Demystifying When Pruning Works via Representation Hierarchies

Shwai He, Guoheng Sun, Haichao Zhang, Yun Fu, Ang Li

University of Maryland, College Park

Northeastern University

Discrepancies in Network Pruning across Tasks

- Non-generative

Stable in embedding and multi-choice tasks

- Generative

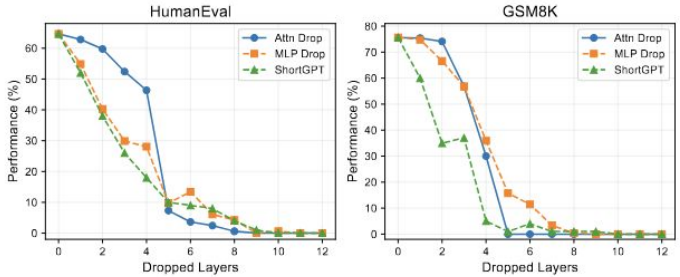
Fails under moderate pruning

E5-Mistral #Params	Full-Model 7.1B	Drop-8A 6.8B	Drop-8M 5.7B
Embedding Tasks			
Arguana	60.9	54.7	58.6
Climate-FEVER	36.8	31.9	38.4
DBPedia	47.9	43.6	44.1
FEVER	87.6	82.9	88.7
FiQA	56.4	50.9	52.8
HotpotQA	74.9	66.8	74.2
NFCorpus	38.1	35.4	36.9
NQ	66.3	56.1	65.4
Quora	88.6	86.5	88.2
SCIDOCS	16.2	12.4	14.7
SciFact	75.8	71.4	73.6
TREC-COVID	85.9	84.3	79.6
Touche-2020	22.9	18.1	18.7
Average	58.9	53.4	56.8

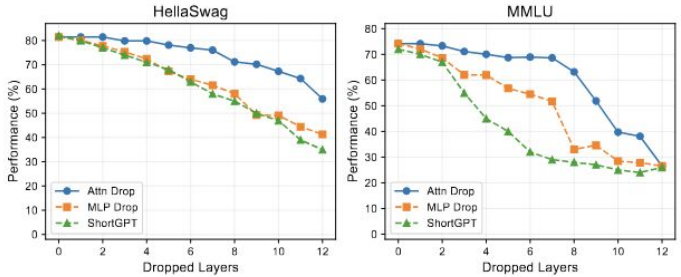
(a) Retrieval performance of E5-Mistral (Wang et al., 2024).

Mistral-7B-Instruct #Params	Full-Model 7.1B	Drop-8A 6.8B	Drop-8M 5.7B
Multi-choice Tasks			
BoolQ	85.9	86.0	78.2
MMLU	62.1	62.0	59.1
OpenBookQA	47.0	46.8	41.2
RTE	72.9	74.0	72.1
Winogrande	78.8	80.0	71.1
Average	69.3	69.8	64.3
Generation Tasks			
GSM8K	48.4	36.2	0.0
HumanEval	4.9	0.0	0.0
MBPP	13.8	0.4	0.0
NarrativeQA	16.3	9.6	2.0
NQ-Open	27.9	20.9	2.0
Average	22.3	13.2	0.8

(b) Benchmarks of Mistral-7B (Jiang et al., 2023).



(a) Generative tasks.



(b) Non-generative tasks.

Case Study of Output Examples After Pruning

- Moderate Pruning:

  The model retains basic language capabilities and produces reasonable answers.

- Aggressive Pruning

  Generation degrades, exhibiting repetitive or irrelevant outputs.

Question	Natalia sold clips to 48 of her friends in April, and then she sold half as many clips in May. How many clips did Natalia sell altogether in April and May?
Method	**Output**
Baseline	Natalia sold 48 clips in April. In May, she sold half as many clips as in April, which amounts to 24 clips. Therefore, the total number of clips sold across April and May is $48 + 24 = 72$.
Drop-4A	Natalia sold 48 clips in April and half as many in May, which is 24 clips. Therefore, the total number of clips sold over April and May is $48 + 24 = 72$.
Drop-8A	Let's denote the number of clips Alicia, the ILUNNIE M '<%=>t \[K1205359752000_1_61.7251015512555646044587492305 2e-08T.1000.001.10115651362430000.2023-01-02 23:21:07 20227744796744571555 6447
Drop-4M	Natalia sold 48 clips in April and half that amount in May (24 clips), resulting in a total of $48 + 24 = 72$ clips sold across the two months.
Drop-8M	To calculate the total number of clips, we are adding the result of first you and your a year and then the second or your and your a year and your and your and your and your and your and your and your...

