The Representation Landscape of Few-Shot Learning and Fine-Tuning in Large Language Models

Diego Doimo, Alessandro Serra, Alessio Ansuini, Alberto Cazzaniga

Area Science Park, Trieste, Italy

NeurIPS 2024

Research Question

Fine-tuning (FT) and in-context learning (ICL) are the central paradigms for solving domain-specific language tasks.

Few-shot learning

1. **Does Not Modify parameters**
2. Sensitivity to prompt format,
3. Order of the shots, choice of the shots

Fine-tuning

1. **Changes parameters**
2. Affected by training training instabilities,
3. Sensitive to the amount of training data

The choice of which is the “best approach” depends on the amount of **data available**, **model size**, …

Research Question

Fine-tuning (FT) and in-context learning (ICL) are the central paradigms for solving domain-specific language tasks.

Few-shot learning

1. **Does Not Modify parameters**
2. Sensitivity to prompt format,
3. Order of the shots, choice of the shots

Fine-tuning

1. **Changes parameters**
2. Affected by training training instabilities,
3. Sensitive to the amount of training data

The choice of which is the "best approach" depends on the amount of **data available**, **model size**, ...

We study how ICL and FT affect the geometry of the representations.

➔ within the same model (*e.g. Llama*)

➔ when they reach the same performance (*MMLU accuracy*)

- How ICL and FT reach similar performance?
- Do they affect the representation landscape the same?

Methods: Advanced Density Peaks Clustering

1. Compute the local density around each data point

k-NN density estimation

$$\rho_i = \frac{k}{NV_{k_i}}$$

k=16 hyper-parameter

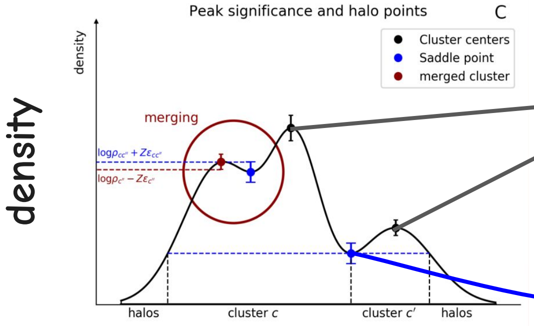
The volume is computed using the intrinsic dimension

intrinsic dimension = 2

F Denti, D Doimo, A Laio, A Mira
The generalized ratios intrinsic dimension estimator
Scientific Reports, 2022

2. Find the density peaks. Keep only “significant” peaks

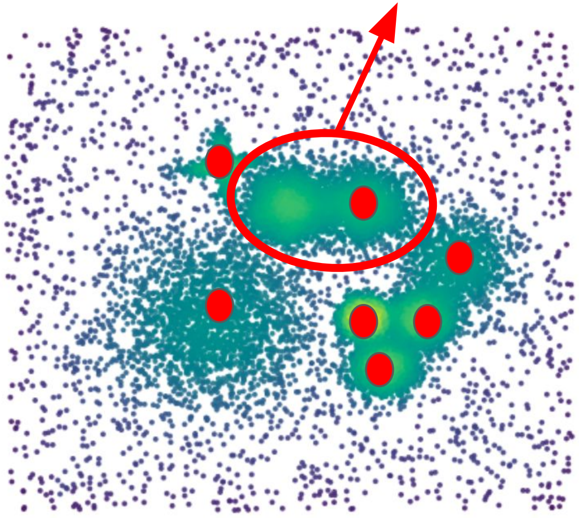
M d’Errico, E Facco, A Laio, A Rodriguez
Automatic topography of high-dimensional data sets by non-parametric density peak clustering,
Information Sciences, *2021*


Peak significance and halo points
C
● Cluster centers
● Saddle point
● merged cluster
density
density
merging
$\log\rho_{cc'} + Z\varepsilon_{cc'}$
$\log\rho_{c'} - Z\varepsilon_{c'}$
halos
cluster c
cluster c'
halos

Significant peaks

Saddle points

merged peaks



Models and Datasets

Pretrained models

Llama-2	7b	13b	70b
Llama-3	8b	70b	
Mistral	7b		

Dataset: MMLU

57 subjects:

abstract algebra, physics, philosophy, medical science, biology economy, ...

