 2023; Xing, 2024; Ahmed et al., 2024).

We introduce SENTRYLLAMA, a specialized Llama model that leverages layers up to $L_{\leq i}$ for sentiment tasks, where i denotes the most representative layer for the task. Based on our

Model	Dataset	Layer	Prober	Pooling	Accuracy
Llama 3.2 1B (Instruct)	SST2	10	Non-linear SVM	Attn	0.9450
		10	Non-linear SVM	Mean	0.9450
		8	Non-linear SVM	Last-Token	0.9352
	IMDB	8	Logistic Reg.	Mean	0.9400
		8	Logistic Reg.	Attn	0.9396
		8	Linear SVM	Last-Token	0.9009
	Rotten	8	Linear SVM	Concat	0.8939
		8	Logistic Reg.	Concat	0.8939
		8	Non-linear SVM	Last-Token	0.8789
	Emotion	1	Linear SVM	Concat	0.7880
		10	LightGBM	Max	0.7880
		7	Linear SVM	Last-Token	0.6830
Llama 3.2 3B (Instruct)	SST2	5	Non-linear SVM	Concat	0.9594
		12	Non-linear SVM	Concat	0.9577
		13	Linear SVM	Last-Token	0.9522
	IMDB	4	Logistic Reg.	Attn	0.9523
		14	Logistic Reg.	Mean	0.9522
		13	Non-linear SVM	Last-Token	0.9177
	Rotten	3	Logistic Reg.	Concat	0.9090
		3	CNN	Mean	0.9071
		13	Linear SVM	Last-Token	0.8949
	Emotion	1	MLP	Concat	0.8220
		0	LightGBM	Concat	0.8195
		4	Linear SVM	Last-Token	0.6940
Llama 3.1 8B (Instruct)	SST2	3	Logistic Reg.	Concat	0.9605
		14	Logistic Reg.	Concat	0.9599
		14	Linear SVM	Last-Token	0.9533
	IMDB	4	Non-linear SVM	Concat	0.9579
		16	Non-linear SVM	Mean	0.9576
		11	Linear SVM	Last-Token	0.9273
	Rotten	9	MLP	Mean	0.9203
		14	CNN	Mean	0.9203
		14	Logistic Regression	Last-Token	0.9183
	Emotion	0	Linear SVM	Concat	0.8685
		0	LightGBM	Max	0.8655
		4	Linear SVM	Last-Token	0.6885

Table 3: Probing Results Across Different Pooling Methods. See Appendix F for confidence level plots across all layers and datasets.

earlier analysis, we identified the optimal layer L_i , which retains the most salient features for downstream tasks. By attaching a lightweight, task-specific classification head—such as a LinearSVM—SENTRI L LAMA efficiently repurposes the expressive power of the Llama model while discarding unnecessary layers. This approach significantly reduces computational requirements for inference, making it both efficient and task-specific without compromising performance.

Specifically, the total model parameters are calculated as the sum of the input embedding, N -layer, and LM head parameters. The proposed SENTRY L LAMA approach reduces model complexity by retaining only the input embedding and the layers up to the most representative layer (i), identified in previous experiments, while replacing the LM head with a lightweight classification head, such as a linear SVM. This reduces parameter usage and adapts the architecture for downstream tasks.

To benchmark the effectiveness of SENTRY L-

LAMA, we compare its performance against fine-tuned DeBERTaV3-large³ and RoBERTa-large⁴ across all datasets. Additionally, we evaluate SENTRY L LAMA against Llama models under zero-shot, few-shot, and Chain-of-Thought prompting scenarios using carefully designed templates inspired by prior work (Deng et al., 2023). Details on the prompt used are provided in Appendix C. Furthermore, to understand the effect of instruction fine-tuning on the sentiment task, we also include in the comparison the non-instructed version of Llama3.2 (1B). Table 4 presents a comparative analysis of the accuracy of SENTRY L LAMA, DeBERTa, RoBERTa, and a prompt-based method. The best-performing model is highlighted in **bold**, while the runner-up is underlined.

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
SENTRI L LAMA 3.2 (1B)	0.9308	0.9445	0.8912	0.8015
SENTRI L LAMA 3.2 (1B) Instruct	0.9450	0.9400	0.8940	0.7880
SENTRI L LAMA 3.2 (3B) Instruct	0.9594	0.9523	<u>0.9090</u>	0.8220
SENTRI L LAMA 3.1 (8B) Instruct	0.9605	0.9579	0.9203	<u>0.8685</u>
DeBERTa V3 Large (418M)	<u>0.9599</u>	<u>0.9534</u>	0.8671	0.8765
RoBERTa Large (355M)	0.9038	0.9430	0.8808	0.8416

Table 4: Comparison of SENTRY L LAMA against DeBERTa, RoBERTa, and prompt-based method.

For the sentiment downstream task, SENTRY L LAMA 8B outperforms all other approaches, achieving approximately 96% accuracy on SST-2, 96% on IMDB, 92% on Rotten Tomatoes, and 87% on the Emotion dataset. Only DeBERTa surpasses this performance on the Emotion dataset, with an accuracy of 88%. The prompt-based approach reveals the effectiveness of few-shot and Chain-of-Thought (CoT), achieving comparable results but falling short on the Emotion dataset. These results align with expectations, revealing several insights: (1) increasing the base model size improves performance, (2) the non-instruction version of the 1B model exhibits minimal differences compared to the instructed ones, and (3) DeBERTa and RoBERTa continue to lead the leaderboard in sentiment analysis, outperforming prompt-based

³Microsoft/deberta-v3-large

⁴FacebookAI/roberta-large

methods. However, it is worth noting that the actual size of SENTRILLAMA differs from the original Llama model. We represent the effective sizes in Table 5.

Model	SST2	IMDB	Rotten Tomatoes	Emotion
SENTRI L LAMA 3.2 (1B)	810M	811M	811M	384M
SENTRI L LAMA 3.2 (1B) Instruct	932M	810M	810M	384M
SENTRI L LAMA 3.2 (3B) Instruct	2B	1.9B	1.8B	595M
SENTRI L LAMA 3.1 (8B) Instruct	3.5B	3.8B	3.8B	743M

Table 5: Comparison of SENTRILLAMA Sizes Across Datasets.