Demystifying When Pruning Works via Representation Hierarchies

Embedding:

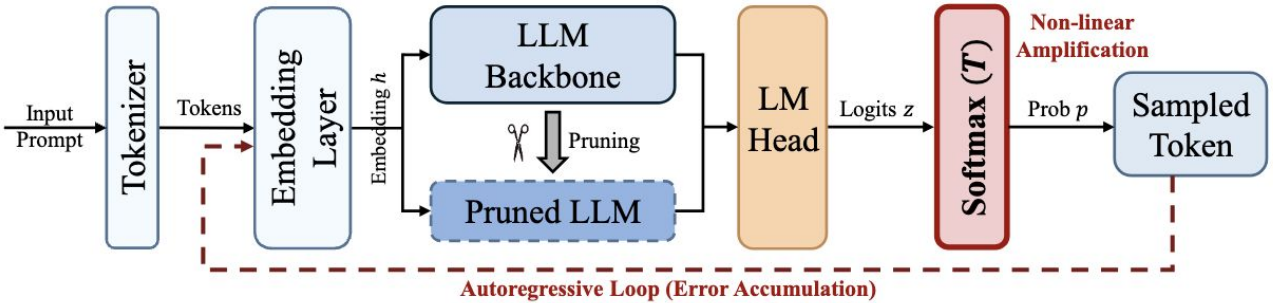
$$h^{(l)} = f^{(l)}\left(h^{(l-1)}\right)$$

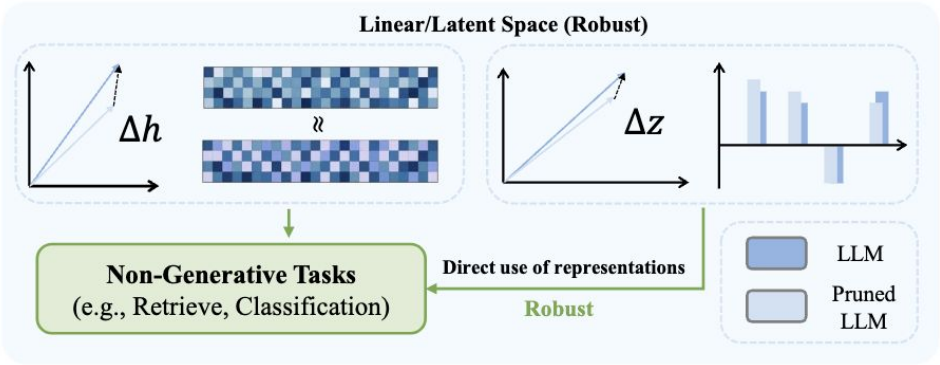
Logits:

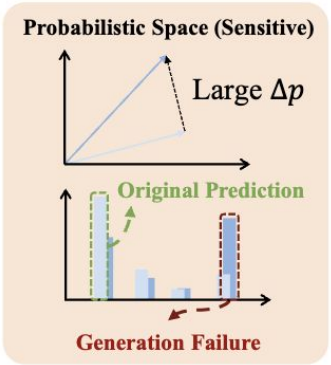
$$z = Wh^{(L)}$$

Vocabulary:

$$p_{i+1} = \mathrm{softmax}(z_i/T)$$







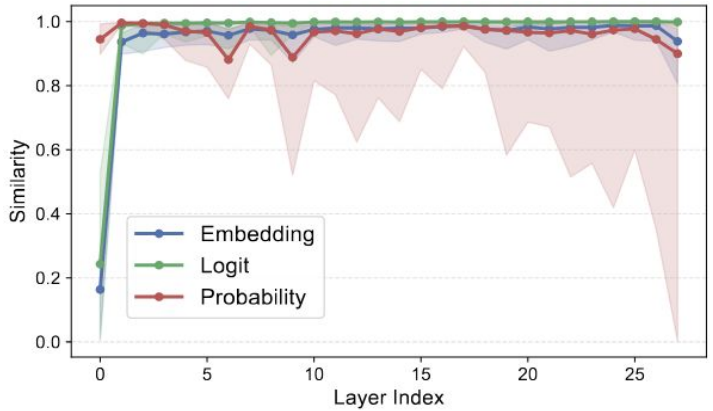
Distinct Behaviors across Representation Spaces

$$z = Wh^{(L)} \quad \rightarrow \quad p_{i+1} = \mathrm{softmax}(z_i/T) \quad \rightarrow \quad \begin{aligned} \hat{x}_{i+1} &= \mathrm{Sample}(p_{i+1}) \\ \hat{\mathcal{T}}_{i+1} &= \tau^{-1}(\hat{x}_{i+1}) \end{aligned}$$

