

# A Comparative Study of Learning Paradigms in Large Language Models via Intrinsic Dimension

**Saahith Janapati**  
University of Virginia  
jax4zk@virginia.edu

**Yangfeng Ji**  
University of Virginia  
yj3fs@virginia.edu

## Abstract

The performance of Large Language Models (LLMs) on natural language tasks can be improved through both supervised fine-tuning (SFT) and in-context learning (ICL), which operate via distinct mechanisms. SFT updates the model’s weights by minimizing loss on training data, whereas ICL leverages task demonstrations embedded in the prompt, without changing the model’s parameters. This study investigates the effects of these learning paradigms on the hidden representations of LLMs using Intrinsic Dimension (ID). We use ID to estimate the number of degrees of freedom between representations extracted from LLMs as they perform specific natural language tasks. We first explore how the ID of LLM representations evolves during SFT and how it varies due to the number of demonstrations in ICL. We then compare the IDs induced by SFT and ICL and find that ICL consistently induces a higher ID compared to SFT, suggesting that representations generated during ICL reside in higher dimensional manifolds in the embedding space.

## 1 Introduction

Large Language Models (LLMs) have transformed the field of Natural Language Processing through their general natural language understanding capabilities which can be applied to a broad range of tasks. The performance of an LLM on a specific task can be improved through two primary learning paradigms: supervised fine-tuning (SFT) and in-context learning (ICL). SFT adapts pre-trained models to specific tasks by updating their parameters, while ICL requires no parameter updates, relying instead on task-specific demonstrations within the model’s context window. Despite their widespread success, how these methods influence a model’s internal representation space is still not fully understood.

A useful metric for studying the geometric complexity of representations of a model is intrinsic di-

mension (ID). ID quantifies the number of degrees of freedom in the representation space learned by the model, providing a measure of complexity of the manifolds in which embeddings lie in.

In this work, we analyze the intrinsic dimension of hidden representations across model layers during task execution under SFT and ICL. Specifically, we explore:

- How fine-tuning duration influences ID of representations on both training and validation data
- How the number of demonstrations used in ICL affects ID of representations

We find that (1) ID sometimes decreases during the early stages of fine-tuning, but then generally increases throughout the later stages and (2) ICL increases initially as we increase number of demonstrations, but then either plateaus or decreases for higher numbers of demonstrations.

We then conduct experiments directly comparing the intrinsic dimensions of ICL and fine-tuning across several models and datasets. We find that the ID of representations of fine-tuned models are generally lower than the ID of representations of models performing ICL, while the fine-tuned models achieved higher accuracy than models performing ICL. We also find evidence that ID may serve as a practical heuristic to select the optimal number of demonstrations in ICL to maximize performance while minimizing input length. These findings shed light on the differing impacts that the two learning paradigms have on the representation space of LLMs.

## 2 Background

### 2.1 Decoder Transformer Architecture

LLMs are built on the Transformer decoder architecture, which processes token sequences through a series of Transformer layers. Each layer updates

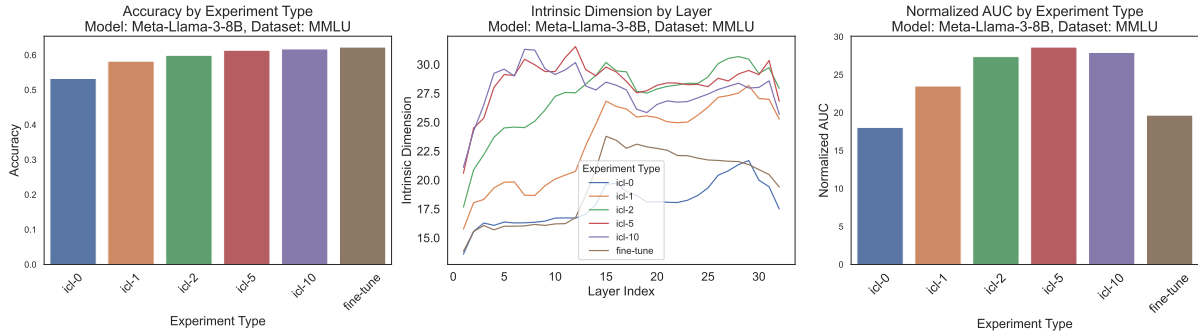


Figure 1: Accuracy, intrinsic dimension, and normalized AUC for the Llama-3-8B model on the MMLU dataset. (a) Fine-tuning achieves the highest accuracy. (b) ICL produces intermediate representations with higher intrinsic dimensions across model layers compared to zero-shot (ic-0) and fine-tuned models. (c) Normalized AUC increases with demonstration counts in ICL, while fine-tuned models show lower AUC.

token representations by attending to preceding tokens from the previous layer, progressively encoding information for the next-token-prediction task. The final layer then uses the representation of the last token to predict the next token in the sequence. In this work, we analyze the intrinsic dimension of representations corresponding to the last token of sequences in which LLMs are prompted to perform a specific natural language task.

## 2.2 Intrinsic Dimension Estimation

Intrinsic dimension (ID) refers to the minimal number of variables required to capture the essential structure of high-dimensional data. Although the representations of modern neural networks lie in high-dimensional spaces (e.g. hidden representations of Llama-3-8B have 4096 dimensions), the representations corresponding to a certain dataset or task often lie on a manifold of far lower dimension as the network disentangles relevant lower-dimensional features to complete a task.

The manifold hypothesis states that this occurs because real-world data often lies on a low-dimensional manifold (Goodfellow, 2016); in order to effectively solve tasks (such as next-token-prediction) on data from these high-dimensional spaces, neural networks must learn representations aligned with the low-dimensional manifold of the data. The intrinsic dimension of data representations can thus provide a unique insight into the complexity of representation spaces constructed across the layers by a neural network.

The method we use to estimate ID is the **TwoNN estimator**, introduced by Facco et al. (2017). We chose this method because of its simplicity, computational efficiency, and robustness in han-

dling datasets with non-uniform densities and high-dimensional curvature, which are common challenges in neural network representations.

The TwoNN method operates on a set of points and outputs an estimate of the ID of the points. The ID is estimated using the distances to the first and second nearest neighbors for each point. For a given point  $x$ , the ratio  $\mu = \frac{r_2}{r_1}$  is calculated, where  $r_1$  and  $r_2$  are the distances to the first and second nearest neighbors, respectively. The intrinsic dimension  $d$  is then derived from the empirical cumulative distribution function (CDF) of  $\mu$ . The log-linear relationship between  $\log(\mu)$  and  $\log(1 - F_{\text{emp}}(\mu))$ , where  $F_{\text{emp}}(\mu)$  is the empirical CDF, is used to estimate  $d$ :

$$d = -\frac{\log(1 - F_{\text{emp}}(\mu))}{\log(\mu)}$$

The TwoNN estimator has been successfully applied in several prior works analyzing the intrinsic dimension of neural network representations, such as Sharma and Kaplan (2022), Ansuini et al. (2019), Valeriani et al. (2024), and specifically in large language models (LLMs) by Cheng et al. (2023) and Lee et al. (2024). We also verify the correlation between the TwoNN and another widely used Intrinsic Dimension Estimator, the Maximum Likelihood Estimator introduced by Levina and Bickel (2004), in Appendix F as a sanity check.

## 3 Related Works

### 3.1 Supervised Fine-Tuning in LLMs

Pre-trained LLMs can quickly be adapted to improved performance on natural language tasks through supervised fine-tuning by performing gradient updates on training examples of a task.

Aghajanyan et al. (2020) shows that fine-tuning large language models requires updating only a low-dimensional subspace of parameters to achieve near-optimal performance. (Note that this work is focused on the intrinsic dimension of parameter space, while our work analyzes the intrinsic dimension of the representation space). Building on this work, Hu et al. (2021) introduces Low-Rank Adaptation (LoRA), a method to inject low-rank matrices to weight matrices instead of performing fine-tuning across all the parameters of a model, which is often intractable for LLMs. We utilize LoRA to do all fine-tuning in our experiments.