Model	Peak GPU Usage	Avg. Time per Sample	Throughput (Samples/sec)
Instruct-Llama 3.2 (1B)	2.4 GB	11.17 ms	90
Instruct-Llama 3.2 (3B)	6.2 GB	18.19 ms	55
Instruct-Llama 3.1 (8B)	15.4 GB	37.73 ms	48
SENTRI L LAMA 3.2 (1B)	1.5 GB	6.08 ms	164
SENTRI L LAMA 3.2 (1B) Instruct	1.7 GB	7.98 ms	125
SENTRI L LAMA 3.2 (3B) Instruct	1.7 GB	5.09 ms	196
SENTRI L LAMA 3.1 (8B) Instruct	3.2GB	5.31 ms	182
DeBERTa V3 Large (418M)	845 MB	22.03 ms	45
RoBERTa Large (355M)	692 MB	8.35 ms	120

Table 6: Computational efficiency comparison on the SST-2 dataset. See Appendix G for additional dataset results.

Using SENTRILLAMA reduces the size of the models. For the 1B model (both the standard and instruction-finetuned versions) on SST-2, IMDB, and Rotten Tomatoes, approximately 19% of the parameters are removed, while for the Emotion dataset, the reduction is up to 61.6%. For the 3B model on SST-2, IMDB, and Rotten Tomatoes, the reduction is approximately 36.6%, and for the Emotion dataset, it is 80%. For the 8B model on SST-2, IMDB, and Rotten Tomatoes, approximately 53.7% of the parameters are removed, while for the Emotion dataset, the reduction is 90.7%. This result indicates that the amount of parameters removed depends heavily on the specific task and that it is possible to reuse the highly relevant representations of Llama to improve downstream tasks, going beyond prompting and resulting in a competitive model.

Furthermore, we compare the computational efficiency of SENTRILLAMA with DeBERTa, RoBERTa, and a prompt-based method. In Table 6, we present the results for SST-2.

Comparing GPU usage, SENTRILLAMA models demonstrate greater memory efficiency than their Instruct-Llama counterparts. SENTRILLAMA 3.2 (1B) uses 37.5% less GPU memory than Instruct-Llama 3.2 (1B), while SENTRILLAMA 3.1 (8B)

reduces memory usage by 79.2% compared to Instruct-Llama 3.1 (8B). However, DeBERTa and RoBERTa remain the most memory-efficient, with RoBERTa requiring only 692 MB.

In terms of inference speed, SENTRILLAMA models exhibit a significantly lower average time per sample, indicating they are faster than their Instruct-Llama counterparts. Specifically, SENTRILLAMA 3.2 (1B) is 45.5% faster than Instruct-Llama 3.2 (1B), SENTRILLAMA 3.2 (3B) is 71.9% faster than Instruct-Llama 3.2 (3B), and SENTRILLAMA 3.1 (8B) is 85.9% faster than Instruct-Llama 3.1 (8B). Meanwhile, DeBERTa has a 75.9% longer inference time compared to SENTRILLAMA 3.2 (8B), while RoBERTa is 36.4% slower.

Finally, SENTRILLAMA models achieve significantly higher throughput (samples per second) than Instruct-Llama, DeBERTa, and RoBERTa.

Overall, this analysis reveals that SENTRILLAMA models are more efficient in terms of GPU memory usage and inference speed, consistently outperforming Instruct-Llama models across all aspects. Smaller models, such as SENTRILLAMA 3.2 (1B) and SENTRILLAMA 3.2 (3B), offer the best trade-off between memory consumption and speed. Meanwhile, RoBERTa Large provides competitive throughput and latency compared to larger Llama-based models, whereas DeBERTa struggles with slower inference.

6 Discussion

Our results show that our method is still far from providing a smaller Llama model that performs as well as similarly sized models like DeBERTa or RoBERTa. The primary limitation stems from the need to retain all preceding layers up to the most representative one. This limitation introduces an element of randomness: if circumstances align favorably and the most representative layer is among the early layers (as observed with the Emotion dataset), the final model’s size becomes highly competitive with DeBERTa and RoBERTa. However, if the most representative layer is in the middle layers, the number of parameters—although still lower than the full model—can exceed those of the baseline models.

While we have automated the discovery of the most representative layer and pooling methods (see Table 3), the dependence on preceding layers results in a suboptimal outcome. A more robust solution could involve leveraging only the input

embeddings and the most representative layer, potentially complemented by layer compression or pruning techniques to reduce computational overhead while preserving performance.

Is this method generalizable for different downstream tasks beyond sentiment analysis? While we cannot yet provide a definitive answer on the method’s generalizability, prior studies have shown that LLMs capture a broad range of semantic concepts, such as truthfulness, honesty, and factual knowledge (Burns et al., 2023; Azaria and Mitchell, 2023; Li et al., 2023). Our findings suggest that sentiment and emotion are similarly well-represented in these models. We hypothesize that this approach could extend to tasks requiring nuanced linguistic analysis, such as sarcasm detection or intent recognition, assuming the availability of labeled datasets.

However, significant challenges remain. Extending this method to more complex tasks may require additional adaptations. Future research should systematically evaluate this approach across diverse tasks and datasets to uncover its broader potential and limitations.

7 Conclusion

In this work, we present a systematic analysis of hidden representations in Llama for sentiment analysis. Our findings demonstrate that for positive/negative sentiment, the middle layers are the most representative across dimensions, whereas for emotion analysis, the most representative layers are found in the initial ones. Additionally, we show that while the current standard for representing a sentence involves using the last token, alternative methods, such as concatenating the max, min, and mean of the representations, yield better results.

Finally, we introduce a novel approach called SENTRILLAMA, which utilizes only a subset of the model’s parameters. This approach achieves results comparable to state-of-the-art models and outperforms prompting-based methods. We hope that the introduction of this layer selection approach for Llama will inspire the development of more robust techniques, facilitating the reuse of these large models for downstream tasks and extending their usability beyond text generation.

