Layer by Layer

Uncovering Hidden Representations in Language Models

Presented by: Oscar Skean

Oscar Skean

University of Kentucky

Md Rifat Arefin

Université de Montréal Mila

Dan Zhao

NYU

Niket Patel

UCLA

Jalal Naghiyev

Yann LeCun

NYU Meta

Ravid Shwartz-Ziv

NYU wand

Overview of the Work

- Our work **challenges common assumptions** in modern ML folklore
  - ✗ Myth 1: Final layers always give the best embeddings
  - ✗ Myth 2: Middle layers are useless for downstream tasks
- ✓ Reality : Intermediate layers often outperform final layers

Overview of the Work

- Embeddings of **intermediate layers outperform final layers** on downstream tasks

- Rigorous **empirical testing** across model architectures, scales, tasks, and modalities

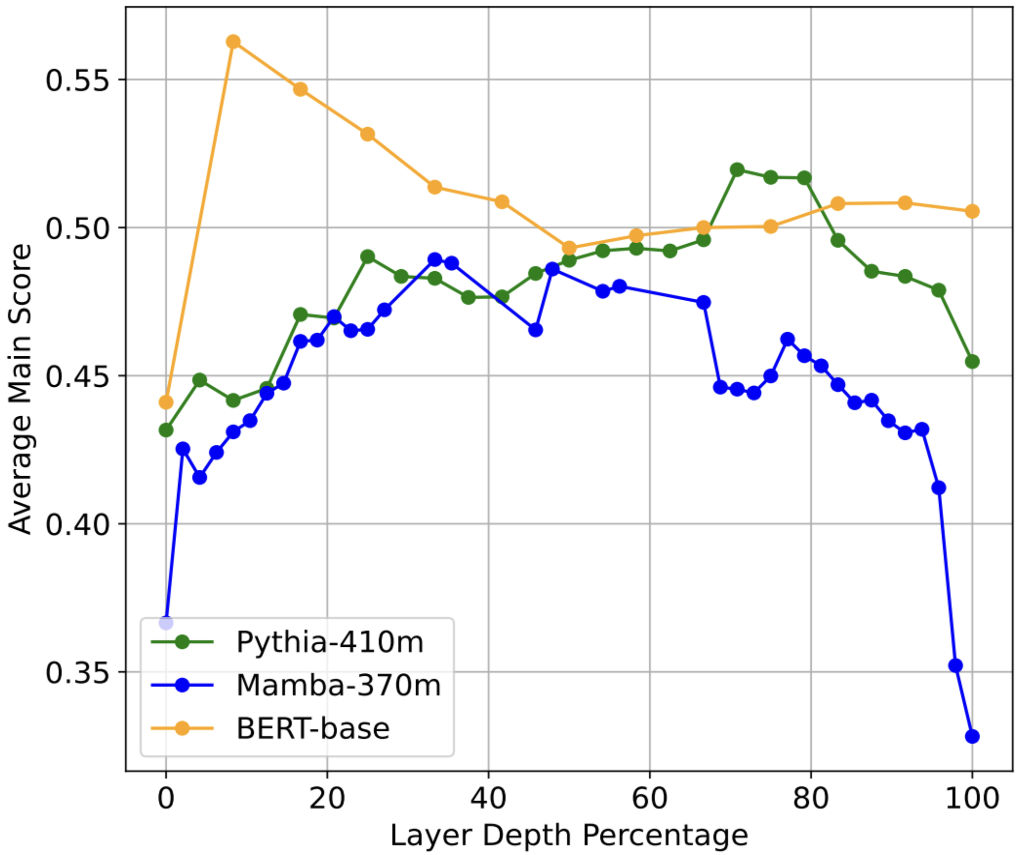
- **Theoretical toolkit** of evaluation metrics to explain internal phenomena and explore *why* intermediate layers are strong

Evaluate Intermediate Layers Performance

- **MTEB Benchmark:**
  - SoTA benchmark (Muennighoff et al., 2022) 🤗
  - Used 32 diverse tasks spanning 5 different domains
  - Probed every model layer