### 3.2 In-Context Learning

Introduced in GPT-3 by Brown (2020), ICL (or few-shot learning) refers to the ability of LLMs to learn to perform a task in a single forward pass based on (input, output) pairs of that task embedded in a prompt.

Dai et al. (2022) provides evidence that ICL operates as implicit meta-optimization, where GPT models perform a gradient-like update using attention mechanisms during its forward pass. This suggests that ICL is replicating fine-tuning behavior. Specifically, they demonstrate attention outputs and weights are updated in a direction similar to fine-tuning during ICL.

Xie et al. (2021) explain in-context learning as implicit Bayesian inference, where large language models infer latent document-level concepts during pretraining. These inferred concepts are leveraged at test time to solve tasks based on input-output examples embedded in prompts.

Expanding on the ICL paradigm to long-context models, Agarwal et al. (2024) studied many-shot ICL, where hundreds or thousands of task examples are used to improve performance of frontier models; the work finds that more demonstrations generally improves model performance on a variety of complex tasks such as mathematical problem-solving.

### 3.3 Intrinsic Dimension in Deep Learning

Ansuini et al. (2019) investigated the intrinsic dimensionality (ID) of data representations across various convolutional neural networks (CNNs) for image classification. They observed a consistent 'hunchback' pattern in ID evolution: an initial increase in the early layers followed by a progressive decrease in later layers.

Valeriani et al. (2024) extended the analysis of intrinsic dimension to protein language models and image transformers, finding that the evolution of representations across layers is marked by distinct phases of ID growth and compression.

Yin et al. (2024) investigate the use of Local Intrinsic Dimension (LID) to detect untruthful outputs from LLMs. Their study reveals that truthful outputs typically exhibit lower LIDs compared to hallucinated, suggesting that LID can be used as a signal for truthfulness in LLM generations. Yin et al. (2024) also identifies a positive relationship between ID of data representations and validation performance during fine-tuning.

Cheng et al. (2023) show that intrinsic dimension correlates with fine-tuning ease and perplexity, with low-dimensional representations enabling faster task adaptation. They also find that ID values are consistent across model sizes, supporting the manifold hypothesis and suggesting LLMs trained on similar data recover comparable intrinsic dimensions.

Of particular relevance to our study is the concurrent work of Doimo et al. (2024), which analyzes the internal representations of LLMs when solving tasks from the MMLU dataset through ICL and SFT. They find that ICL forms semantic clusters in early layers, while SFT sharpens clusters in later layers for task-specific answers. They also identify that ID increases when using higher number of demonstrations in ICL. Furthermore, they find that SFT induces a higher ID compared to ICL (a finding which our results consistently contradict).

To our knowledge, our work is the first to systematically analyze and compare Intrinsic Dimension across the two learning paradigms for numerous datasets and models. We also provide in-depth analyses of how ID is affected by different factors of the two learning paradigms (such as number of gradient steps in SFT and number of demonstrations in ICL).

### 3.4 Intrinsic Dimension and Neural Network Scaling Laws

Recent studies emphasize the role of intrinsic dimension in neural network scaling behavior. (Sharma and Kaplan, 2022) showed that test loss follows a power-law with model size, where the scaling exponent is inversely related to the intrinsic dimension of the data manifold. They extended this to model representations in hidden layers, demon-

strating that representation complexity and data manifold structure jointly influence model scaling and generalization.

## 4 Methods

We perform experiments with subsets of the following 8 datasets: AG News (Zhang et al., 2015), CoLA (Warstadt et al., 2018), CommonsenseQA (Talmor et al., 2018), MMLU (Hendrycks et al., 2020), MultiNLI (Williams et al., 2017), QNLI (Wang, 2018), QQP (Wang et al., 2017), SST2 (Socher et al., 2013).

For each experiment, we also use subsets of the following open-source LLMs: Llama-3-8B (Dubey et al., 2024), Llama-2-13b, Llama-2-7b (Touvron et al., 2023), and Mistral-7B-v0.3 (Jiang et al., 2023). We run experiments using 6 NVIDIA A600s.

For each dataset, we create a training set of 1000 examples and a validation set of 5000 validation examples. We use 5000 validation elements to ensure stability of the TwoNN estimator. Details regarding dataset creation can be found in Appendix G.

We calculate accuracy a model of a model’s response using the logit probabilities assigned to the tokens corresponding to the possible answers for that question. We mark a response as correct if the probability corresponding to first token of the correct label is highest.

### 4.1 Computing Intrinsic Dimension

In both the SFT and ICL learning paradigms, the model is provided an input sequence of tokens and must generate an output sequence containing its answer to the posed task. To measure the ID of a model’s representations for a certain dataset, we collect the activations corresponding to the **last token of every input sequence** in the dataset at every layer of the LLM. For a model with  $L$  layers and a dataset with  $N$  elements, we collect  $L$  groups of  $N$  hidden state representations. We then compute the intrinsic dimension of each of the  $L$  sets of  $N$  vectors, which yields an estimate of the ID of the representation space at every layer of the model. We refer to the line created when plotting Layer Index vs ID as the **Intrinsic Dimension Curve**.

To obtain a single aggregated metric representing the intrinsic dimension (ID) across all the layers of a model, we compute the **Normalized Area Under the Curve (AUC)** of the Intrinsic Dimension Curve, which is defined as:

$$\text{Normalized AUC} = \frac{1}{L} \sum_{i=1}^{L-1} \frac{1}{2} (\text{ID}_i + \text{ID}_{i+1})$$

Here,  $\text{ID}_i$  represents the intrinsic dimension estimate at layer  $i$ . This formula uses the trapezoidal rule to compute the area under the Intrinsic Dimension Curve. We normalize the summation to allow comparison between models with different numbers of layers.

## 5 Dynamics of ID during Supervised Fine-Tuning

### 5.1 Supervised Fine-Tuning Experimental Setup

To investigate the impact of supervised fine-tuning at a granular level, we conduct experiments using the 8 datasets discussed in Section 4 and the Llama-3-8B and Llama-2-13B models.

Using the training split for each of the datasets, we perform LoRA fine-tuning on the query, key, value, and output projection matrices of attention heads across all layers of the model. For all models, we fine-tune with a batch size of 16 for 15 epochs. For all fine-tuning runs, we use LoRA hyperparameters of  $r = 64$ ,  $\text{lora\_alpha} = 16$ ,  $\text{lora\_dropout} = 0.1$ , no LoRA bias, and a learning rate of  $1e^{-4}$ .

During the fine-tuning process for a specific model and dataset, we save a checkpoint of the model after every epoch (~62 gradient update steps). For each checkpoint, we evaluate the model’s accuracy and measure the intrinsic dimension (ID) of the hidden representations on prompts from the training and validation splits for the dataset.

### 5.2 Intrinsic Dimension Generally Increases Through Fine-Tuning

As demonstrated in Figure 2c, we find that ID of representations corresponding to both training data and validation data sometimes decreases during the initial stages of fine-tuning, but then generally increases as fine-tuning progresses.

We also observe larger changes in ID values for later layers of the models, despite LoRA adaptation being applied on all the layers with the same configuration (Figure 2a).