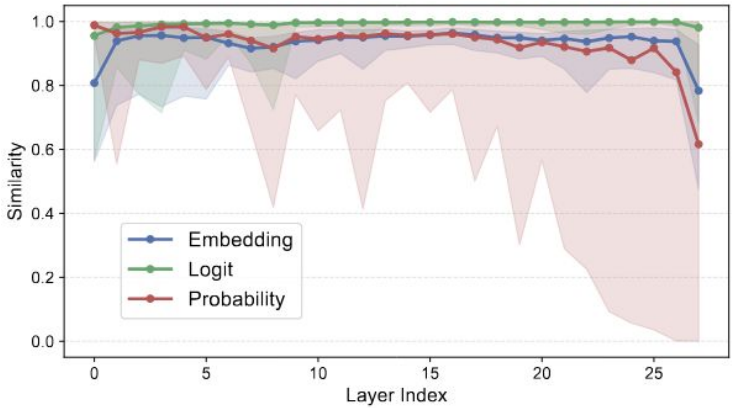
Metric: $\mathrm{CosineSim}(h_l, h_l + \Delta h_l)$ – similarity between outputs from baseline and pruned model.

Embedding / Logit: Consistently high similarity, indicating strong robustness to pruning.

Probability: Substantially lower similarity, which can drop to extremely low values.



(a) Attention.



(b) MLP.

Pruning-Induced Perturbations in Embedding Space

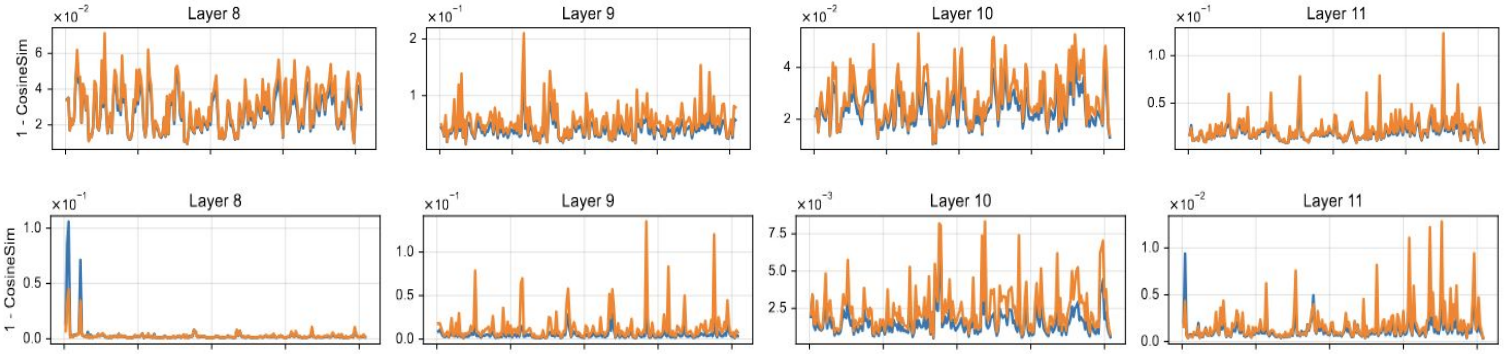
Given the high similarity observed, pruning-induced deviations lie in a local neighborhood of the original representation. Therefore, they can be analyzed using a second-order Taylor expansion:

$$1 - \mathrm{CosineSim}(h, h + \Delta h) \approx \frac{\|\Delta h_\perp\|^2}{2\|h\|^2} \qquad 1 - \mathrm{CosineSim}(z, z + \Delta z) \approx \frac{\|\Delta z_\perp\|^2}{2\|z\|^2}$$

The estimated values closely match the ground truth in most cases.