Limitations

This work has two primary limitations. First, the study does not evaluate the proposed approach on

domain-specific datasets, such as those from financial, healthcare, or social media domains, nor does it address multilingual sentiment analysis. This gap restricts the generalizability of the findings to a broader range of real-world applications. Second, while the study introduces a method for identifying the most representative layer, there is scope for exploring alternative methodologies within this selection mechanism.

Acknowledgments

This study was supported by the MOST – Sustainable Mobility National Research Center funded by the European Union Next- GenerationEU (Italian National Recovery and Resilience Plan (NRRP) – M4C2, Investment 1.4 – D.D. 1033 17/06/2022, CN00000023 - CUP: D93C22000410001). We acknowledge ISCRA for awarding this project access to the LEONARDO supercomputer, owned by the EuroHPC Joint Undertaking, hosted by CINECA (Italy). Patti Territoriali WP1, OVS: Fashion Retail Reloaded, Lutech Digitale 4.0, Natuzzi S.p.A. del Contratto di Sviluppo Industriale ai sensi dell’art. 9 del Decreto del Ministro dello Sviluppo Economico del 09.12.2014. EPANSA (FAIR) - Enhancing Personal Assistants with Neuro-Symbolic AI and Knowledge Graphs, funded by the European Union Next- GenerationEU (NRRP – M4C2, Investment 1.3, D.R. No. 123 of 16/01/2024, PE00000013, CUP: H97G22000210007). XAI4AMELIA - Implementation of Explainable Artificial Intelligence Technologies for Advancing Interoperability and Analysis in AMELIA, funded by the European Union Next- GenerationEU (NRRP – M4C2, Investment 1.3, D.D. 341 del 15/03/2022, CUP: J33C22002910001). Furthermore, this work has been carried out while Giovanni Servedio was enrolled in the Italian National Doctorate on Artificial Intelligence run by Sapienza University of Rome in collaboration with Politecnico di Bari.

References

- Rabbia Ahmed, Sadaf Abdul-Rauf, and Seemab Latif. 2024. Leveraging large language models and prompt settings for context-aware financial sentiment analysis. In *ICACS*, pages 1–9. IEEE.
- Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

- Kamla Al-Mannai, Hanan Alshikhabobakr, Sabih Bin Wasi, Rukhsar Neyaz, Houda Bouamor, and Behrang Mohit. 2014. Cmuq-hybrid: Sentiment classification by feature engineering and parameter tuning. In *SemEval@COLING*, pages 181–185. The Association for Computer Linguistics.
- Guillaume Alain and Yoshua Bengio. 2017. Understanding intermediate layers using linear classifier probes. In *ICLR (Workshop)*. OpenReview.net.
- Vito Walter Anelli, Giovanni Maria Biancofiore, Alessandro De Bellis, Tommaso Di Noia, and Eugenio Di Sciascio. 2022. Interpretability of BERT latent space through knowledge graphs. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management, Atlanta, GA, USA, October 17-21, 2022*, pages 3806–3810. ACM.
- Dogu Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models. *CoRR*, abs/1908.10063.
- A. Ardit and O. B. Obeso. 2023. Refusal mechanisms: Initial experiments with llama-2-7b-chat. <https://www.lesswrong.com/posts/pYcEhoAoPfHhgJ8YC>. Accessed: 2024-11-18.
- Amos Azaria and Tom M. Mitchell. 2023. The internal state of an LLM knows when it’s lying. In *EMNLP (Findings)*, pages 967–976. Association for Computational Linguistics.
- Yinhao Bai, Zhixin Han, Yuhua Zhao, Hang Gao, Zhuowei Zhang, Xunzhi Wang, and Mengting Hu. 2024. Is compound aspect-based sentiment analysis addressed by LLMs? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7836–7861, Miami, Florida, USA. Association for Computational Linguistics.
- Giovanni Maria Biancofiore, Dario Di Palma, Claudio Pomo, Fedelucio Narducci, and Tommaso Di Noia. 2025. Conversational user interfaces and agents. In *Human-Centered AI: An Illustrated Scientific Quest*, pages 399–438. Springer.
- Monali Bordoloi and Saroj K. Biswas. 2023. Sentiment analysis: A survey on design framework, applications and future scopes. *Artif. Intell. Rev.*, 56(11):12505–12560.
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, and Jared Kaplan et al. 2020. Language models are few-shot learners. In *NeurIPS*.
- Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. 2023. Discovering latent knowledge in language models without supervision. In *ICLR*. OpenReview.net.
- José Camacho-Collados, Kiamehr Rezaee, Talayeh Riahi, Asahi Ushio, and Daniel Loureiro et al. 2022. Tweetnlp: Cutting-edge natural language processing for social media. In *EMNLP (Demos)*, pages 38–49. Association for Computational Linguistics.
- Pimwadee Chaovalit and Lina Zhou. 2005. Movie review mining: a comparison between supervised and unsupervised classification approaches. In *HICSS*. IEEE Computer Society.
- Nuo Chen, Ning Wu, Shining Liang, Ming Gong, Linjun Shou, Dongmei Zhang, and Jia Li. 2023. Is bigger and deeper always better? probing llama across scales and layers. *CoRR*, abs/2312.04333.
- Gladson de Araujo, Tiago de Melo, and Carlos Mauricio S. Figueiredo. 2024. Is chatgpt an effective solver of sentiment analysis tasks in portuguese? A preliminary study. In *PROPOR*, pages 13–21. Association for Computational Linguistics.
- Alessandro De Bellis, Vito Walter Anelli, Tommaso Di Noia, and Eugenio Di Sciascio. 2024. PRONTO: prompt-based detection of semantic containment patterns in mlms. In *ISWC (2)*, volume 15232 of *Lecture Notes in Computer Science*, pages 227–246. Springer.
- Xiang Deng, Vasilisa Bashlovkina, Feng Han, Simon Baumgartner, and Michael Bendersky. 2023. Llm to the moon? reddit market sentiment analysis with large language models. In *WWW (Companion Volume)*, pages 1014–1019. ACM.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Dario Di Palma. 2023. Retrieval-augmented recommender system: Enhancing recommender systems with large language models. In *RecSys*, pages 1369–1373. ACM.
- Dario Di Palma, Felice Antonio Merra, Maurizio Sfilio, Vito Walter Anelli, Fedelucio Narducci, and Tommaso Di Noia. 2025. Do llms memorize recommendation datasets? a preliminary study on movielens-1m. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2025, Padua, Italy July 13-18, 2025*. ACM.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, and Angela Fan et al. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.
- Georgios Fatouros, John Soldatos, Kalliopi Kouroumali, Georgios Makridis, and Dimosthenis Kyriazis. 2023. Transforming sentiment analysis in the financial domain with chatgpt. *Machine Learning with Applications*, 14:100508.
- Yusra Ghafoor, Shi Jinping, Fernando H. Calderon, Yen-Hao Huang, Kuan-Ta Chen, and Yi-Shin Chen. 2023.

