Ranking Attention Heads by Operational Contribution Using Few-Shot In-Context Learning

Priyanshu Raj Mall

June 21, 2025

1 Approach

To identify and rank the operational importance of attention heads in the final layer of a transformer-based language model, we adopted a perturbation-based evaluation framework in a few-shot in-context learning (ICL) setup. The process was implemented in Python using the HuggingFace `transformers` libraries. The major steps of our approach are as follows:

1.1 Dataset Preparation

We used the given IMDB sentiment classification dataset (99 samples) containing labeled movie reviews. The dataset was divided into two subsets:

- A few-shot context set of 4 samples (2 positive, 2 negative), used to build a prompt for in-context learning.
- A test set, comprising the remaining examples, used to evaluate model predictions.

Each test prompt was constructed by concatenating the few-shot context with the new review, prompting the model to generate the sentiment label (`positive` or `negative`).

1.2 Model Selection and Loading

We selected `SmolLM2-135M` a lightweight LLaMA-based causal language model as our base model, suitable for full attention head access for experimentation, and also with context length 8192 to be able to accommodate 4-shot to 8-shot examples in the prompt. The corresponding tokenizer and model were loaded from HuggingFace and deployed to a GPU for faster inference.

1.3 Baseline Evaluation

We evaluated the model's original performance on the test set using the constructed few-shot prompts. The evaluation metrics—accuracy and weighted F1-score—were recorded as the baseline (`result_original`).

1.4 Attention Head Perturbation

To analyze each attention head's contribution in the last layer, we modified the last layer's attention mechanism by introducing a custom subclass `PerturbedLlamaAttention`, which extends the original `LlamaAttention` class. This subclass allows injection of controlled Gaussian noise into the output of a specific attention head during the forward pass.

The perturbation process involved:

- Iteratively selecting each attention head `i` in the final layer.
- Injecting random noise into that head's output with magnitude set to a fixed fraction (default 0.5) of the maximum absolute activation value in that head.
- Evaluating the perturbed model's performance on the same test set with identical prompts as the baseline.

After each perturbation, the original attention layer was restored to maintain isolation of effects.

1.5 Contribution Scoring

For each head `i`, we measured the drop in performance metrics (accuracy and F1-score) relative to the baseline. These drops served as the operational contribution scores, quantifying the importance of each head to the model's task-specific performance.

1.6 Systematic Evaluation Across Noise Levels and Noise Seeds

To account for stochasticity in noise injection and obtain more reliable importance estimates, we conducted a grid of experiments with:

- Noise levels ranging from 0.1 to 0.7 (in steps of 0.1)
- Random seeds for random gaussian noise ranging from 0 to 20

This resulted in 210 runs (7 noise levels $\times$ 21 seeds). Each run generated a CSV file logging performance drops for all attention heads under that specific configuration.

1.7 Averaging and Robust Ranking

To mitigate the effect of random variations across seeds, we aggregated results for each attention head across all 21 seeds at a fixed noise level (e.g., 0.7). We computed the mean drop in F1-score and accuracy per head and re-ranked them based on these average drops. This produced a more stable and noise-resilient ranking of attention heads.

The aggregation process involved:

- Computing mean F1 and accuracy drops for each head.
- Sorting the heads by their average accuracy drop.

The final averaged rankings were written to a CSV and printed in tabular format for analysis.