200 prompts per subject → 10k samples

Example of two-shot learning setup (MMLU)

"The following are multiple choice questions (with answers) about abstract algebra.

shot 1

Find all c in Z_3 such that Z_3[x]/(x^2 + c) is a field.
A. 0
B. 1
C. 2
D. 3
Answer: B

shot 2

Find the characteristic of the ring 2Z.
A. 0
B. 3
C. 12
D. 30
Answer: A

Question

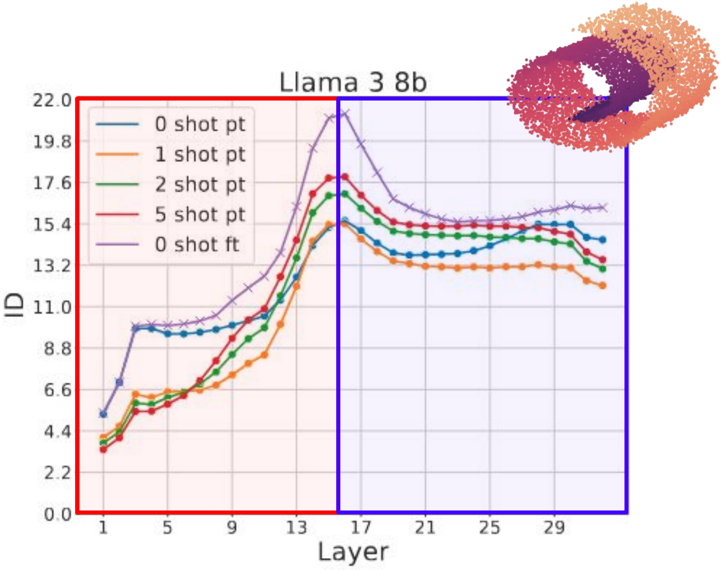
The cyclic subgroup of Z_24 generated by 18 has order
A. 4
B. 8
C. 12
D. 6
Answer: □

How do last token embeddings change in the hidden layers?

The geometry of the probability landscape shows a two-phased behavior

The intrinsic dimension has a peak in the middle of the network

intrinsic dimension

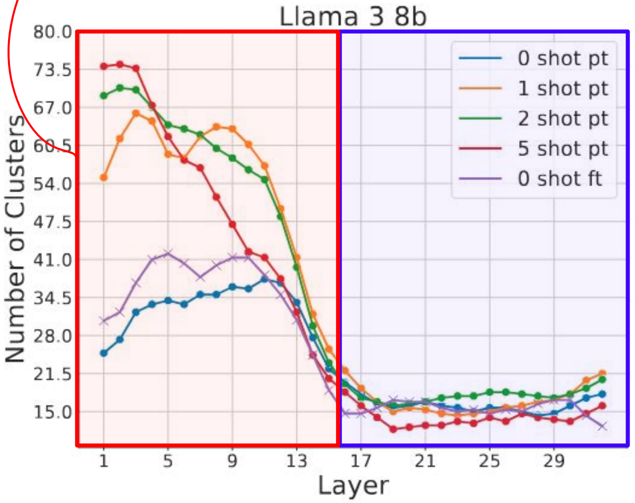


The number of clusters decreases

the number of subjects is 57!

few shot

fine-tuned

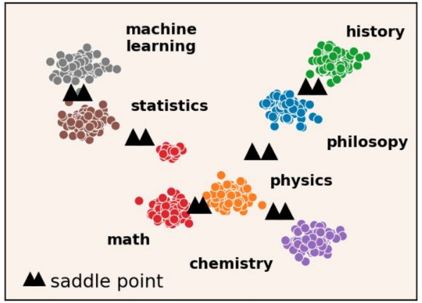


The unsupervised analysis of the geometry of the representation landscape allows to split the networks in two parts

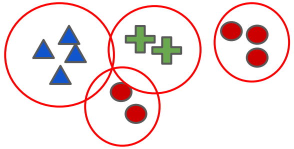
The semantics probability landscape before the transition