- TERMS: textual emotion recognition in multidimensional space. *Appl. Intell.*, 53(3):2673–2693.
- Wes Gurnee and Max Tegmark. 2024. Language models represent space and time. In *ICLR*. OpenReview.net.
- Md. Arif Hasan, Shudipta Das, Afyat Anjum, Firoj Alam, Anika Anjum, Avijit Sarker, and Sheak Rashed Haider Noori. 2024. Zero- and few-shot prompting with llms: A comparative study with fine-tuned models for bangla sentiment analysis. In *LREC/COLING*, pages 17808–17818. ELRA and ICCL.
- Vasileios Hatzivassiloglou and Kathleen R. McKeown. 1997. Predicting the semantic orientation of adjectives. In *ACL*, pages 174–181. Morgan Kaufmann Publishers / ACL.
- Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. In *ICLR*. OpenReview.net.
- Khairullah Khan, Baharum Baharudin, Aurangzeb Khan, and Ashraf Ullah. 2014. Mining opinion components from unstructured reviews: A review. *J. King Saud Univ. Comput. Inf. Sci.*, 26(3):258–275.
- Siwon Kim, Sangdoo Yun, Hwaran Lee, Martin Gubri, Sungroh Yoon, and Seong Joon Oh. 2023. Propile: Probing privacy leakage in large language models. In *Advances in Neural Information Processing Systems*, volume 36, pages 20750–20762. Curran Associates, Inc.
- Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *EMNLP*, pages 1746–1751. ACL.
- Filip Klubicka and John D. Kelleher. 2022. Probing with noise: Unpicking the warp and weft of embeddings. In *BlackboxNLP@EMNLP*, pages 404–417. Association for Computational Linguistics.
- Jan Ole Krugmann and Jochen Hartmann. 2024. Sentiment analysis in the age of generative ai. *Customer Needs and Solutions*, 11(1):3.
- Kenneth Li, Oam Patel, Fernanda B. Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Inference-time intervention: Eliciting truthful answers from a language model. In *NeurIPS*.
- Shuying Lin, Huosheng Xie, Liang-Chih Yu, and K. Robert Lai. 2017. Sentinlp at IJCNLP-2017 task 4: Customer feedback analysis using a bi-lstm-cnn model. In *IJCNLP (Shared Tasks)*, pages 149–154. Asian Federation of Natural Language Processing.
- Fuxiao Liu, Paiheng Xu, Zongxia Li, Yue Feng, and Hyemi Song. 2023. Towards understanding in-context learning with contrastive demonstrations and saliency maps. *arXiv preprint arXiv:2307.05052*.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.
- Zhiheng Lyu, Zhijing Jin, Fernando Gonzalez Adauto, Rada Mihalcea, Bernhard Schölkopf, and Mrinmaya Sachan. 2024. Do llms think fast and slow? a causal study on sentiment analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9353–9372.
- Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *ACL*, pages 142–150. The Association for Computer Linguistics.
- Alex Mallen and Nora Belrose. 2023. Eliciting latent knowledge from quirky language models. *CoRR*, abs/2312.01037.
- Francesco Manigrasso, Stefan F. Schouten, Lia Morra, and Peter Bloem. 2024. Probing llms for logical reasoning. In *NeSy (1)*, volume 14979 of *Lecture Notes in Computer Science*, pages 257–278. Springer.
- Samuel Marks and Max Tegmark. 2023. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *CoRR*, abs/2310.06824.
- Tomás Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *HLT-NAACL*, pages 746–751. The Association for Computational Linguistics.
- Roser Morante and Eduardo Blanco. 2012. *sem 2012 shared task: Resolving the scope and focus of negation. In **SEM@NAACL-HLT*, pages 265–274. Association for Computational Linguistics.
- nostalgebraist. 2020. Interpreting gpt: The logit lens. <https://www.lesswrong.com/posts/AcKRB8wDpdAN6v6ru>.
- OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.
- Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the ACL*.
- Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? sentiment classification using machine learning techniques. In *EMNLP*, pages 79–86.
- Vincenzo Paparella, Dario Di Palma, Vito Walter Anelli, and Tommaso Di Noia. 2023. Broadening the scope: Evaluating the potential of recommender systems beyond prioritizing accuracy. In *RecSys*, pages 1139–1145. ACM.

- Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, and Mathieu Blondel et al. 2011. Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.*, 12:2825–2830.
- Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick S. H. Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander H. Miller. 2019. Language models as knowledge bases? In *EMNLP/IJCNLP (1)*, pages 2463–2473. Association for Computational Linguistics.
- Paulo Pirozelli, Marcos Menon José, Paulo de Tarso P. Filho, Anarosa A. F. Brandão, and Fábio Gagliardi Cozman. 2024. Assessing logical reasoning capabilities of encoder-only transformer models. In *NeSy (1)*, volume 14979 of *Lecture Notes in Computer Science*, pages 29–46. Springer.
- Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is chatgpt a general-purpose natural language processing task solver? In *EMNLP*, pages 1339–1384. Association for Computational Linguistics.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.
- N. Rimsky. 2023. Reducing sycophancy and improving honesty via activation steering. <https://www.lesswrong.com/posts/zT6hRsDE84HeBKh7E>.
- Sarthak Roy, Ashish Harshvardhan, Animesh Mukherjee, and Punyajoy Saha. 2023. Probing llms for hate speech detection: strengths and vulnerabilities. In *EMNLP (Findings)*, pages 6116–6128. Association for Computational Linguistics.
- Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. 2018. CARER: contextualized affect representations for emotion recognition. In *EMNLP*, pages 3687–3697. Association for Computational Linguistics.
- Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *EMNLP*, pages 1631–1642. ACL.
- Anders Søgaard. 2011. Semi-supervised condensed nearest neighbor for part-of-speech tagging. In *ACL (2)*, pages 48–52. The Association for Computer Linguistics.
- William Stigall, Md Abdullah Al Hafiz Khan, Dinesh Chowdary Attota, Francis Nweke, and Yong Pei. 2024. Large language models performance comparison of emotion and sentiment classification. In *ACM Southeast Regional Conference*, pages 60–68. ACM.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*, pages 5998–6008.
- Dimitri von Rütte, Sotiris Anagnostidis, Gregor Bachmann, and Thomas Hofmann. 2024. A language model’s guide through latent space. In *ICML*. OpenReview.net.
- Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. 2020. Probing pretrained language models for lexical semantics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7222–7240, Online. Association for Computational Linguistics.
- Hetong Wang, Pasquale Minervini, and Edoardo M. Ponti. 2024. Probing the emergence of cross-lingual alignment during LLM training. In *ACL (Findings)*, pages 12159–12173. Association for Computational Linguistics.
- Jerry Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, et al. 2023. Larger language models do in-context learning differently. *arXiv preprint arXiv:2303.03846*.
- Jiaxin Wen, Pei Ke, Hao Sun, Zhixin Zhang, Chengfei Li, Jinfeng Bai, and Minlie Huang. 2023. Unveiling the implicit toxicity in large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1322–1338, Singapore. Association for Computational Linguistics.
- Janyce Wiebe. 2000. Learning subjective adjectives from corpora. In *AAAI/IAAI*, pages 735–740. AAAI Press / The MIT Press.
- WeiQi Wu, Chengyue Jiang, Yong Jiang, Pengjun Xie, and Kewei Tu. 2023. Do PLMs know and understand ontological knowledge? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3080–3101, Toronto, Canada. Association for Computational Linguistics.

- Qingrong Xia, Bo Zhang, Rui Wang, Zhenghua Li, Yue Zhang, Fei Huang, Luo Si, and Min Zhang. 2021. A unified span-based approach for opinion mining with syntactic constituents. In *NAACL-HLT*, pages 1795–1804. Association for Computational Linguistics.
- Frank Xing. 2024. Designing heterogeneous llm agents for financial sentiment analysis. *ACM Transactions on Management Information Systems*.
- Yadollah Yaghoobzadeh, Katharina Kann, Timothy J. Hazen, Eneko Agirre, and Hinrich Schütze. 2019. Probing for semantic classes: Diagnosing the meaning content of word embeddings. In *ACL (1)*, pages 5740–5753. Association for Computational Linguistics.
- Lin Yao, Chengjie Sun, Lu Li, Zhixin Hao, and Xiaolong Wang. 2010. Hybrid learning of syntactic and semantic dependencies. *Comput. Inf. Sci.*, 3(4):187–199.
- Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Jialin Pan, and Lidong Bing. 2024. Sentiment analysis in the era of large language models: A reality check. In *NAACL-HLT (Findings)*, pages 3881–3906. Association for Computational Linguistics.
- Kaiyuan Zheng, Qinghua Zhao, and Lei Li. 2025. Re-assessing the role of chain-of-thought in sentiment analysis: Insights and limitations. *arXiv preprint arXiv:2501.08641*.
- Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, and Alexander Pan et al. 2023. Representation engineering: A top-down approach to AI transparency. *CoRR*, abs/2310.01405.

Appendix

A Detailed Preprocessing Workflow

The IMDB and Emotion datasets were preprocessed due to their original sizes—50K for IMDB and 20K for Emotion—which made them impractical for the extensive number of experiments. To address this, reduced versions of these datasets were created while preserving their statistical properties. A structured preprocessing workflow was used to prepare the datasets for the experiments, ensuring statistical consistency and adherence to specific constraints, such as sentence length and label distribution. The key steps are outlined below:

1. **Initialize Random Seed:** To achieve reproducible datasets, we initialized a random seed (Appendix B) at the beginning of the workflow, ensuring that the sampling process could be replicated exactly in future runs.
2. **Calculate Dataset Statistics:**
 - (a) Compute the label distribution to maintain the proportion of samples for each label in the original dataset.
 - (b) Compute the average sentence length across all samples to preserve similar statistics.
3. **Generate the Dataset:** For each label in the dataset, we followed a sampling process as follows:
 - (a) Filter samples by sentence length for the given label.
 - (b) Random sample to select the required number of samples for each label.
 - (c) Combine the selected samples for all labels into a single dataset. To ensure randomness, the combined dataset was shuffled before saving.
4. **Validate Reduced Dataset:** The label distribution and average sentence length were recalculated for the reduced dataset. If the average sentence length exceeded the specified constraint, an error was raised, indicating the need for further adjustments.

To provide transparency and facilitate analysis, a summary of the preprocessing steps is reported:

1. The total number of samples in both the original and reduced datasets.
2. The distribution of labels before and after reduction.
3. The average sentence length in both datasets.

This comprehensive preprocessing workflow ensured that the reduced datasets retained the statistical properties of the original datasets while adhering to predefined constraints. The code used to generate the datasets is provided in `dataset/reduce_dataset.py`, along with a detailed PDF report containing all the statistics.

B Reproducibility and Determinism

To ensure reproducibility, we consistently set random seeds to 42 across Python’s ‘random’, NumPy, and PyTorch libraries. Furthermore, deterministic behavior was enforced for CUDA-enabled operations by configuring PyTorch’s cuDNN backend and settings environment variables. Further details can be found by examining the `set_seed()` function in `utils.py`. This rigorous control of randomness and parallelism ensures consistent experimental results, addressing the stochastic nature of the workflows.

C Prompting Approach

In this section, we illustrate the prompts adopted for the experiments with the Llama models, covering zero-shot, few-shot and Chain-of-Thought prompting scenarios.

Zero-Shot Prompting for Binary Sentiment

System: "You are an assistant trained to perform strict sentiment and emotion classification. You MUST only respond with the numeric label corresponding to the classification. Do not provide any explanations, reasoning, or any text other than the required numeric value."