We also find that that the AUC values of the model on the training set and validation set are often highly correlated with each other during the

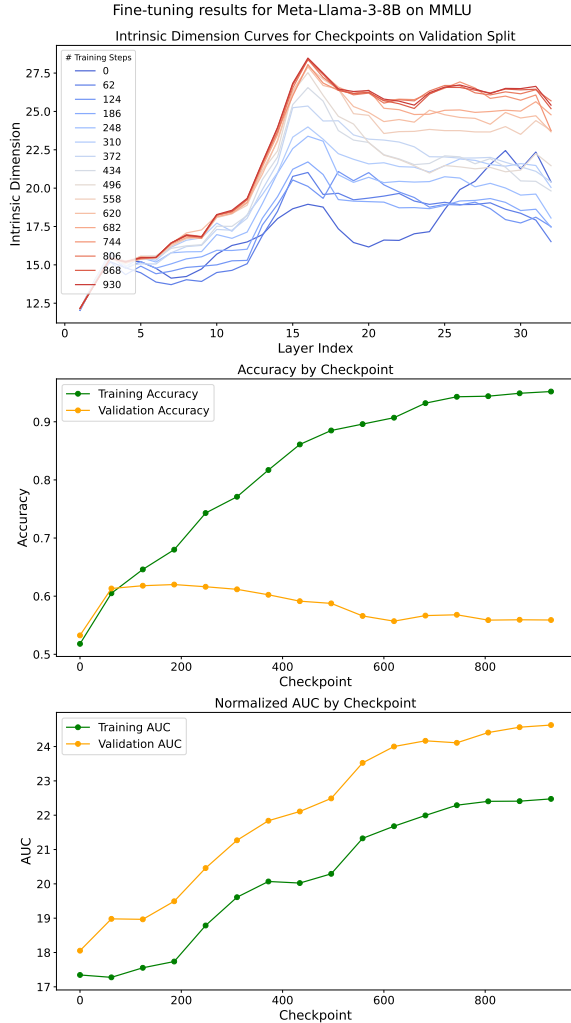


Figure 2: Fine-tuning results for Meta-Llama-3-8B on the MMLU dataset. (a) Intrinsic Dimension curves on the validation split increase across training steps. (b) Training accuracy improves steadily, while validation accuracy plateaus. (c) Normalized AUC for training and validation sets increases throughout fine-tuning.

training process (Figure 2c). Experimental results for all models and datasets can be found in Appendix B.

Prior work by Yin et al. (2024) found that on Question-Answering datasets, intrinsic dimension of representations is correlated with validation performance and can therefore be used as a heuristic to select final checkpoints. In general, we do not find this trend to hold on the datasets and models we tested. In fact, as shown in Figure 2, large increases in validation accuracy sometimes coincide with drops in ID on both the training and validation datasets.

## 6 Relationship of ID in ICL with Different Numbers of Shots

### 6.1 In-Context Learning Experimental Setup

To investigate the impact of ICL on the ID of model representations, we conduct experiments using the Llama-3-8B and Llama-2-13B models. The datasets included in our evaluation are CommonsenseQA, MMLU, and QNLI.

We evaluate ICL performance using various values of  $k$ , where  $k$  denotes the number of demonstrations in the ICL prompt. The values considered are  $k \in \{0, 1, 2, 5, 10, 12, 14, 16, 18, 20\}$ . Note that  $k = 0$  serves as a baseline, representing the model’s performance in the absence of both ICL and SFT.

For each  $k$  and dataset, we generate 5000 ICL prompts (one for each element of the validation split of the dataset). Each ICL prompt includes  $k$  unique demonstrations, or (input, output) pairs, randomly sampled from the training set. While we ensure that demonstrations within a single prompt are unique, they may be reused across different prompts.

### 6.2 ID Has a Non-Linear Relationship with Number of Demonstrations

We observe that ID values across layers can fluctuate up till a certain value of  $k$  (usually up to 5 or 10), and then either plateau or steadily decreases for larger values of  $k$  (Figure 3c). Results for all model and dataset configurations can be found in Appendix A. This is agreement with prior work by Doimo et al. (2024), which found that ICL increased as  $k$  was varied from 0,1,2, and 5.

We also find that across most (model, dataset) combinations, the shapes of the intrinsic dimension curves correlate strongly with each other for  $k \geq 2$ .

Due to our procedure of selecting demonstrations with replacement, we suspected that the plateau in ID for larger values of  $k$  may be due to greater number of demonstrations shared across prompts. We hypothesized that shared demonstrations could make representations corresponding these prompts artificially similar to each other and thus skew ID results. To test this, we perform additional experiments using larger number of dataset elements from the CommonsenseQA, QNLI, and AG News datasets, which contain enough training elements to ensure that demonstrations are not used in prompts for more than one element of the validation set. We observe the same trend of an increase,

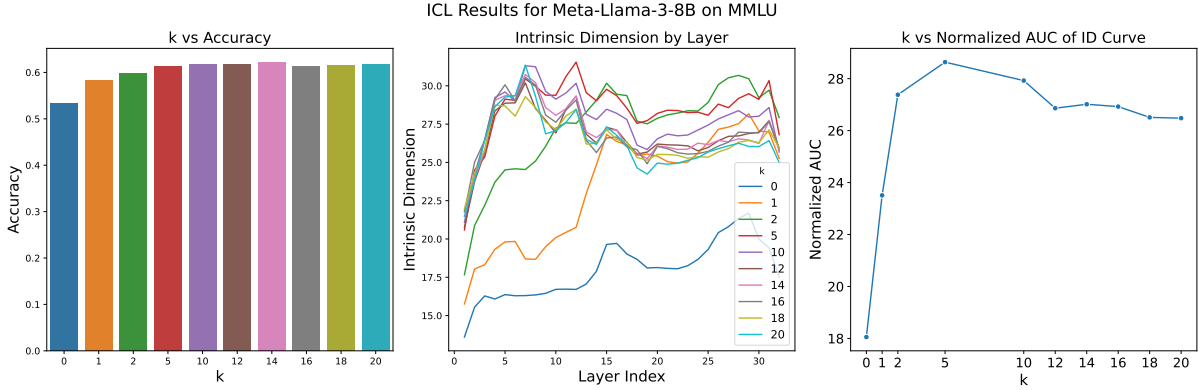


Figure 3: (ICL) results for Meta-Llama-3-8B model on MMLU dataset. (a) Accuracy increases, then plateaus as number of demonstrations increases (b) Intrinsic Dimension (ID) curves for different values of  $k$ . (c) Normalized AUC of the ID curves peaks at  $k=5$ , which also aligns with saturation of accuracy.

and then general plateau in the ID for these experiments as well, suggesting that the plateau is likely not due to the reuse of demonstrations among the prompts. Full results for this experiment can be found in Appendix ??.

We also find that peaks in the  $k$  vs AUC relationship align with peak (or near-peak) accuracy in 5 out of the 6 ICL experiments we conducted. Thus, the  $k$  value corresponding to the peak ID may serve as a practical indicator of the optimal number of demonstrations to use for ICL to maximize performance while minimizing input length.

A hypothesis as to why ID plateaus or slightly decreases as  $k$  increases is that more demonstrations allows the model to more effectively capture the underlying task put forth in the demonstrations, allowing for representations corresponding to different inputs to be more similar to each other. This idea is supported by previous theoretical analysis of ICL by Xie et al. (2021), which posits that a greater number of demonstrations allows for a model to more effectively infer the latent concept across demonstrations.

We also find that across most experiments, accuracy either steadily increases or plateaus with higher numbers of demonstrations (Figure 3a).

## 7 Comparing Intrinsic Dimension of In-Context Learning and Supervised Fine-Tuning

### 7.1 Experiment Setup for Comparative Analysis

We perform a series of experiments directly comparing ID curves from both SFT and ICL. We utilize similar setups as discussed in Sections 5 and

6. However, for the fine-tuning experiments in this section, we only train for 4 epochs and only measure the accuracy and ID for the final checkpoint. We choose this number of epochs because the previous experiments in Section 5 indicated that models tended to begin over-fitting past 4 epochs across the datasets tested.

For ICL experiments in this section, we only consider 0,1,2,5,10 as values for  $k$ . We choose these values for the number of demonstrations because they are popular choices for ICL in practice and also because prior experiments in 6 indicate that ID curves tended to plateau when  $k \geq 10$ . We perform experiments on all 8 datasets and 4 models discussed in Section 4.

### 7.2 In-context Learning Induces Higher IDs Compared to Fine-Tuning

We find that across all datasets and models, ICL with  $k \geq 5$  consistently induces higher intrinsic dimensions (IDs) across all layers compared to SFT and 0-shot prompts (see Figures 1b and 1c). This contrasts with the findings of Doimo et al. (2024), who report that ID induces higher ID in later layers.