Adjusted Rand Index (ARI): measures how well the clusters represent the subjects

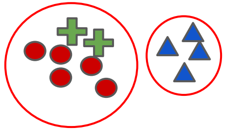
Few-shot learning



homogeneity

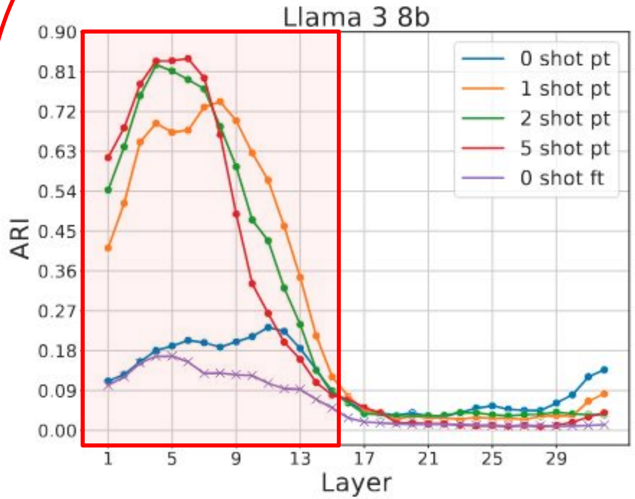


completeness



A high ARI means that the clusters are homogeneous and complete

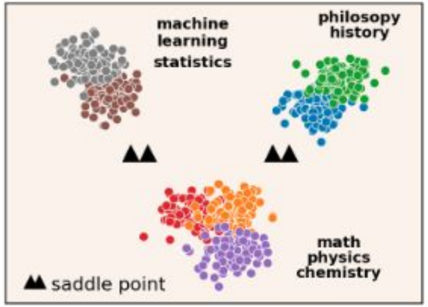
few shot



fine-tuned

Fine-tuned clusters are less homogeneous: the subjects are more mixed

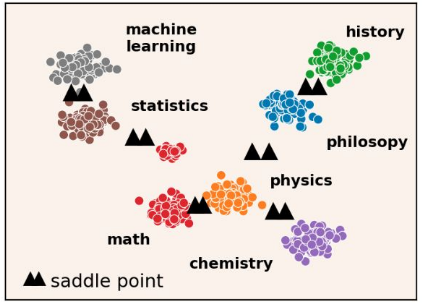
Fine-tuning

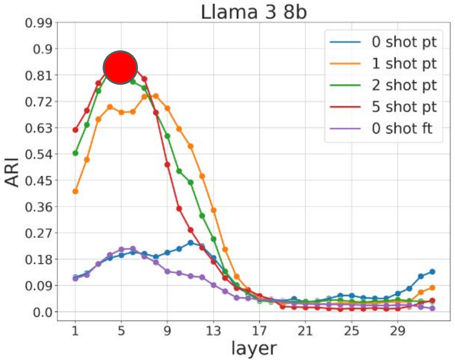


ICL modifies the a lot early layers!

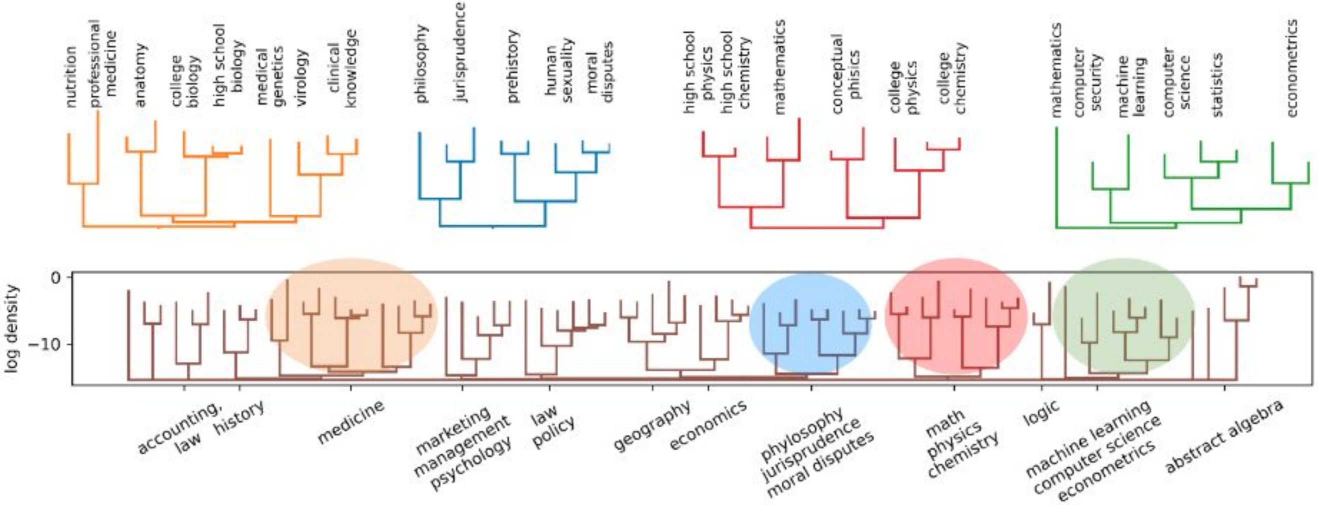
Hierarchical organization of the density peaks in few-shot representations

