Towards the Use of Layer-to-Layer Stability Patterns for Early Accuracy Estimation in Question Answering

Minjoon Choi

Seoul National University, Seoul, South Korea

minjoonchoi08@snu.ac.kr

❑ Summary

The similarity between consecutive hidden representations across Large Language Model (LLM) transformer layers follows a consistent trajectory: the similarity is low (unstable) in early layers, rises to a peak (stable) around the 70-80th percentile layers, and then drops sharply at the final layers. But similarity alone weakly predicts accuracy in LLM question answering (QA) tasks.

❑ Why is early estimation in QA beneficial?

	Natural Questions	TriviaQA	GOOAQ
Generating	1.18 iter / sec	1.30 iter / sec	1.19 iter / sec
Probing Hidden States Before Generating	61.21 iter / sec	61.08 iter / sec	63.98 iter / sec

Model: meta-llama/Meta-Llama-3-8B

	Natural Questions	TriviaQA	GOOAQ
Generating	0.16 iter / sec	0.18 iter / sec	0.16 iter / sec
Probing Hidden States Before Generating	24.20 iter / sec	23.93 iter / sec	26.10 iter / sec

Model: meta-llama/Meta-Llama-3-70B

❑ Why should we look at mid-to-late-layer hidden representations?

- Can evaluate how knowledgeable an LLM is about a given subject entity by only considering how it processes the name of that entity, before generating tokens (Gottesman et al., 2024)[1].
- Intermediate layers often surpass the final layer by up to 16% in downstream task accuracy (Skean et al., 2025)[2].
- While middle layers capture essential reasoning information, it may not be fully utilized or maintained by the later layers, potentially impacting the model's reasoning performance (Xie et al., 2024)[3].
- The attributes rate at the last-subject position is substantially high in the intermediate-upper layers (Geva et al., 2023)[4].

❑ Datasets

- Natural Questions (Kwiatkowski et al., 2019)[5]
- TriviaQA (Joshi et al., 2017)[6]
- GOOAQ (Khashabi et al., 2021)[7]

❑ Models

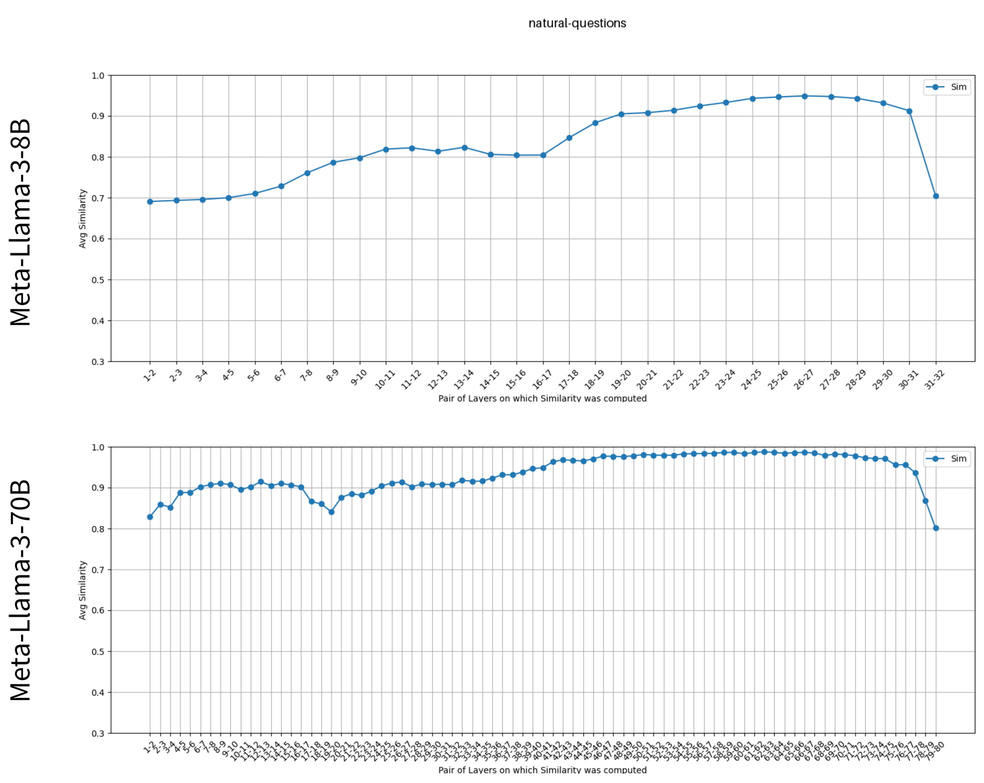
- Llama: Meta-Llama-3-8B, Meta-Llama-3-70B
- Mistral: Mistral-7B-v0.3, Mistral-Nemo-Base-2407 (12B), Mistral-Small-24B-Base-2501
- Qwen: Qwen3-8B, Qwen3-14B, Qwen3-32B

