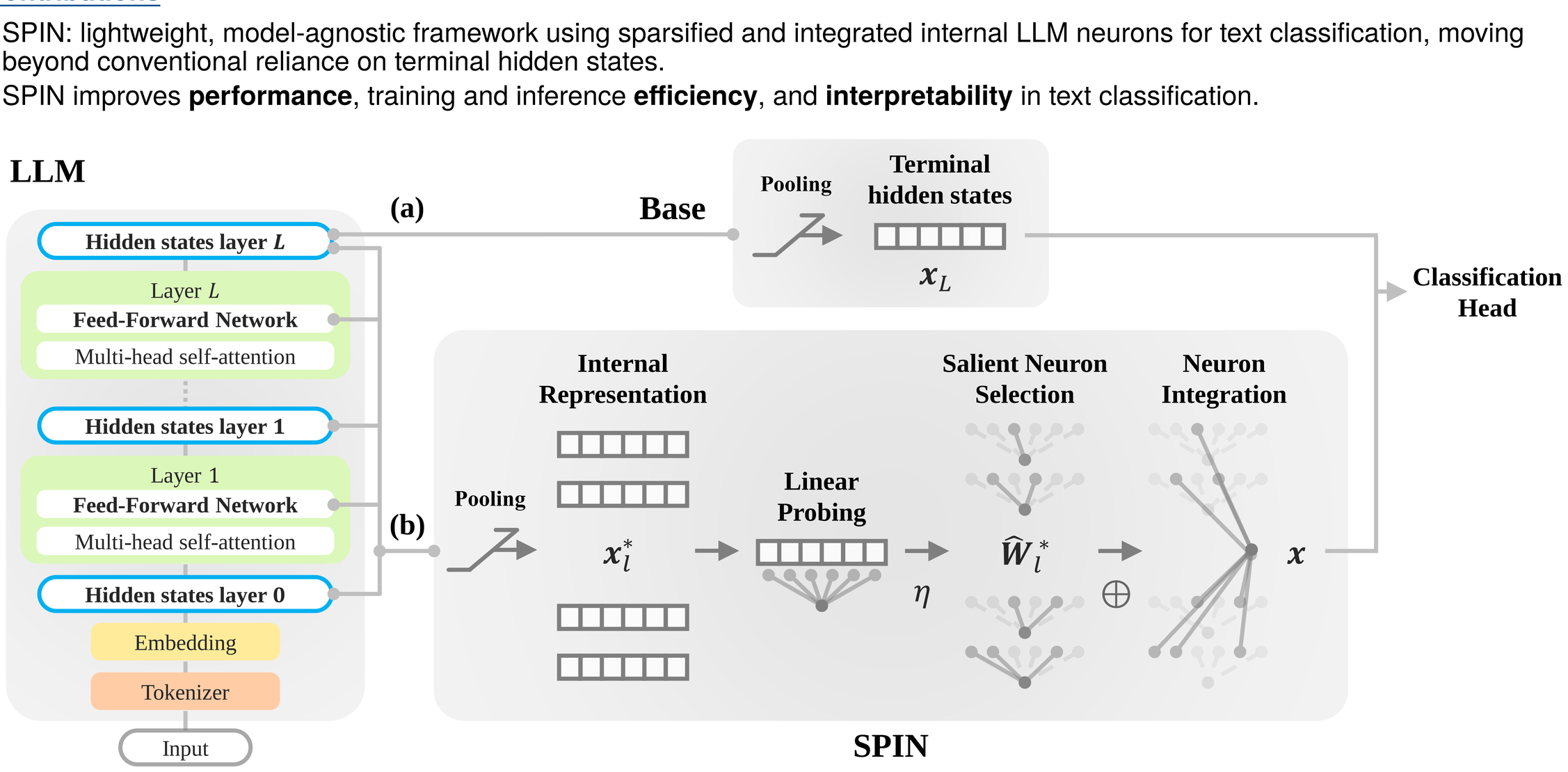
# SPIN: Sparsifying and Integrating Internal Neurons in Large Language Models for Text Classification

**Difan Jiao**[♣†] **Yilun Liu**[♦†]
**Zhenwei Tang**[♣] **Daniel Matter**[♦] **Jürgen Pfeffer**[♦] **Ashton Anderson**[♣]
[♣]University of Toronto, Canada [♦]Technical University of Munich, Germany
https://github.com/difanj0713/SPIN https://liuyilun2000.github.io/spin-visualization
[†] *Equal contribution.* difanjiao@cs.toronto.edu yilun.liu@tum.de

## Contributions

1. SPIN: lightweight, model-agnostic framework using sparsified and integrated internal LLM neurons for text classification, moving beyond conventional reliance on terminal hidden states.
2. SPIN improves **performance**, training and inference **efficiency**, and **interpretability** in text classification.




## Methodology

1. **Sparsification**: Internal neurons are useful, but not all of them.
   - Extract layer-wise internal representations from LLMs with mean-, max-pooling and single token embeddings
   - Train **layer-wise liner probes** with $\mathcal{L}^1$-regularized Logistic Regression for label over neuron activations
   - Select **salient neurons** with highest weights of linear probes until meeting the cumulative contribution threshold $\eta$
2. **Integration**: Gathering multi-grained features for classification.
   - Concatenate selected salient neurons across layers
   - Train with model classification head