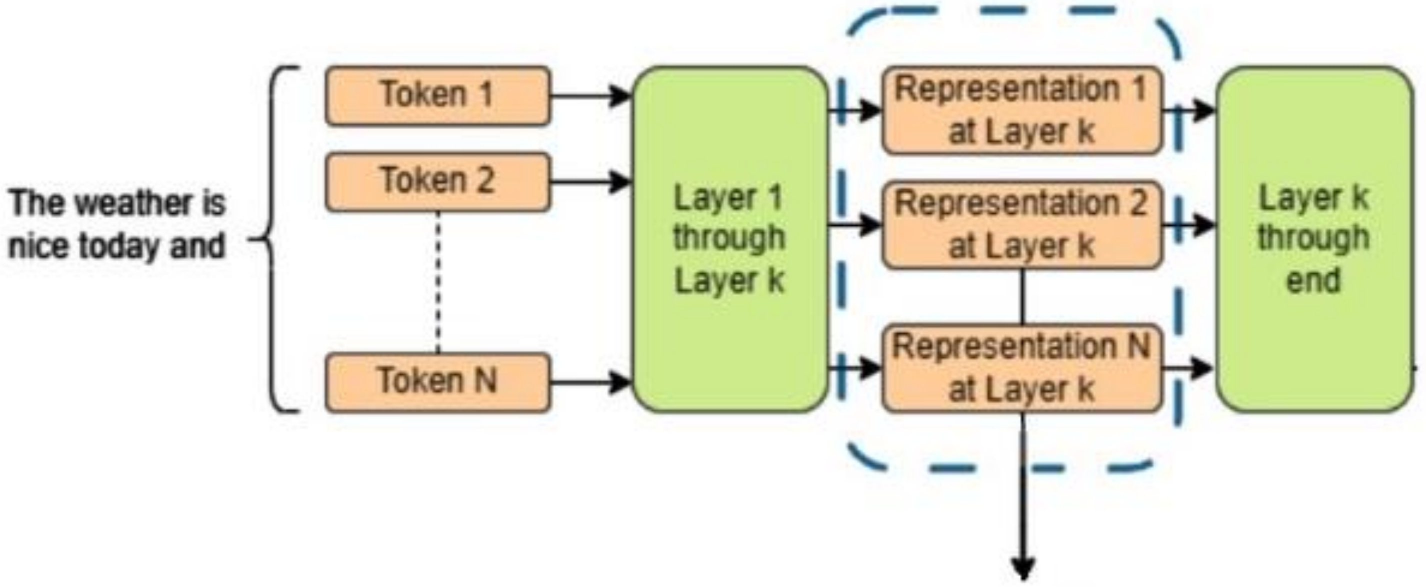
- **Goal:** Find which layers create the best embeddings

Middle Layers Win



Peak performance occurs at intermediate depth, not at the final layer

Our Experimental Pipeline



The weather is nice today and

Token 1

Token 2

Token N

Layer 1 through Layer k

Representation 1 at Layer k

Representation 2 at Layer k

Representation N at Layer k

Layer k through end

Choose evaluation metric

Information-Theoretic

- Prompt Entropy
- Dataset Entropy

Augmentation Invariance

- DiME
- LiDAR
- infoNCE

Geometric

- Curvature

The Metrics Zoo: Three Ways to Evaluate Hidden Representations

Information-Theoretic

How much data is preserved?

Augmentation Invariance

How stable are the representations?

Curvature

What is the shape of the data?

The Metrics Zoo: Three Ways to Evaluate Hidden Representations

Information-Theoretic

How much data is preserved?

Augmentation Invariance

How stable are the representations?

Curvature

What is the shape of the data?