Few-shot learning





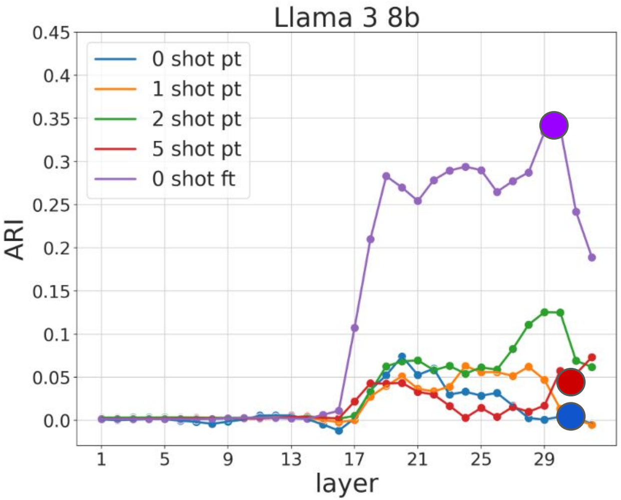
The density of the saddle points between clusters can be used to assess the similarity between clusters



ICL induces a semantically meaningful hierarchical organization of the representations

The probability landscape of late layers

ARI with answers

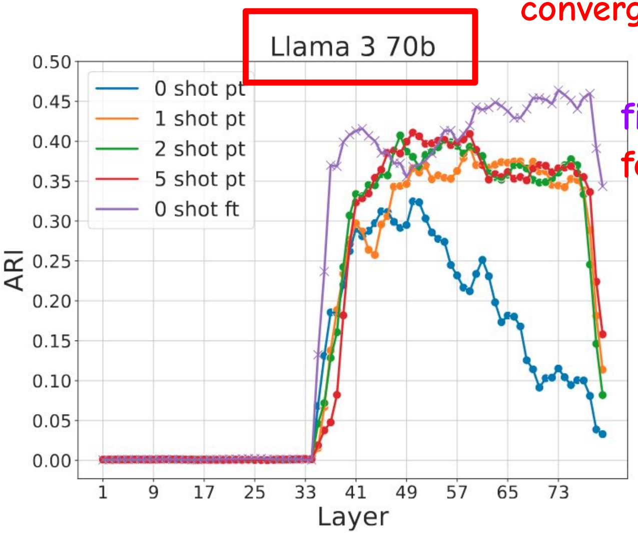


fine-tuned

5 shot

0 shot

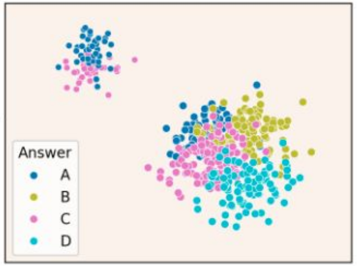
Increasing the model size to 70b the geometry of ICL and SFT converge



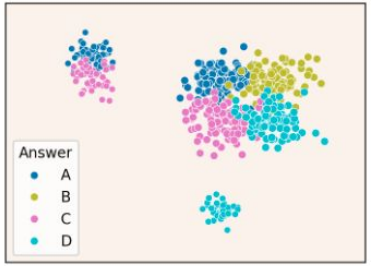
fine tuned

few-shot

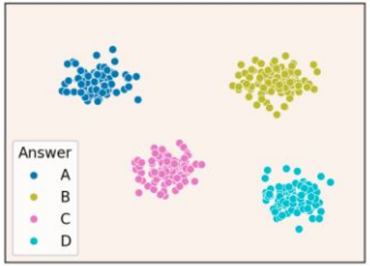
0 shot



5 shot



fine-tuned



The Representation Landscape of Few-Shot Learning and Fine-Tuning in Large Language Models

Diego Doimo, Alessandro Serra, Alessio Ansuini, Alberto Cazzaniga

diego.doimo@areasciencepark.it

AREA SCIENCE PARK

NEURAL INFORMATION PROCESSING SYSTEMS

NeurIPS 2024