User: "Classify the sentiment of the following text: '{text}'"

Assistant: "If the sentiment is positive, respond with '1'. If the sentiment is negative, respond with '0'. No other text, explanation, or formatting."

Zero-Shot Prompting for Emotion

System: "You are an assistant trained to perform strict sentiment and emotion classification. You MUST only respond with the numeric label corresponding to the classification. Do not provide any explanations, reasoning, or any text other than the required numeric value."

User: "Classify the sentiment of the following text: '{text}'"

Assistant: "Classify the text into one of the following emotions and respond only with the corresponding number:
0: Sadness, 1: Joy, 2: Love, 3: Anger, 4: Fear, 5: Surprise.
No explanation or additional text."

Chain-of-Thought Prompting for Binary Sentiment

System: "You are an assistant specialized in sentiment and emotion analysis. Think step-by-step through the reasoning process (chain-of-thought) privately, but provide only the final numeric classification as instructed. Do not include reasoning steps in the output."

User: "Analyze the sentiment of the following text: '{text}'"
Carefully reason step-by-step to determine the sentiment.
Output only '1' for positive sentiment or '0' for negative sentiment as your final response."

Assistant: "I will reason step-by-step internally to determine the sentiment.
However, my final response will be '1' for positive sentiment or '0' for negative sentiment, with no explanation included in the output."

Few-Shot Prompting for Binary Sentiment

System: "You are an assistant trained for sentiment and emotion analysis. You MUST only respond with the correct numeric label. Do not provide explanations or any additional text."

User: "Examples:
'I love this product!' => 1
'I am disappointed with the service.' => 0
Classify the following text sentiment: '{text}'"

Assistant: "If the sentiment is positive, respond with '1'. If the sentiment is negative, respond with '0'.
No other text, explanation, or formatting."

Chain-of-Thought Prompting for Emotion

System: "You are an assistant specialized in sentiment and emotion analysis. Think step-by-step through the reasoning process (chain-of-thought) privately, but provide only the final numeric classification as instructed. Do not include reasoning steps in the output."

User: "Analyze the emotion of the following text: '{text}'"
Carefully reason step-by-step to identify the best-matching emotion.
Output only the corresponding number as your final response: 0: Sadness, 1: Joy, 2: Love, 3: Anger, 4: Fear, 5: Surprise."

Assistant: "I will reason step-by-step internally to determine the most appropriate emotion.
My final response will be one of the following numbers: 0: Sadness, 1: Joy, 2: Love, 3: Anger, 4: Fear, 5: Surprise.
No reasoning will be included in the output."

Few-Shot Prompting for Emotion

System: "You are an assistant trained for sentiment and emotion analysis. You MUST only respond with the correct numeric label. Do not provide explanations or any additional text."

User: "Examples:
'This is the worst day of my life.' => 0
'I feel so joyful and alive!' => 1
'I feel so deeply connected and grateful for you in my life.' => 2
'I am so angry right now.' => 3
'I'm really scared and worried about what might happen next.' => 4
'Wow, I didn't expect that at all! This is completely unexpected!' => 5
Classify the following text emotion: '{text}'"

Assistant: "Respond only with one of the following numbers: 0: Sadness, 1: Joy, 2: Love, 3: Anger, 4: Fear, 5: Surprise.
No other text, explanation, or formatting."

D Pooling Methods

In this section, we provide a visual clarification of the different pooling strategies adopted in our work. Further details can be found in Section 3.