Prompt Entropy, an information-theoretic metric, is a central link

Prompt Entropy

Captures the compression level of representations

For any layer, measure the “effective rank” of the DxD token covariance matrix Σ

$$R(\Sigma) = -\sum_{i=1}^{\min(N,D)} \lambda_i \log \lambda_i$$

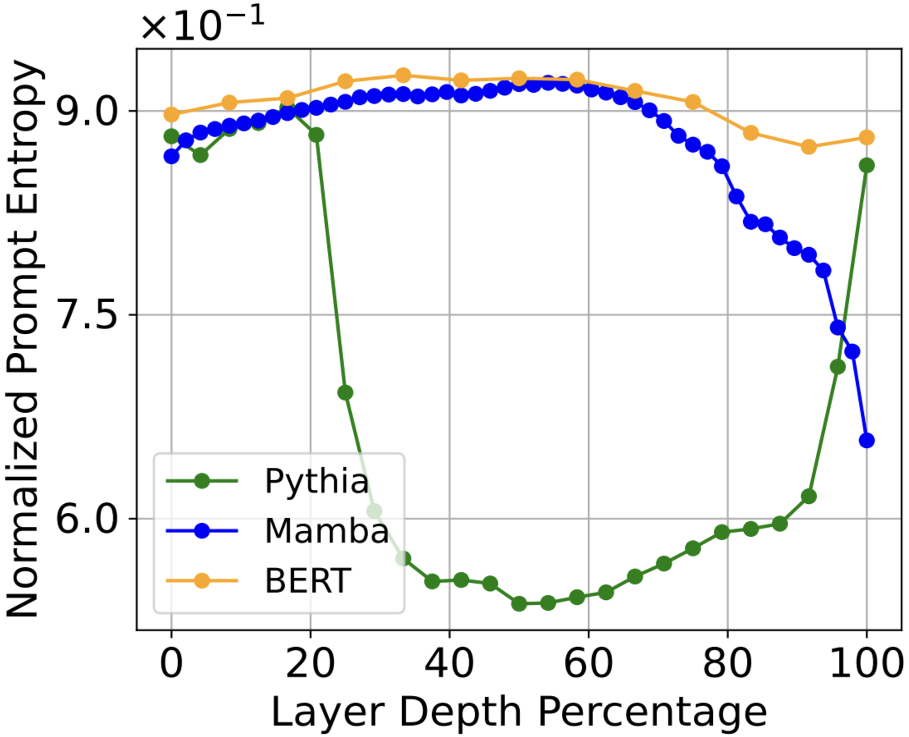
High entropy

- high rank
- tokens are very spread out
- a lot of information

Low entropy

- tokens are very compressed

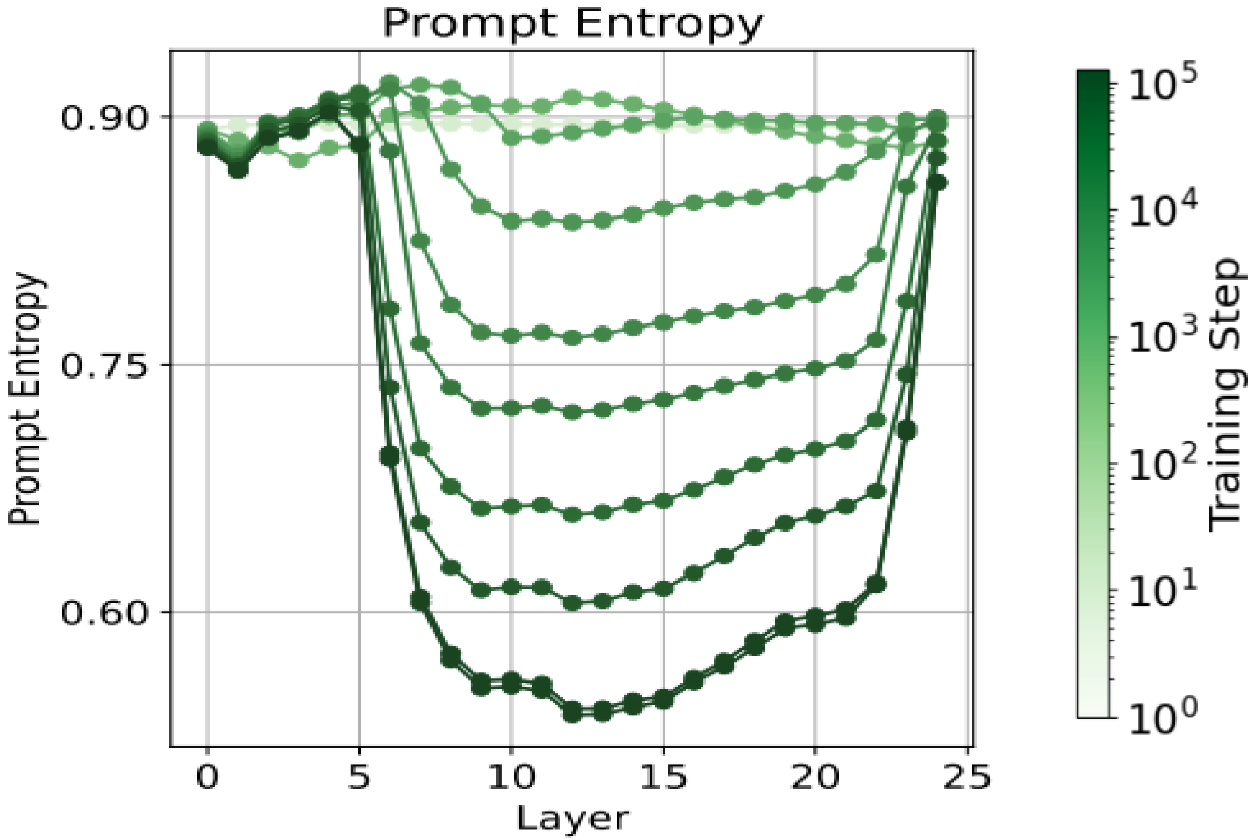
Layerwise Prompt Entropy across Architectures