We also find that the ID values of models fine-tuned with 1000 samples tend to remain similar to the original ID of the baseline model on a zero-shot prompt (designated by icl-0). We present a heatmap displaying average differences in normalized AUC between learning paradigms in Figure 4, and a boxplot depicting the distribution of normalized AUC values for the different paradigms in Figure 5.

Dataset	ICL-0	ICL-1	ICL-2	ICL-5	ICL-10	Finetune 1K
SST-2	0.685	0.633	0.731	0.807	0.832	<b>0.944</b>
CoLA	0.720	0.723	0.735	0.746	0.742	<b>0.750</b>
QNLI	0.517	0.513	0.555	0.590	0.585	<b>0.761</b>
QQP	0.417	0.462	0.485	0.508	0.519	<b>0.707</b>
MNLI	0.374	0.367	0.387	0.414	0.431	<b>0.676</b>
AGNews	0.638	0.573	0.712	0.772	0.809	<b>0.881</b>
CommonsenseQA	0.199	0.375	0.417	0.470	0.492	<b>0.500</b>
MMLU	0.449	0.488	0.511	0.524	0.531	<b>0.542</b>

Table 1: Average accuracy results for Datasets across ICL and SFT settings. SFT obtains the highest average accuracy for all datasets. Accuracy increases and then plateaus for higher number of demonstrations.

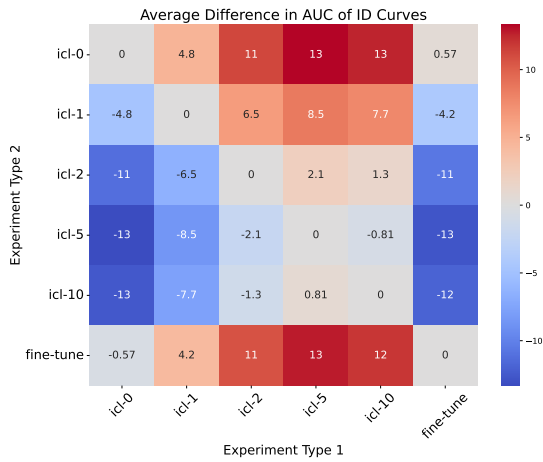


Figure 4: Heatmap showing the average differences in normalized AUC of ID curves between pairs of learning paradigms. Each value represents the average difference (Experiment Type 1 - Experiment Type 2), computed across all (model, dataset) pairs.

### 7.3 Analysis of Intrinsic Dimension Curves

#### 7.3.1 Differing Shapes of Intrinsic Dimension Curves

We observe that the exact shape of the Intrinsic Dimension curves is highly dependent upon the dataset. For some datasets, such as AG News, we observe a consistent 'hunchback' shape, where ID increases and then progressively lowers throughout the later layers of the model, across all models and learning paradigms (Figure 36). This shape has been reported on by previous work by Yin et al. (2024) in QA datasets. However, we find that this pattern does not consistently hold across all models, datasets, and learning paradigms. For example, on the QQP dataset, we do not observe a consistent hunchback shape for icl-0, icl-1 or fine-tune learning paradigms (33). Thus, in contrast to Convolutional Neural Networks (Ansuini et al.,

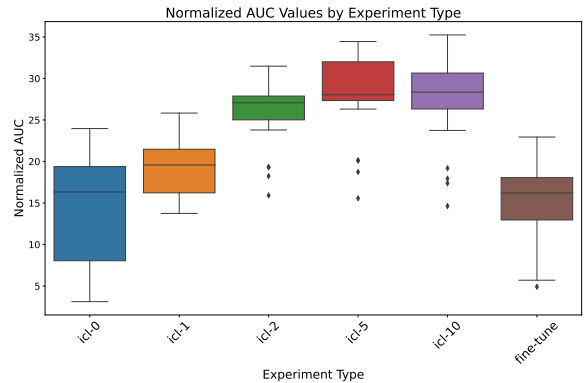


Figure 5: Boxplot displaying the distribution of normalized AUC values for different learning paradigms. Each point corresponds to the normalized AUC value for a (model, dataset) pair. Median normalized AUC peaks with 5-shot ICL, while values for SFT are closer to zero-shot baseline (icl-0).

2019), Image Generation Transformers such as ImageGPT and Protein Language Models (Valeriani et al., 2024), which have been found to exhibit consistent patterns Intrinsic Dimension patterns across their layers for inputs of their respective data modality. This difference suggests that LLMs encode data into more diverse manifolds in their representation space, potentially reflecting their generality and the complexity of their learning task compared to other neural networks.

We also find that for a specific learning paradigm, the range of normalized AUC values across datasets for the four different models we tested to be in similar ranges, despite the fact that the models come from different families and have different embedding dimensions (e.g. Llama-2-13b has a hidden dimension of 5120, while the other three models have hidden dimensions of 4096). Figure 6 depicts the range of normalized AUC values for the icl-5 learning paradigm and shows that all the values fall

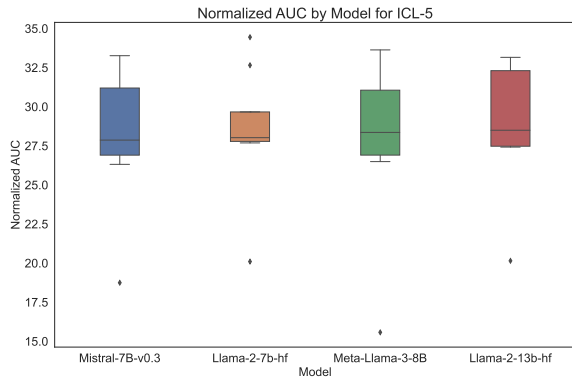


Figure 6: Boxplot displaying the distribution of normalized AUC values across datasets for each model in the ICL-5 shot setting. Each point corresponds to the value corresponding to a (model, dataset) pair. ID values all lie in a narrow range, highlighting similarity in representation spaces across models.

within a range of 20. We view this as evidence that different models may be generating representations with similar geometric complexity for a specific dataset, despite differences in model size or pre-training schemes. We include similar boxplots for normalized AUC values of other experiments in Appendix C. These findings are in agreement with results from Cheng et al. (2023), which find that LLMs of different sizes and families create representations with similar ID values for a variety of text corpora.

#### 7.4 Comparing Performance of Different Learning Paradigms

We found that models fine-tuned with 1k samples obtained the highest accuracy, and models performing ICL with 10-samples followed closely. This implies ID may not be closely related to accuracy, as the models fine-tuned with 1k samples had a closer ID to the base models than to ICL models, which had substantially lower accuracies. See Table 1 for average performance of each of the learning paradigms across the models and datasets tested.

### 8 Summary

We present a detailed analysis of the intrinsic dimension (ID) induced by the SFT and ICL learning paradigms. We identify a trend where the normalized AUC of ID curves sometimes decreases during the initial stages of SFT but generally increases during the later stages of fine-tuning.

Additionally, we observe a trend in the number of demonstrations used in ICL and the resulting ID

curves: the normalized AUC of the ID curves initially increases for small values of  $k$  (the number of demonstrations) but plateaus or slightly decreases as  $k$  increases further. We observe that the value of  $k$  which corresponds to the highest normalized AUC also obtains peak (or close-to-peak) accuracy, suggesting that ID may serve as a useful indicator to select the number of demonstrations to use during ICL.

Finally, we compare the ID curves from ICL and SFT directly and find that representations generated during ICL yield higher ID curves compared to fine-tuning on 1k samples, although SFT with 1k samples achieves the highest performance. This analysis provides evidence that the two learning paradigms induce distinct representational structures in the embedding space, with representations created by ICL occupying higher-dimensional manifolds.

### 9 Limitations

In this study, we limit our analysis to models with sizes between 7B and 13B parameters. Future work may extend this line of investigation to models of larger sizes. We also focus on datasets defined by narrowly-defined tasks and do not consider datasets with long-form answers. Due to compute requirements, we only perform fine-tuning using LoRA adapters and do not study the impacts of full fine-tuning on intrinsic dimension. We also only analyze dense transformer models and do not study architectures such as Mixture-of-Experts (Shazeer et al., 2017) used in current frontier models.