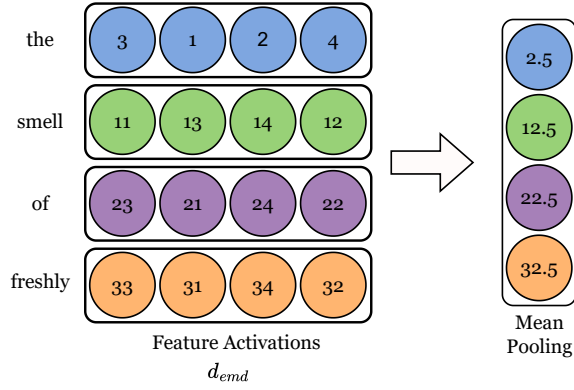


Figure 2: Mean Pooling Visually Explained

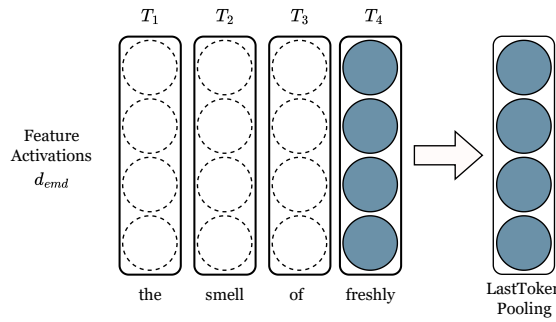


Figure 3: Last Token Pooling Visually Explained

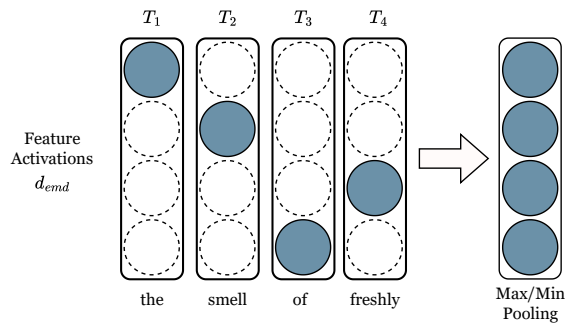


Figure 4: Max and Min Pooling Visually Explained

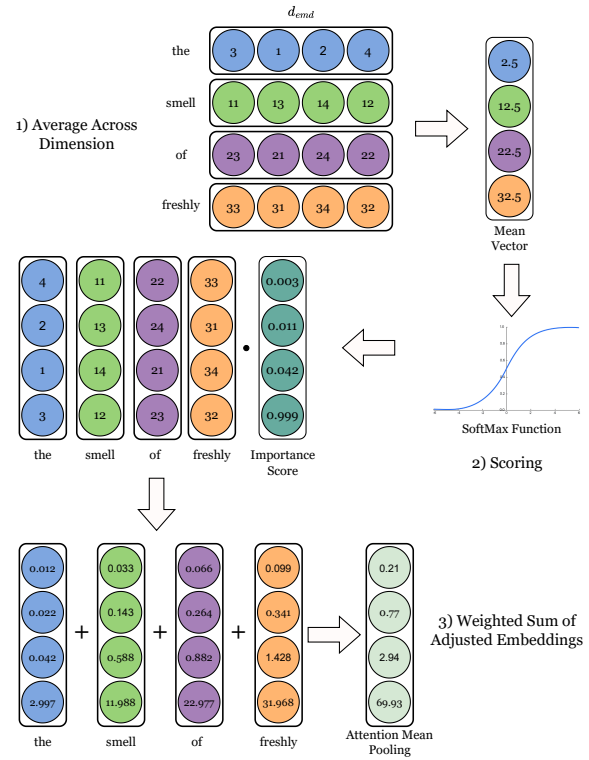


Figure 5: Attention Pooling Visually Explained

E Comprehensive Results for the Last-Token Approach

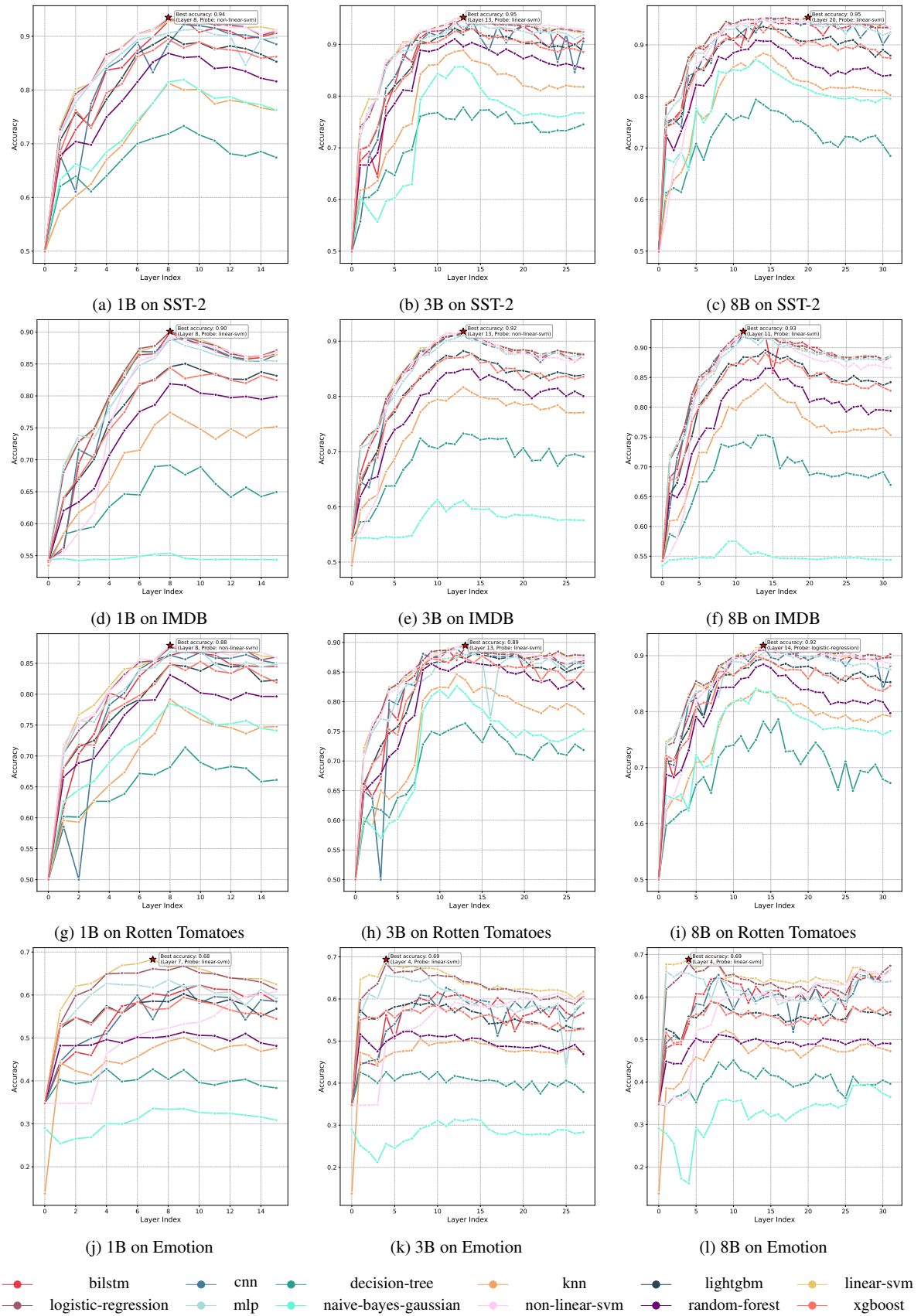


Figure 6: Layer-wise probing accuracy using the last token on SST-2, IMDB, Rotten Tomatoes, and Emotion.

F Layer-wise pooling methods confidence accuracy on SST-2, IMDB, Rotten Tomatoes, and Emotion datasets.