Model architecture and pretext task influence internal behavior

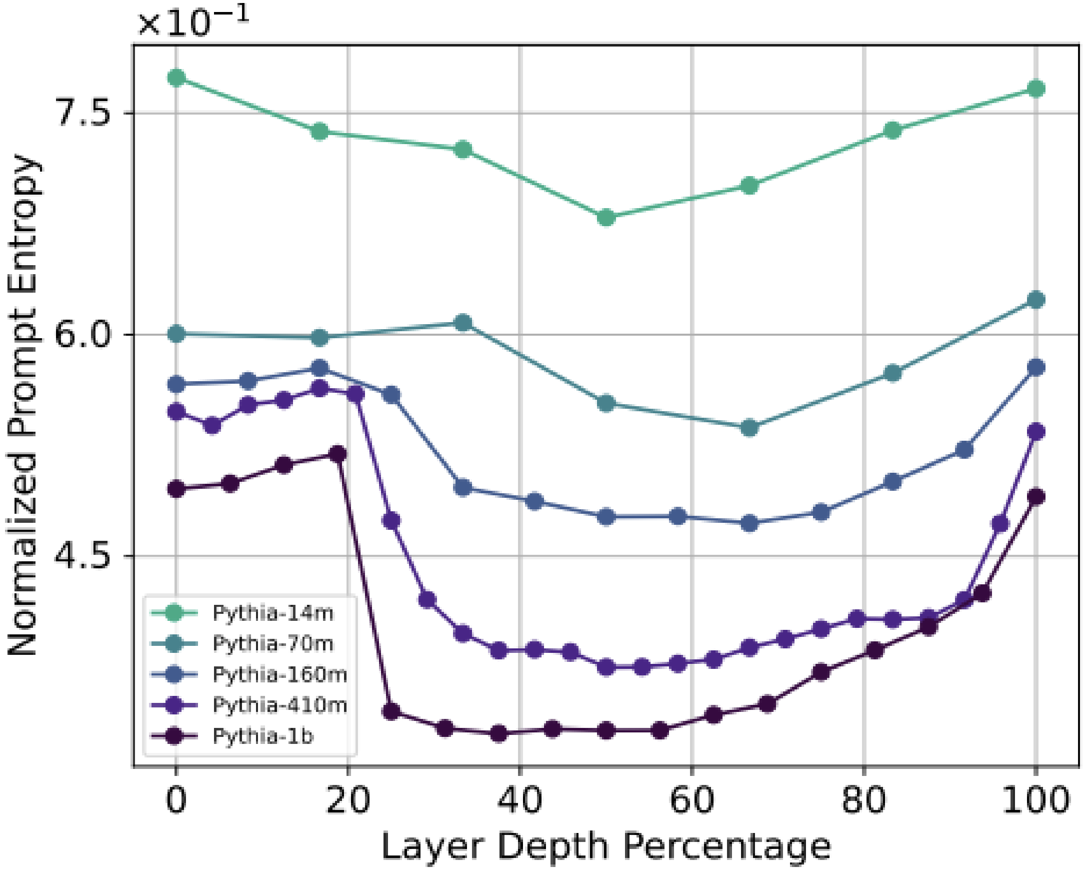
Autoregressive models exhibit a strong intermediate bottleneck