### References

- Rishabh Agarwal, Avi Singh, Lei M Zhang, Bernd Bohnet, Stephanie Chan, Ankesh Anand, Zaheer Abbas, Azade Nova, John D Co-Reyes, Eric Chu, et al. 2024. Many-shot in-context learning. *arXiv preprint arXiv:2404.11018*.
- Armen Aghajanyan, Luke Zettlemoyer, and Sonal Gupta. 2020. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. *arXiv preprint arXiv:2012.13255*.
- Alessio Ansuini, Alessandro Laio, Jakob H Macke, and Davide Zoccolan. 2019. Intrinsic dimension of data representations in deep neural networks. *Advances in Neural Information Processing Systems*, 32.
- Stephen Bach, Victor Sanh, Zheng Xin Yong, Albert Webson, Colin Raffel, Nihal V. Nayak, Abheesht Sharma, Taewoon Kim, M Saiful Bari, Thibault



- Fevry, Zaid Alyafeai, Manan Dey, Andrea Santilli, Zhiqing Sun, Srulik Ben-david, Canwen Xu, Gunjan Chhablani, Han Wang, Jason Fries, Maged Al-shaibani, Shanya Sharma, Urmish Thakker, Khalid Almubarak, Xiangru Tang, Dragomir Radev, Mike Tian-jian Jiang, and Alexander Rush. 2022. [Prompt-Source: An integrated development environment and repository for natural language prompts](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 93–104, Dublin, Ireland. Association for Computational Linguistics.
- Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.
- Emily Cheng, Corentin Kervadec, and Marco Baroni. 2023. Bridging information-theoretic and geometric compression in language models. *arXiv preprint arXiv:2310.13620*.
- Damai Dai, Yutao Sun, Li Dong, Yaru Hao, Shuming Ma, Zhifang Sui, and Furu Wei. 2022. Why can gpt learn in-context? language models implicitly perform gradient descent as meta-optimizers. *arXiv preprint arXiv:2212.10559*.
- Diego Doimo, Alessandro Serra, Alessio Ansuini, and Alberto Cazzaniga. 2024. The representation landscape of few-shot learning and fine-tuning in large language models. *arXiv preprint arXiv:2409.03662*.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Elena Facco, Maria d’Errico, Alex Rodriguez, and Alessandro Laio. 2017. Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific reports*, 7(1):12140.
- Ian Goodfellow. 2016. Deep learning.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.
- Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.
- Jin Hwa Lee, Thomas Jiralerspong, Lei Yu, Yoshua Bengio, and Emily Cheng. 2024. Geometric signatures of compositionality across a language model’s lifetime. *arXiv preprint arXiv:2410.01444*.
- Elizaveta Levina and Peter Bickel. 2004. Maximum likelihood estimation of intrinsic dimension. *Advances in neural information processing systems*, 17.
- Utkarsh Sharma and Jared Kaplan. 2022. Scaling laws from the data manifold dimension. *Journal of Machine Learning Research*, 23(9):1–34.
- Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*.
- Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.
- Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2018. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *arXiv preprint arXiv:1811.00937*.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- Lucrezia Valeriani, Diego Doimo, Francesca Cuturello, Alessandro Laio, Alessio Ansuini, and Alberto Cazzaniga. 2024. The geometry of hidden representations of large transformer models. *Advances in Neural Information Processing Systems*, 36.
- Alex Wang. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.
- Zhiguo Wang, Wael Hamza, and Radu Florian. 2017. Bilateral multi-perspective matching for natural language sentences. *arXiv preprint arXiv:1702.03814*.
- Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. 2018. Neural network acceptability judgments. *arXiv preprint arXiv:1805.12471*.
- Adina Williams, Nikita Nangia, and Samuel R Bowman. 2017. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*.
- Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. 2021. An explanation of in-context learning as implicit bayesian inference. *arXiv preprint arXiv:2111.02080*.
- Fan Yin, Jayanth Srinivasa, and Kai-Wei Chang. 2024. Characterizing truthfulness in large language model generations with local intrinsic dimension. *arXiv preprint arXiv:2402.18048*.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28.