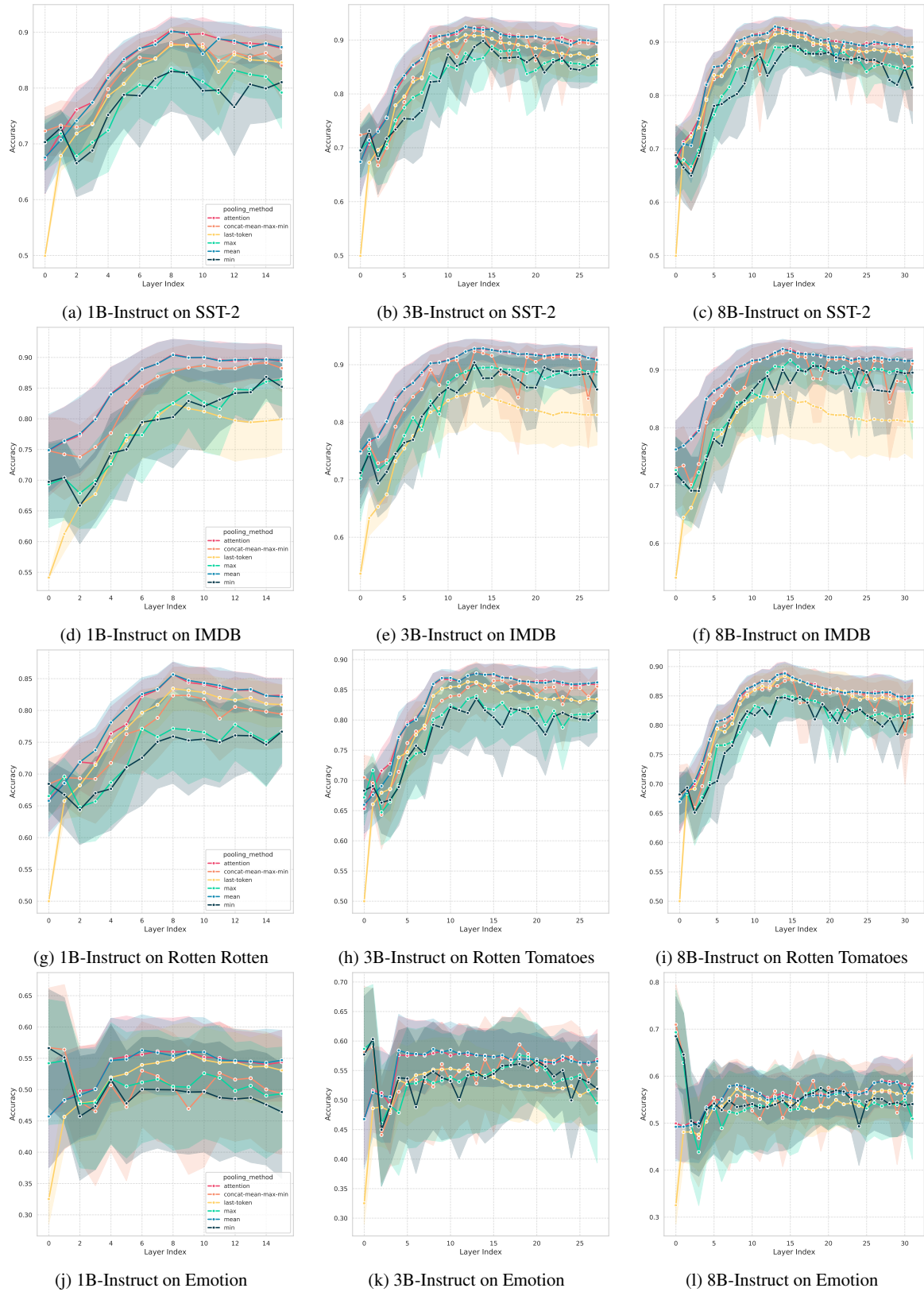


Figure 7: Layer-wise pooling methods confidence accuracy on SST-2, IMDB, Rotten Tomatoes, and Emotion.

G Computational Efficiency of SENTRILLAMA in SST-2, IMDB, Rotten Tomatoes, and Emotion datasets.

Model	Peak GPU Usage	Avg. Time per Sample	Throughput (Samples/sec)
Instruct-Llama 3.2 (1B)	4.8GB	11.33 ms	88
Instruct-Llama 3.2 (3B)	9 GB	20.83 ms	48
Instruct-Llama 3.1 (8B)	18.6 GB	35.33 ms	28
SENTRI LLAMA 3.2 (1B)	3.8 GB	7.13 ms	140
SENTRI LLAMA 3.2 (1B) Instruct	3.7 GB	6.56 ms	152
SENTRI LLAMA 3.2 (3B) Instruct	3.8 GB	4.73 ms	211
SENTRI LLAMA 3.1 (8B) Instruct	3.6GB	6.31 ms	158
DeBERTa V3 Large (418M)	2.7 GB	22.79 ms	44
RoBERTa Large (355M)	2.6 GB	8.90 ms	119

Table 7: Computational efficiency comparison of SENTRILLAMA, DeBERTa, RoBERTa, and a prompt-based method on the IMDB dataset.

Model	Peak GPU Usage	Avg. Time per Sample	Throughput (Samples/sec)
Instruct-Llama 3.2 (1B)	2.4 GB	12.01 ms	83
Instruct-Llama 3.2 (3B)	6.2 GB	18.41 ms	54
Instruct-Llama 3.1 (8B)	15.4 GB	21.17 ms	50
SENTRI LLAMA 3.2 (1B)	1.6 GB	7.08 ms	141
SENTRI LLAMA 3.2 (1B) Instruct	1.5 GB	6.42 ms	156
SENTRI LLAMA 3.2 (3B) Instruct	1.4 GB	4.80 ms	208
SENTRI LLAMA 3.1 (8B) Instruct	5.7GB	19.30 ms	97
DeBERTa V3 Large (418M)	845 MB	23.32 ms	43
RoBERTa Large (355M)	692 MB	9.05 ms	118

Table 8: Computational efficiency comparison of SENTRILLAMA, DeBERTa, RoBERTa, and a prompt-based method on the Rotten Tomatoes dataset.

Model	Peak GPU Usage	Avg. Time per Sample	Throughput (Samples/sec)
Instruct-Llama 3.2 (1B)	2.4 GB	10.97 ms	91
Instruct-Llama 3.2 (3B)	6.2 GB	17.42 ms	57
Instruct-Llama 3.1 (8B)	15.4 GB	20.53 ms	49
SENTRI LLAMA 3.2 (1B)	673 MB	2.70 ms	370
SENTRI LLAMA 3.2 (1B) Instruct	673 MB	2.07 ms	483
SENTRI LLAMA 3.2 (3B) Instruct	1 GB	2.76 ms	363
SENTRI LLAMA 3.1 (8B) Instruct	2.6GB	3.84 ms	260
DeBERTa V3 Large (418M)	844 MB	22.52 ms	44
RoBERTa Large (355M)	692 MB	8.50 ms	120

Table 9: Computational efficiency comparison of SENTRILLAMA, DeBERTa, RoBERTa, and a prompt-based method on the Emotion dataset.