The Bottleneck Emerges During Training



Models learn to compress information as training progresses

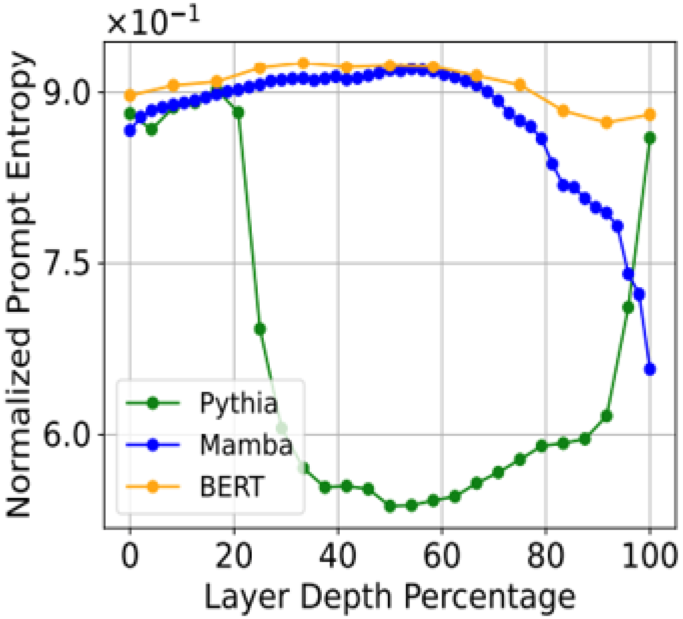
Bigger Models = Stronger Compression



Larger models create deeper bottlenecks

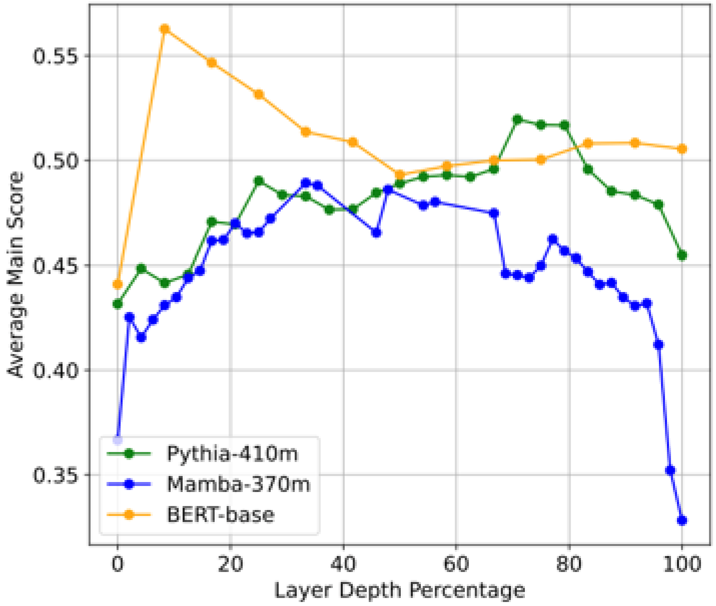
Low Entropy = High Performance in Autoregressive Transformers

Entropy (unsupervised)



Correlations in autoregressive transformer models

Downstream Task Performance (Supervised)



Free Performance Boost: No Training Required

- **The Problem**: Need better embeddings, but no labeled data

- **The Solution**: Find minimum entropy layer

- **The Result**: 5-10% performance improvement with no additional training

Does This Work Beyond Language Models?