# A In-Context Learning Experiments

## A.1 Llama-3-8B In-Context Learning Experiments

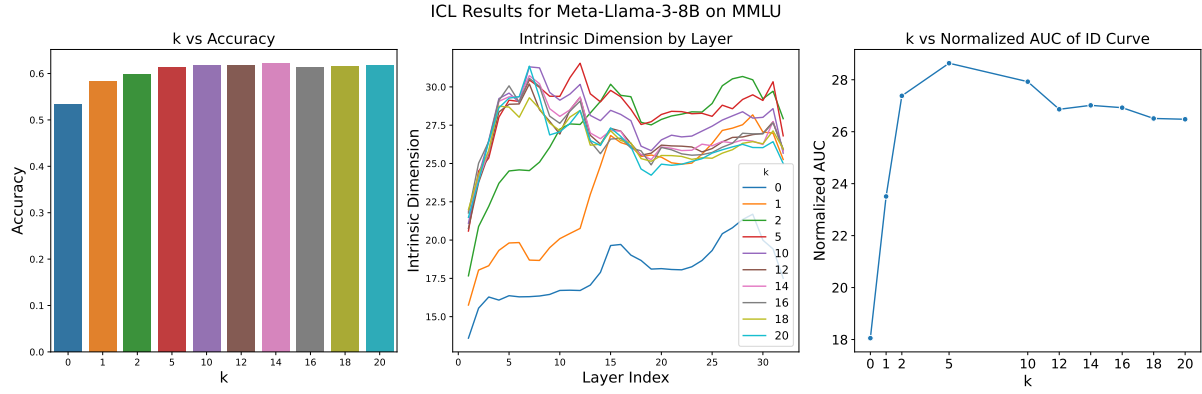


Figure 7: ICL Experiment Results for Meta-Llama-3-8B on MMLU

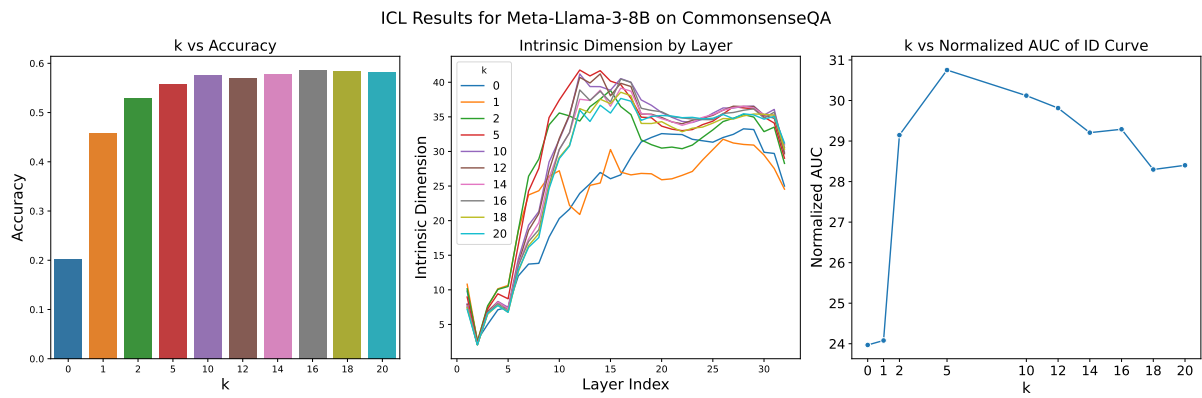


Figure 8: ICL Experiment Results for Meta-Llama-3-8B on CommonsenseQA

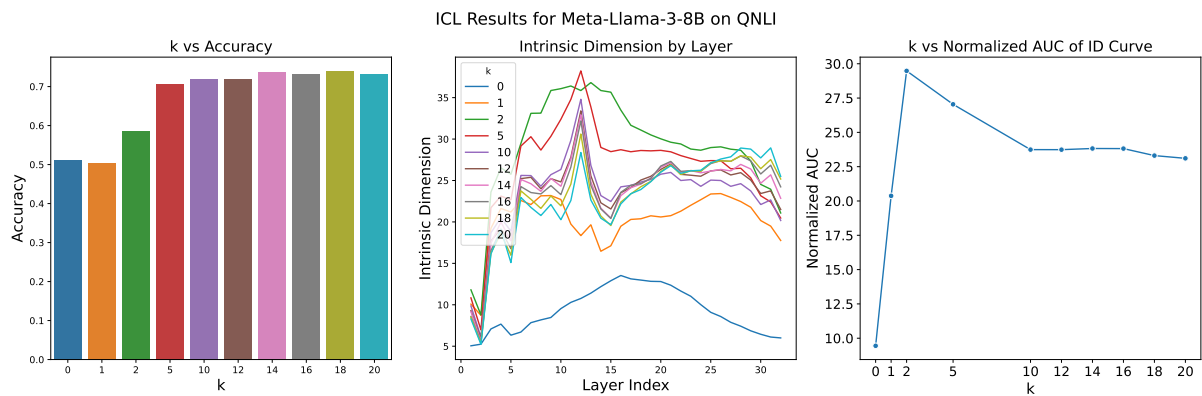


Figure 9: ICL Experiment Results for Meta-Llama-3-8B on QNLI

## A.2 Llama-2-13b In-Context Learning Experiments

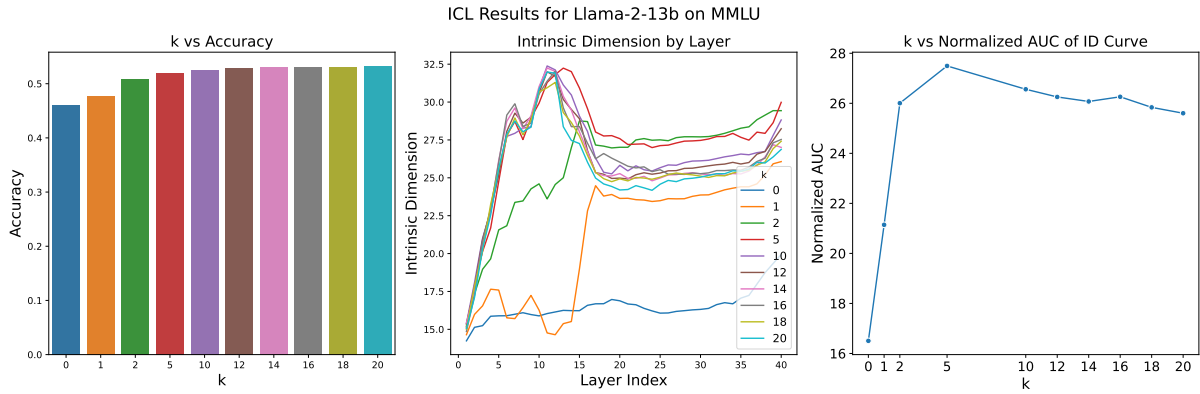


Figure 10: ICL Experiment Results for Llama-2-13b on MMLU

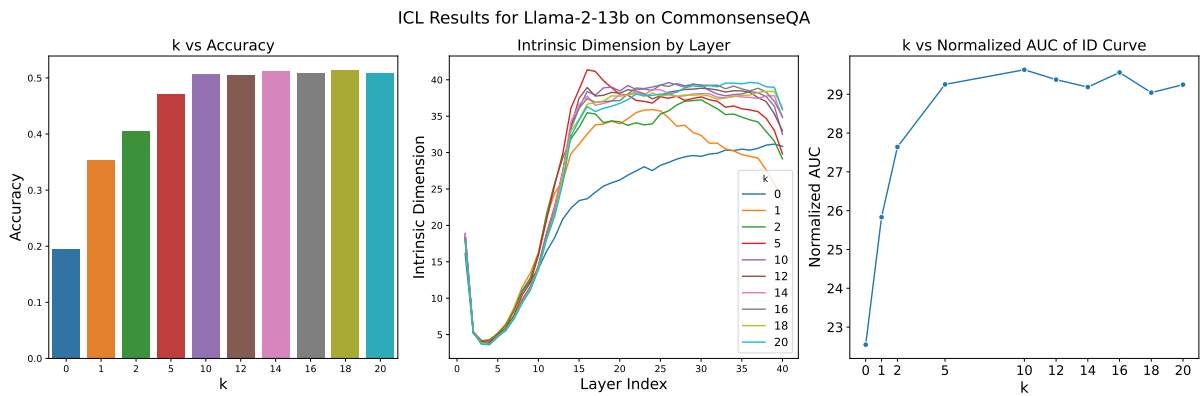


Figure 11: ICL Experiment Results for Llama-2-13b on CommonsenseQA

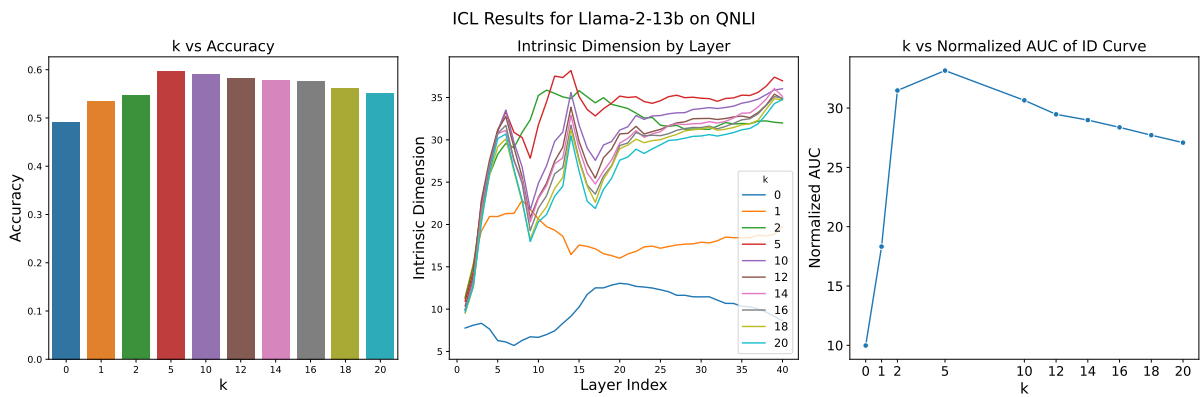


Figure 12: ICL Experiment Results for Llama-2-13b on QNLI

## B Supervised Fine-Tuning Experiments

### B.1 Supervised Fine-Tuning Results for Llama-3-8B

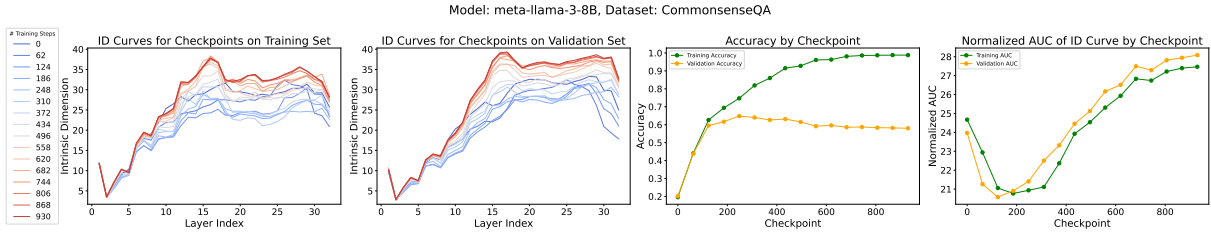