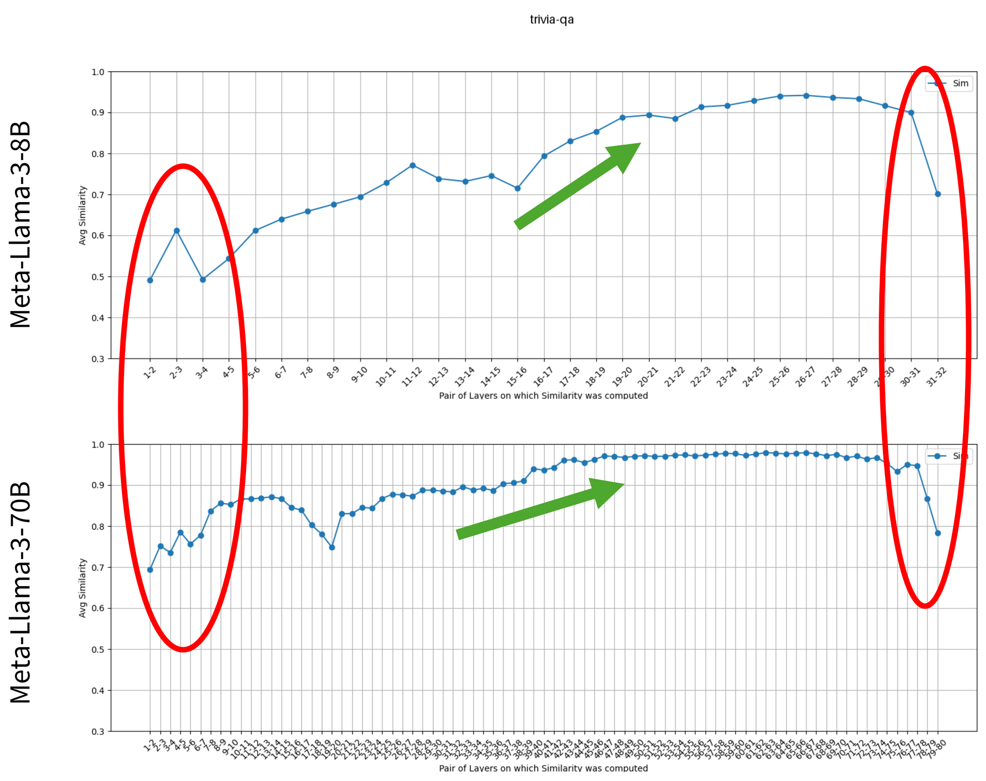
❑ Hidden Representation Probing Method

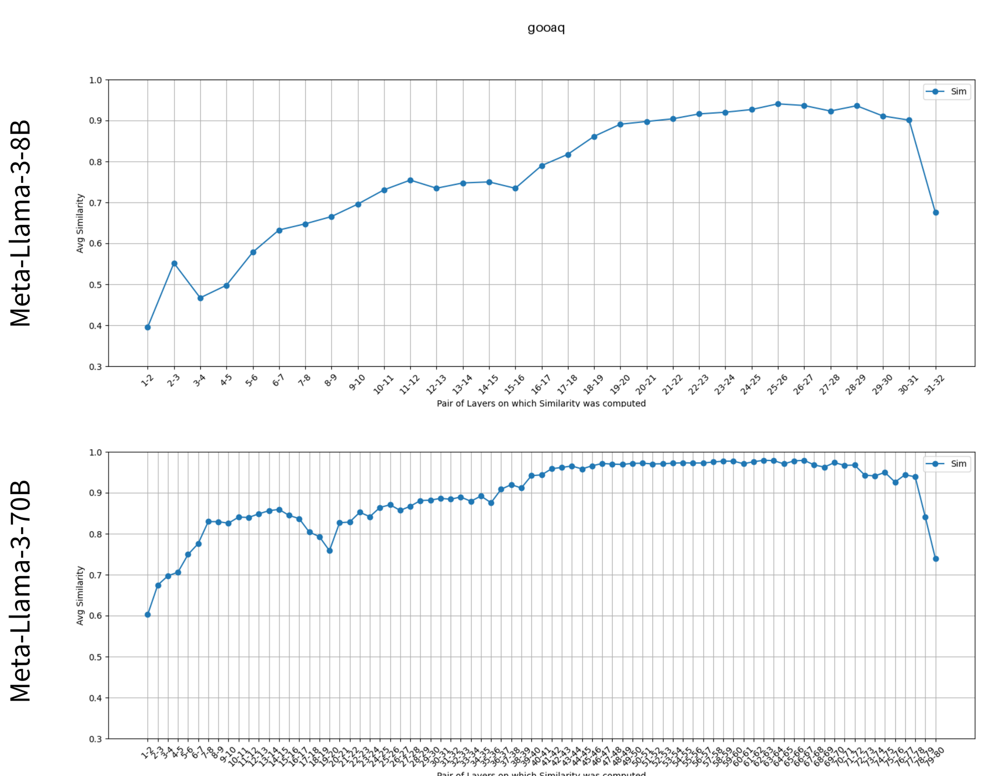
- For each layer $l \in \{1, 2, \dots, L\}$, probe the raw hidden state h_l
  - Obtain the hidden state of the last token of input sequence, following previous works[4,8-9]
- $\forall l \in \{1, 2, \dots, L\}$, apply standardization on h_l vectors along with h_l vectors probed from other examples → Obtain standardized hidden state $\widetilde{h_l}$