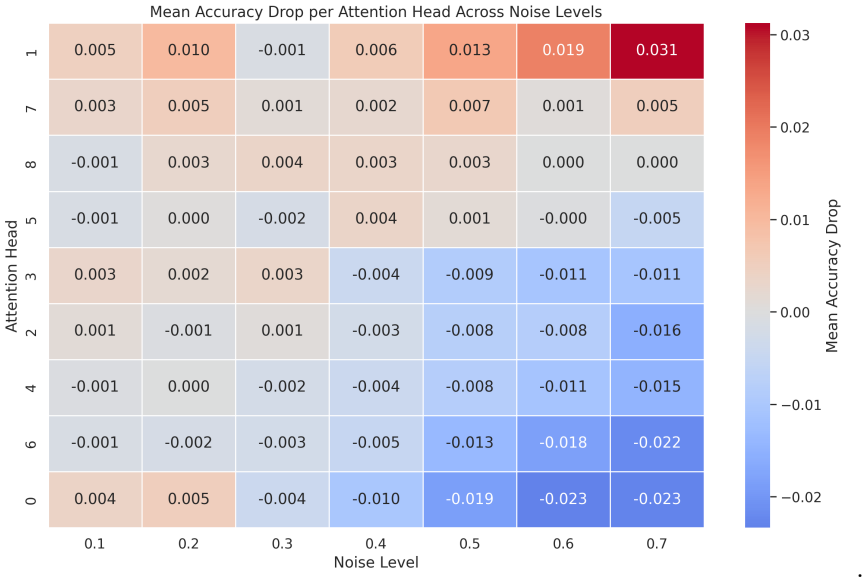
2 Results

Baseline → Accuracy: 0.7158, F1-Score: 0.6747

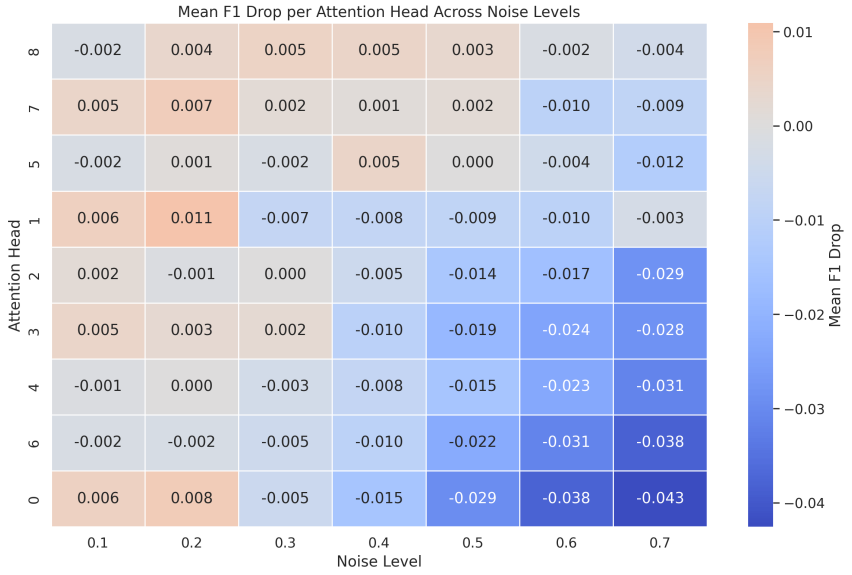
Rank	Head	Mean F1 Drop	Mean Accuracy Drop
1	1	-0.00852	0.01340
2	7	0.00156	0.00670
3	8	0.00346	0.00287
4	5	0.00026	0.00128
5	2	-0.01406	-0.00797
6	4	-0.01494	-0.00829
7	3	-0.01879	-0.00893
8	6	-0.02249	-0.01340
9	0	-0.02948	-0.01850

Table 1: Mean F1 Drop and Mean Accuracy Drop for Different Heads for noise level 0.5

For rest of the noise levels, result tables for each are in the Appendix section.



(a) Mean accuracy drop per attention head across noise levels.



(b) Mean F1-score drop per attention head across noise levels.

Figure 1: **Heatmaps of Mean Performance Drop per Attention Head across Noise Levels.** These visualizations present the operational impact of perturbing individual attention heads in the final layer of the model. Each cell shows the average drop in performance metric—(top) accuracy and (bottom) weighted F1-score—when Gaussian noise is injected into a specific head at a given noise level, averaged over 21 random seeds. Attention heads are sorted by their mean contribution across all noise levels, with the most influential heads (those causing the highest average performance drop) appearing at the top. The redder cells indicate stronger negative impact on performance, signifying high operational importance, whereas blue or near-zero cells indicate minimal or even beneficial effect from perturbation. These heatmaps help identify critical heads responsible for task-specific behavior in a few-shot In-Context Learning (ICL) setting.

3 Observations

The experiment systematically perturbed each attention head in the final layer of the model across a range of noise levels (from 0.1 to 0.7), with the impact averaged over 21 different random seeds to ensure robustness. The following key observations were drawn from the aggregated results:

1. **Performance Degradation Correlates with Noise Level:**
   As expected, the average drop in both F1-score and accuracy tends to increase with higher noise levels. For instance, at low perturbation levels (noise_level = 0.1), the performance impact is minimal (mean accuracy drop ≈ 0.00461). At noise_level = 0.7, the average accuracy drop increases to ≈ 0.03125, indicating stronger disruption to the model's predictions.