Figure 13: Supervised Fine-Tuning Results for Llama-3-8B on Commonsense QA

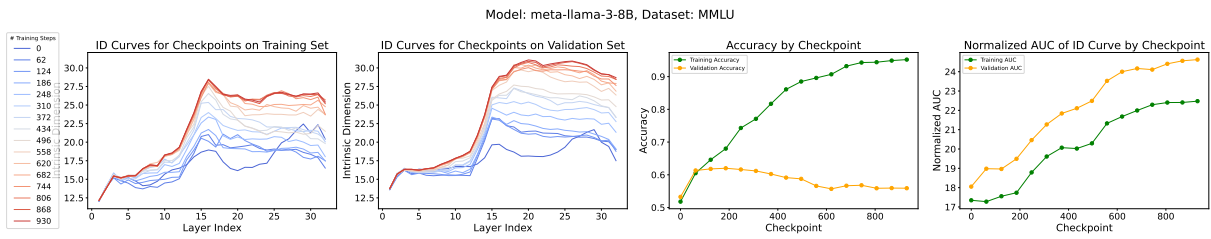


Figure 14: Supervised Fine-Tuning Results for Llama-3-8B on MMLU

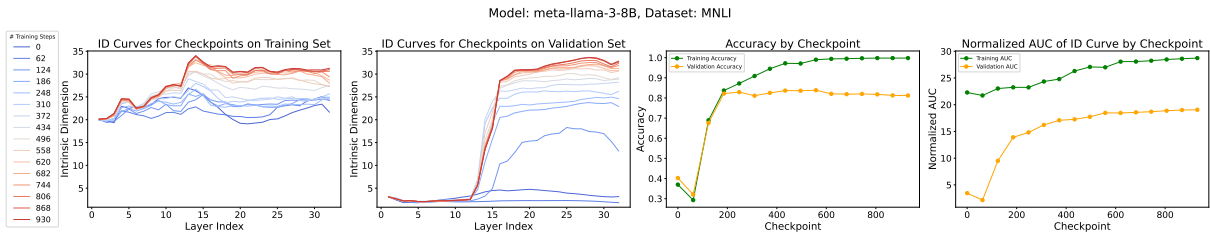


Figure 15: Supervised Fine-Tuning Results for Llama-3-8B on MNL1

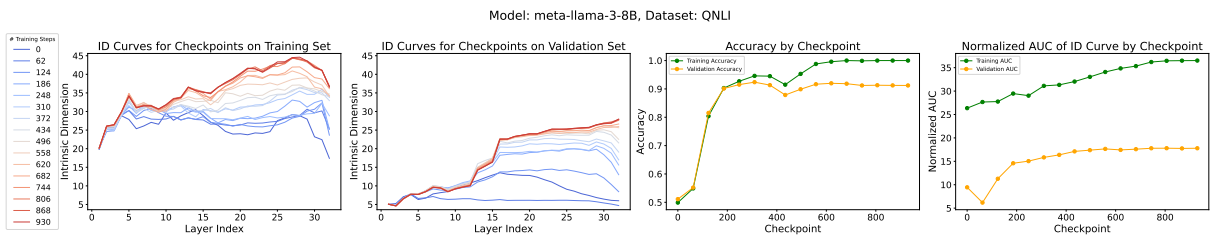


Figure 16: Supervised Fine-Tuning Results for Llama-3-8B on QNLI

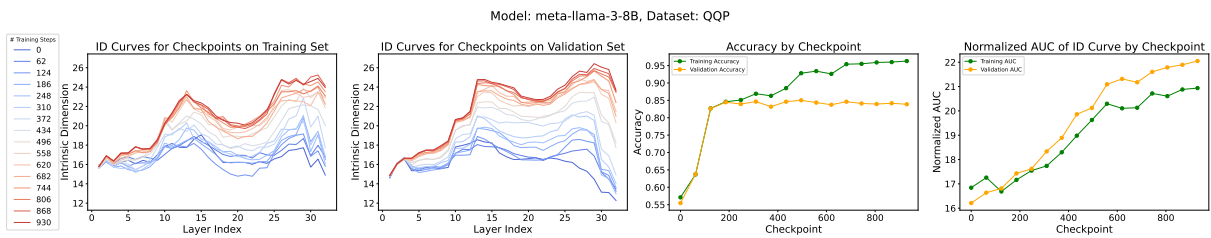


Figure 17: Supervised Fine-Tuning Results for Llama-3-8B on QQP

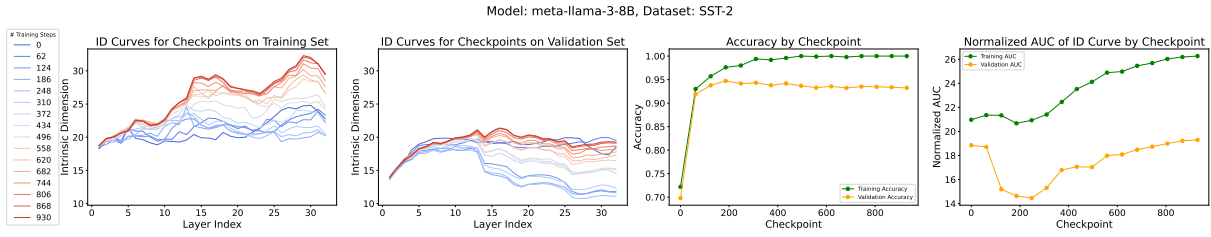


Figure 18: Supervised Fine-Tuning Results for Llama-3-8B on SST-2

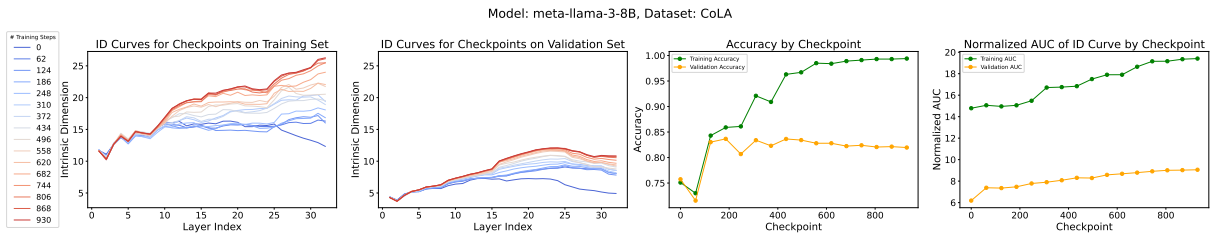


Figure 19: Supervised Fine-Tuning Results for Llama-3-8B on CoLA

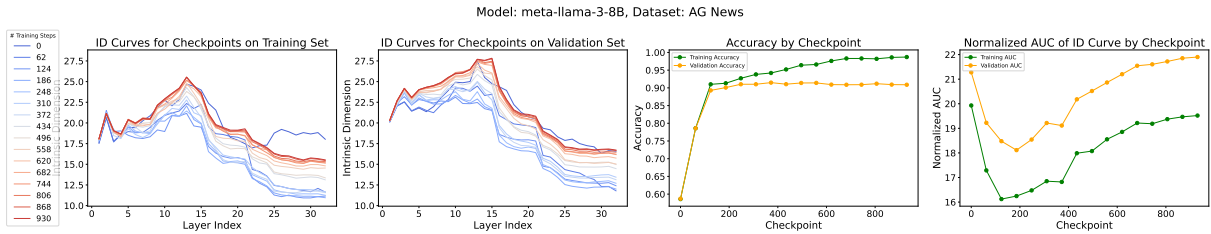


Figure 20: Supervised Fine-Tuning Results for Llama-3-8B on AG News

## B.2 Supervised Fine-Tuning Results for Llama-2-13B

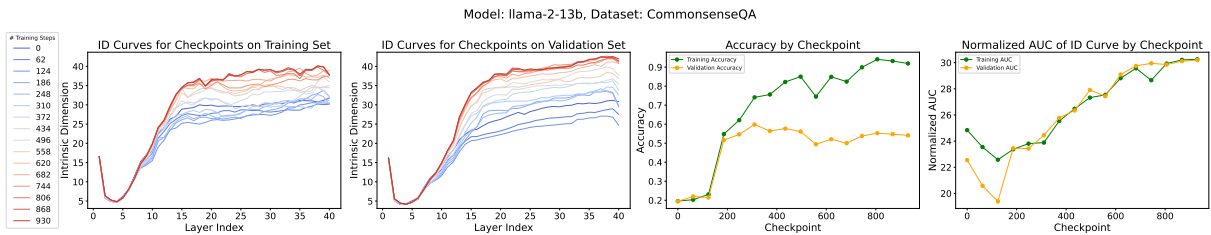