## Benchmarked Performance

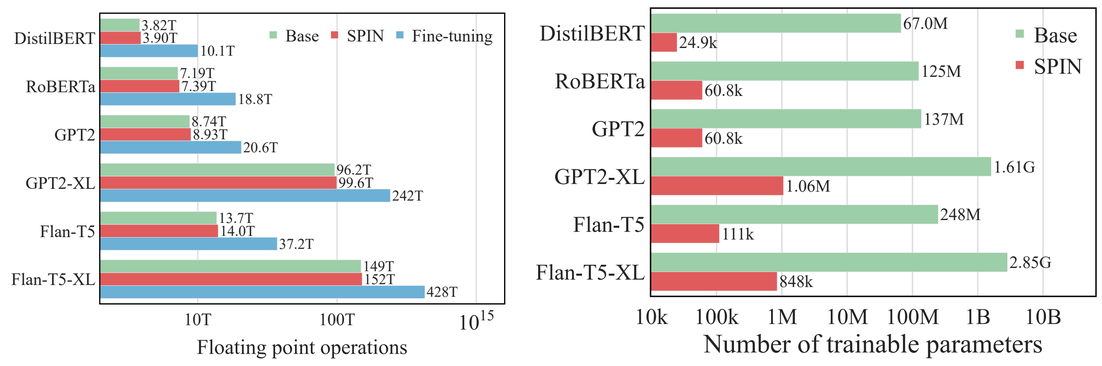
| | IMDb (Acc.) | | | SST-2 (Acc.) | | | EDOS (Macro F1) | | |
|---|---|---|---|---|---|---|---|---|---|
| | Base | SPIN | %impr. | Base | SPIN | %impr. | Base | SPIN | %impr. |
| DistilBERT | 86.95 | 89.78 | +3.25 | 81.88 | 83.94 | +2.52 | 65.09 | 75.79 | +16.44 |
| RoBERTa | 89.67 | 93.61 | +4.39 | 84.06 | 90.59 | +7.77 | 68.81 | 73.50 | +6.82 |
| GPT2 | 87.72 | 91.94 | +4.81 | 85.89 | 87.73 | +2.14 | 68.57 | 76.08 | +10.95 |
| GPT2-M | 88.59 | 93.92 | +6.02 | 86.12 | 90.25 | +4.80 | 71.17 | 75.74 | +6.42 |
| GPT2-XL | 91.86 | 94.92 | +3.33 | 90.02 | 93.23 | +3.57 | 72.56 | 76.79 | +5.83 |
| Flan-T5-S | 84.08 | 91.15 | **+8.41** | 77.17 | 88.99 | +15.32 | 59.62 | 74.51 | **+24.97** |
| Flan-T5 | 90.01 | 94.14 | +4.59 | 78.26 | 92.32 | **+17.97** | 66.64 | 78.04 | +17.11 |
| Flan-T5-XL | 90.50 | **96.12** | +6.21 | 84.75 | **95.64** | +12.85 | 70.08 | **81.48** | +16.27 |
| SoTA | 96.21 | | | 97.50 | | | 82.35 | | |

For different transformer variants, SPIN working on **pretrained** LLMs consistently outperforms baseline across all benchmark datasets, with performance even approaching fine-tuned SoTA.

| | IMDb | | SST-2 | | EDOS | |
|---|---|---|---|---|---|---|
| | Base | SPIN | Base | SPIN | Base | SPIN |
| DistilBERT | 92.80 | **92.88** | 91.05 | **91.19** | 78.74 | **81.12** |
| RoBERTa | 94.67 | **95.68** | 94.03 | **94.38** | 80.48 | **80.88** |
| GPT2 | 94.06 | **94.50** | 91.51 | **92.32** | — | — |

Introducing SPIN to already **fine-tuned** LLMs can further boost their performance.

## Training Efficiency



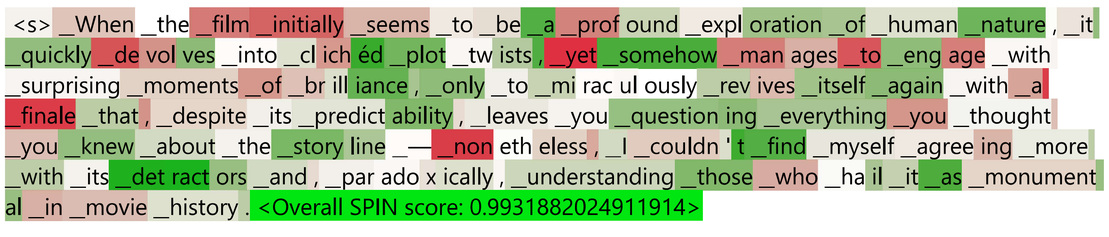

SPIN only introduces **marginal training cost** in floating point operations and trainable parameters.

## Inference Efficiency

| | 20% | 40% | 60% | 80% | 100% |
|---|---|---|---|---|---|
| DistilBERT | 85.20 | 84.73 | 87.45 | 88.67 | 89.78 |
| RoBERTa | 87.06 | 89.72 | 93.13 | 93.50 | 93.61 |
| GPT2 | 87.51 | 89.00 | 91.10 | 91.88 | 91.94 |
| GPT2-M | 88.52 | 91.36 | 93.36 | 93.92 | 93.92 |
| GPT2-XL | 89.66 | 93.15 | 94.73 | 94.92 | 94.92 |
| Flan-T5-S | 82.88 | 87.74 | 90.93 | 91.32 | 91.32 |
| Flan-T5 | 84.58 | 92.55 | 94.14 | 94.14 | 94.14 |
| Flan-T5-XL | 89.21 | 95.28 | 96.12 | 96.12 | 96.12 |

SPIN inherently supports **early-exit** by off-ramping at different percentages of LLM layers used. Averagely, 60% of layers can maintain 99% performance, reducing inference time by 40%.

## Post-hoc Interpretability




After sentence-level training, SPIN can be seamlessly **transferred to token-level classification**, providing a detailed breakdown of cumulative results within a single forward pass.