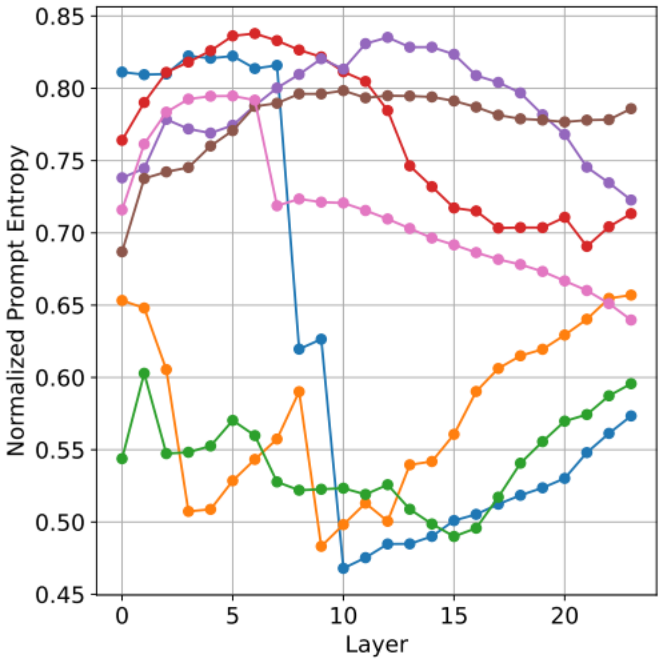
- Vision domain offers a rich selection of models trained on many pretext tasks
    - SimCLR, JEPA, MAE, DINO, supervised models, etc…

- Checked this modality to see:
    - if our findings hold across domains
    - how pretext tasks affect the internal representations

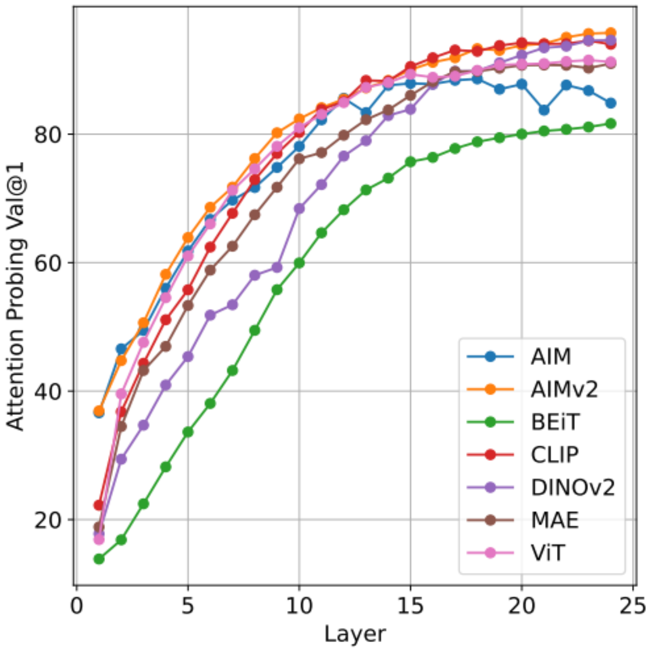
Autoregressive Vision Models Show Same Pattern

- AIMv1 (autoregressive) peaks in middle layers, others don't
- Autoregressive training creates beneficial bottlenecks across modalities

Prompt Entropy



ImageNet Val@1



Key Benefits

- **Performance Boost** - Better embeddings with one line of code
- **Inference Time -** Less layers = less inference time
- **Understanding -** Better understanding of internal model behavior
- **Followup Work -** Seq-VCR (Arefin, et. al 2025) improved GSM8k math reasoning

Take Action: Check Your Middle Layers Today

- Try `model.from_pretrained(output_hidden_states = True)`

- Test layers 20-70% for your tasks

- Measure prompt entropy to find best layers for autoregressive transformers

Thanks for listening!

oscar.skean@uky.edu

Poster today from 11am – 1:30pm at East Exhibition Hall A-B #E-2607

Questions?