2. **Non-Monotonic Behavior at Moderate Noise Levels:**
   Interestingly, the performance trend is not strictly monotonic. Between noise_level = 0.2 and 0.4, the average impact slightly decreases, and even becomes negative at noise_level = 0.3 and 0.4 (i.e., model accuracy improves under perturbation). This could indicate that light perturbations in some attention heads act as a form of regularization or break spurious correlations.

3. **Head Importance Becomes More Distinct at Higher Noise:**
   As the perturbation intensity increases, the range of contribution scores across attention heads widens. At low noise levels, all heads exhibit similar impact, making it difficult to identify the most influential ones. However, at higher noise levels (≥ 0.6), the top-contributing heads emerge clearly due to their consistently large performance degradation.

4. **Averaging Over Seeds Yields Robust Insights:**
   Aggregating the results across 21 random seeds significantly reduces stochastic noise and stabilizes contribution estimates. This ensures that the final ranking of heads reflects genuine importance rather than random variation, improving the reliability of interpretability conclusions.

5. **Accuracy increases for several head perturbations:**
   Counterintuitively, injecting random noise into nearly half of the attention heads results in improved accuracy (e.g., Table 1). This might be because, in few-shot ICL, models can easily overfit to irrelevant context tokens or positional biases through certain attention heads. Perturbing these heads could disrupt such overfitting, thereby enhancing generalization.

4 Appendix

Rank	Head	Mean F1 Drop	Mean Accuracy Drop
1	1	0.00644	0.00461
2	0	0.00591	0.00395
3	3	0.00453	0.00296
4	7	0.00450	0.00296
5	2	0.00152	0.00099
6	4	-0.00082	-0.00066
7	8	-0.00184	-0.00132
8	5	-0.00178	-0.00132
9	6	-0.00180	-0.00132

Table 2: Mean F1 Drop and Mean Accuracy Drop for Different Heads for noise level 0.1

Rank	Head	Mean F1 Drop	Mean Accuracy Drop
1	1	0.01091	0.00987
2	0	0.00800	0.00526
3	7	0.00749	0.00526
4	8	0.00401	0.00263
5	3	0.00269	0.00197
6	5	0.00070	0.00033
7	4	0.00021	0.00000
8	2	-0.00080	-0.00066
9	6	-0.00221	-0.00164

Table 3: Mean F1 Drop and Mean Accuracy Drop for Different Heads for Noise Level 0.2

Rank	Head	Mean F1 Drop	Mean Accuracy Drop
1	8	0.00543	0.00362
2	3	0.00244	0.00329
3	7	0.00186	0.00099
4	2	0.00018	0.00066
5	1	-0.00749	-0.00099
6	5	-0.00203	-0.00164
7	4	-0.00342	-0.00230
8	6	-0.00483	-0.00263
9	0	-0.00543	-0.00395

Table 4: Mean F1 Drop and Mean Accuracy Drop for Different Heads for Noise Level 0.3

Rank	Head	Mean F1 Drop	Mean Accuracy Drop
1	1	-0.00781	0.00592
2	5	0.00479	0.00362
3	8	0.00479	0.00329
4	7	0.00119	0.00230
5	2	-0.00526	-0.00263
6	4	-0.00762	-0.00395
7	3	-0.00984	-0.00428
8	6	-0.00971	-0.00493
9	0	-0.01497	-0.01020

Table 5: Mean F1 Drop and Mean Accuracy Drop for Different Heads for Noise Level 0.4

Rank	Head	Mean F1 Drop	Mean Accuracy Drop
1	1	-0.00969	0.01908
2	7	-0.00974	0.00066
3	8	-0.00159	0.00000
4	5	-0.00431	-0.00033
5	2	-0.01661	-0.00822
6	3	-0.02404	-0.01086
7	4	-0.02311	-0.01118
8	6	-0.03139	-0.01842
9	0	-0.03777	-0.02336

Table 6: Mean F1 Drop and Mean Accuracy Drop for Different Heads for Noise Level 0.6

Rank	Head	Mean F1 Drop	Mean Accuracy Drop
1	1	-0.00298	0.03125
2	7	-0.00866	0.00526
3	8	-0.00408	0.00000
4	5	-0.01190	-0.00526
5	3	-0.02823	-0.01086
6	4	-0.03137	-0.01480
7	2	-0.02856	-0.01579
8	6	-0.03831	-0.02204
9	0	-0.04257	-0.02303

Table 7: Mean F1 Drop and Mean Accuracy Drop for Different Heads for Noise Level 0.7