❑ Computing Layer-by-Layer Hidden Representation Similarities

- $\forall l \in \{1, 2, \dots, L-1\}$, compute $sim(\widetilde{h_l}, \widetilde{h_{l+1}})$
- $sim(\widetilde{h_l}, \widetilde{h_{l+1}})$: cosine similarity between consecutive layers



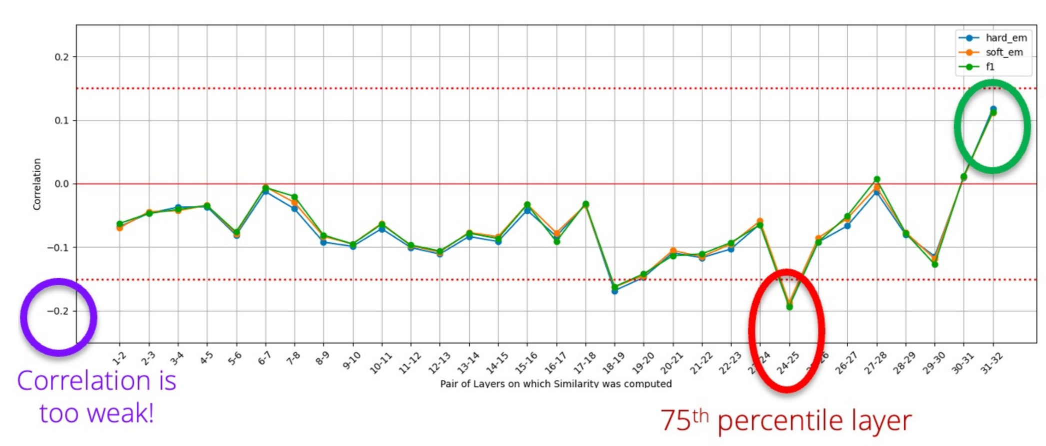




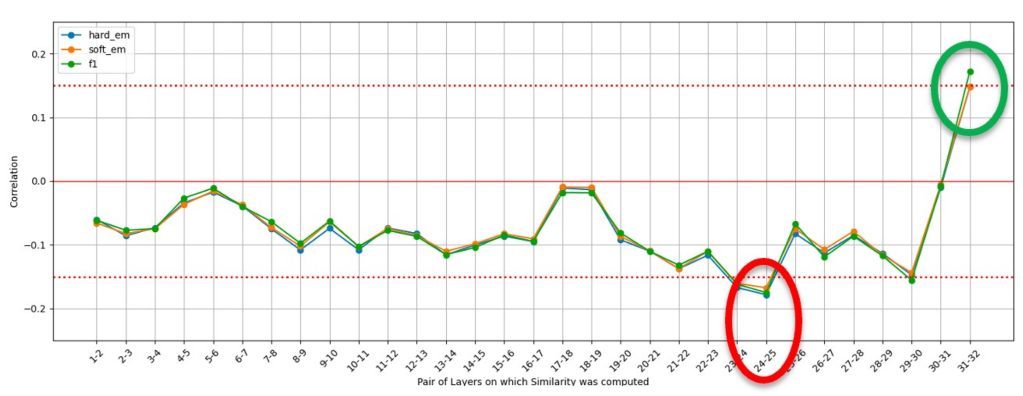
❑ Can these layer-to-layer stability patterns be used to predict QA accuracy?