Embedding Space:

Logit Space:

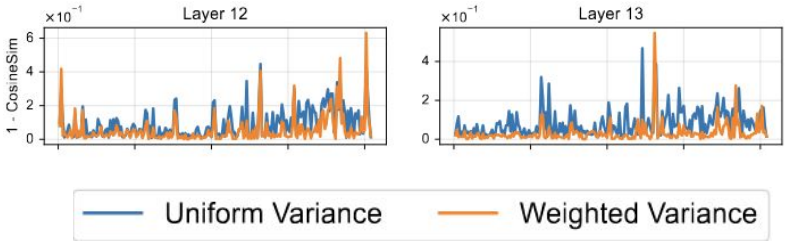


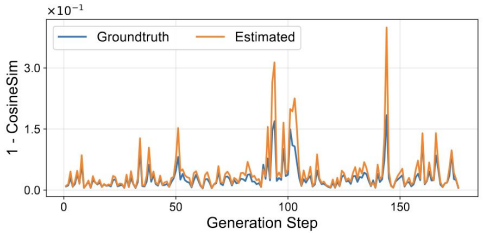
Pruning-Induced Perturbations in Probability Space

The cosine similarity and KL Divergence of **probability space** can be analyzed using a second-order Taylor expansion:

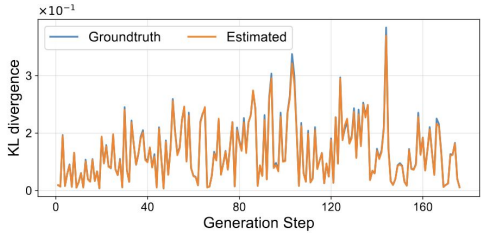
$$1 - \mathrm{CosineSim}(p, p + \Delta p) \approx \frac{\mathrm{Var}_r(\Delta z)}{2T^2}, \; r_i = \frac{p_i^2}{\|p\|^2} \qquad \mathrm{KL}(p\|q) \approx \frac{\mathrm{Var}_{i\sim p}(\Delta z_i)}{2T^2}$$

Not only the orthogonal component but also the parallel component contributes to the deviation. In practice, the variance of the parallel component can be substantial and often exceeds that of the orthogonal component.





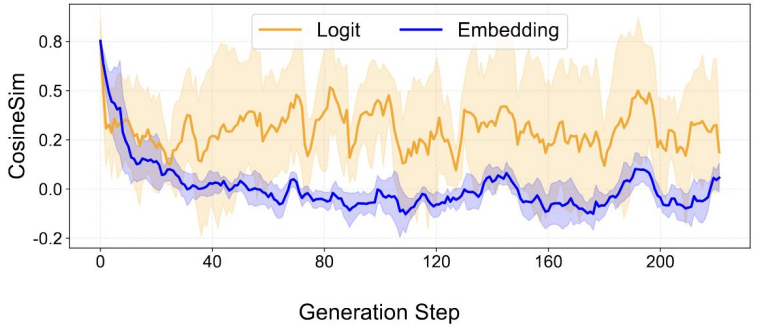
(a) Angular Deviation.



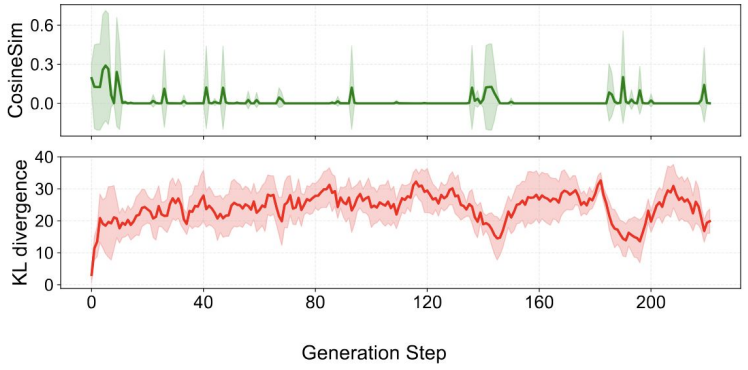
(b) KL divergence.

Temporal Propagation of Perturbations

- At the first generation step, the representations exhibit high similarity across all three spaces.
- As generation proceeds, the similarity gradually decreases in the embedding space but drops sharply in the probability space.
- The logit space, by contrast, sometimes maintains relatively high similarity.
- Meanwhile, the KL divergence remains consistently high across most generation steps.



(a) Embedding and Logit Spaces.



(b) Probability Space.

Conclusion

1. **Network pruning exhibits a clear task-dependent behavior**: it remains effective for non-generative tasks but degrades sharply for generative tasks.

2. **Representation-level analysis reveals the root cause**: pruning impacts embedding, logit, and probability spaces in fundamentally different ways.

3. **For generative tasks**, the nonlinear projection from logits to probabilities (softmax) **amplifies small perturbations**, which accumulate through autoregressive decoding and destabilize generation.

4. **For non-generative tasks**, the relative stability of embedding/logit spaces and the categorical-token probability **subspace** explains why pruning remains effective.

5. **Practical takeaway**: pruning strategies should be **task-aware**: safe for embedding-based or subspace-limited inference, but risky for full-vocabulary autoregressive generation.