- To check the ability of hidden representation similarities to predict QA accuracy,
- $\forall l \in \{1, 2, \dots, L-1\}$, we compute the correlation between
  1) $sim(\widetilde{h_l}, \widetilde{h_{l+1}})$
  2) Closed-book QA Accuracy
  - Possible Metrics: Hard Exact Match (EM), Soft EM, 1-gram (word-level) F1 Score

Model: Mistral-7B-v0.3 Dataset: TriviaQA



Model: Meta-Llama-3-8B Dataset: TriviaQA



- We trained Logistic Regression models that predict the Hard/Soft Exact Match label of a question example, given the observed layer-to-layer pair similarities.
- Performance of trained regression models

```
metric_to_df['accuracy']
```

Low Accuracy

	nq	triv	nq-triv
meta-llama/Meta-Llama-3-8B	0.8112 / 0.7529	0.6483 / 0.669	0.6921 / 0.6924
mistralai/Mistral-7B-v0.3	0.7692 / 0.7366	0.6726 / 0.688	0.6831 / 0.6914
Qwen/Qwen3-8B	0.831 / 0.7774	0.627 / 0.6262	0.6777 / 0.6649

```
metric_to_df['precision']
```

	nq	triv	nq-triv
meta-llama/Meta-Llama-3-8B	nan / nan	0.6566 / 0.6772	0.6697 / 0.6869
mistralai/Mistral-7B-v0.3	nan / nan	0.6783 / 0.6935	0.6769 / 0.6968
Qwen/Qwen3-8B	nan / nan	0.6477 / 0.6461	0.6489 / 0.6519

```
metric_to_df['recall']
```

The trained model always predicted FALSE

	nq	triv	nq-triv
meta-llama/Meta-Llama-3-8B	0.0 / 0.0	0.936 / 0.9629	0.7881 / 0.813
mistralai/Mistral-7B-v0.3	0.0 / 0.0	0.9522 / 0.9701	0.7883 / 0.8045
Qwen/Qwen3-8B	0.0 / 0.0	0.5463 / 0.66	0.4515 / 0.539

```
metric_to_df['f1']
```

	nq	triv	nq-triv
meta-llama/Meta-Llama-3-8B	0.0 / 0.0	0.7718 / 0.7952	0.7241 / 0.7447
mistralai/Mistral-7B-v0.3	0.0 / 0.0	0.7923 / 0.8088	0.7284 / 0.7468
Qwen/Qwen3-8B	0.0 / 0.0	0.5927 / 0.653	0.5325 / 0.5901

Each cell corresponds to:
("when predicting Hard EM" / "when predicting Soft EM")

❑ Conclusions

- Across Llama, Mistral, and Qwen3 model families on commonsense knowledge-intensive benchmarks such as Natural Questions, TriviaQA, and GOOAQ, the layer-to-layer hidden representation similarities exhibit a consistent pattern: low similarity values are observed in early layers, these values peak around 70-80th percentile layers, and drop sharply at the final layers. The observation where hidden representations exhibit solid stability around mid-to-late-layers aligns with previous works[2-4] that claim the significance of these layers.
- Consecutive layer similarities do not have significant correlation with accuracy in QA tasks. Regression models that were trained with a purpose of estimating QA accuracy before generation also underperformed. These results indicate that while layer-to-layer stability is an easy-to-access and fast-to-compute signal, it is insufficient on its own for reliable early accuracy estimation.

❑ Acknowledgements

This work was supported by Korea Institute for Advancement of Technology (KIAT) grant funded by the Korea Government (Ministry of Education) (P0025681-G02P22450002201-10054408, Semiconductor-Specialized University)

❑ References

[1] Gottesman et al., Estimating Knowledge in Large Language Models Without Generating a Single Token, EMNLP 2024

[2] Skean et al., Layer by Layer: Uncovering Hidden Representations in Language Models, ICML 2025

[3] Xie et al., Calibrating Reasoning in Language Models with Internal Consistency, NeurIPS 2024

[4] Geva et al., Dissecting Recall of Factual Associations in Auto-Regressive Language Models, EMNLP 2023

[5] Kwiatkowski et al., Natural Questions: A Benchmark for Question Answering Research, TACL 2019

[6] Joshi et al., TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension, ACL 2017

[7] Khashabi et al., GOOAQ : Open Question Answering with Diverse Answer Types, EMNLP Findings 2021

[8] Goloviznina et al., I've got the "Answer"! Interpretation of LLMs Hidden States in Question Answering, NLDB 2024

[9] Meng et al., Locating and Editing Factual Associations in GPT, NeurIPS 2022

서 울 대 학 교
SEOUL NATIONAL UNIVERSITY