Figure 21: Supervised Fine-Tuning Results for Llama-2-13B on Commonsense QA

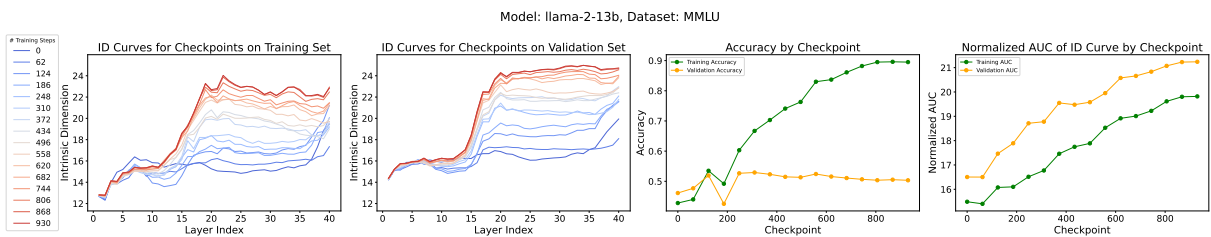


Figure 22: Supervised Fine-Tuning Results for Llama-2-13B on MMLU

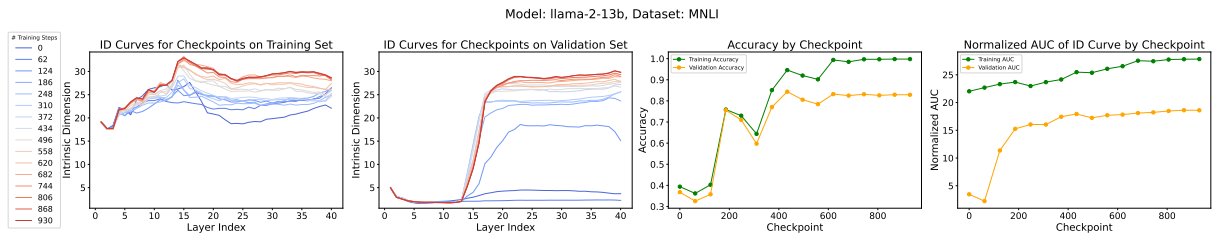


Figure 23: Supervised Fine-Tuning Results for Llama-2-13B on MNL

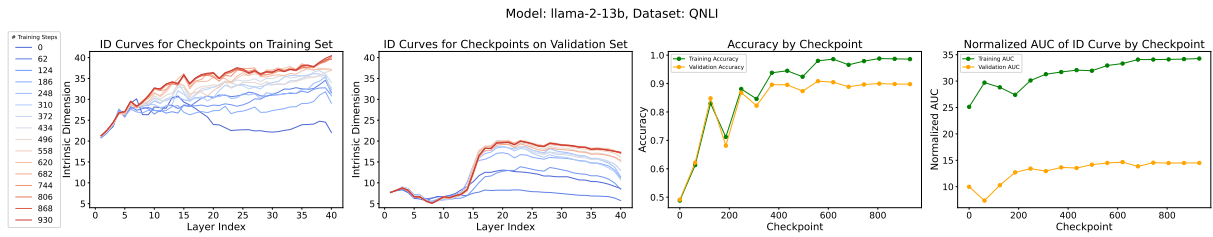


Figure 24: Supervised Fine-Tuning Results for Llama-2-13B on QNL

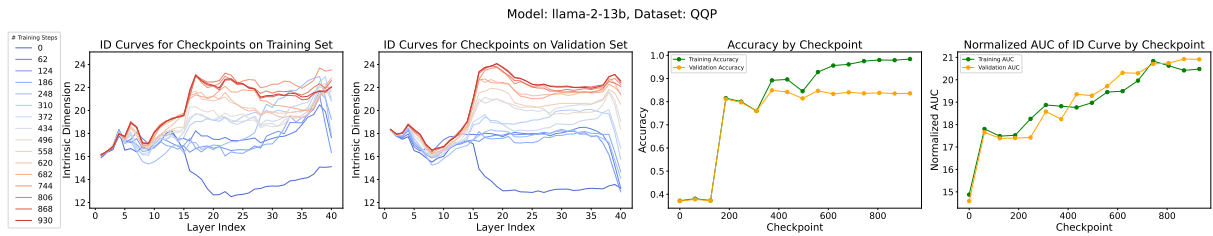


Figure 25: Supervised Fine-Tuning Results for Llama-2-13B on QQP

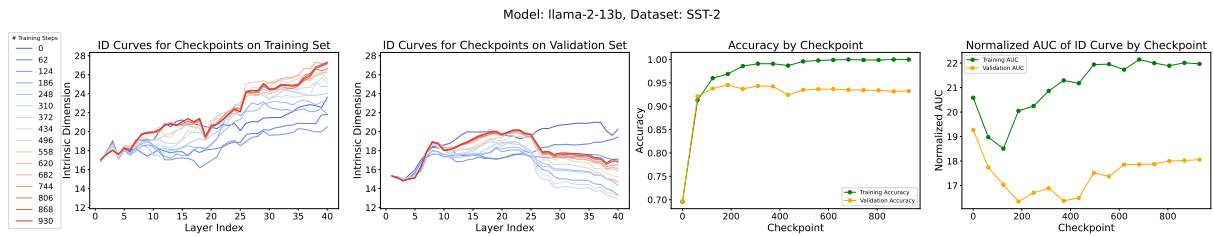


Figure 26: Supervised Fine-Tuning Results for Llama-2-13B on SST-2

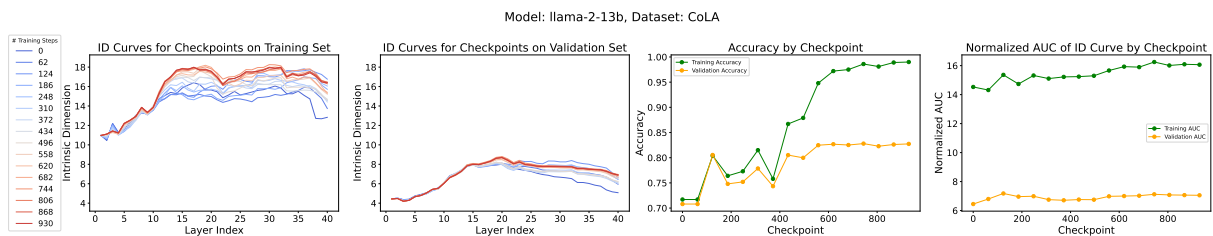


Figure 27: Supervised Fine-Tuning Results for Llama-2-13B on CoLA

Model: llama-2-13b, Dataset: AG News

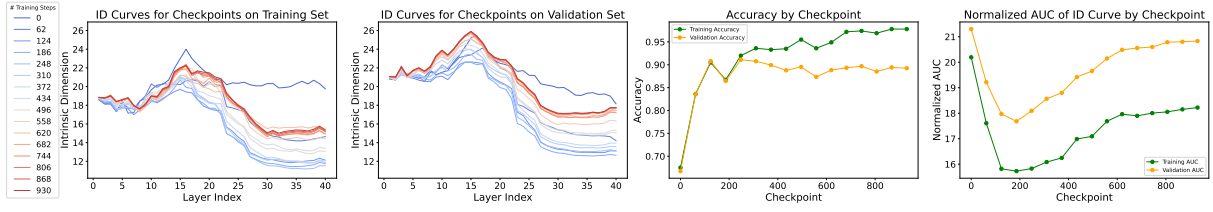


Figure 28: Supervised Fine-Tuning Results for Llama-2-13B on AG News



## C Comparisons of Supervised Fine-Tuning and In-Context Learning

Dataset: CommonsenseQA

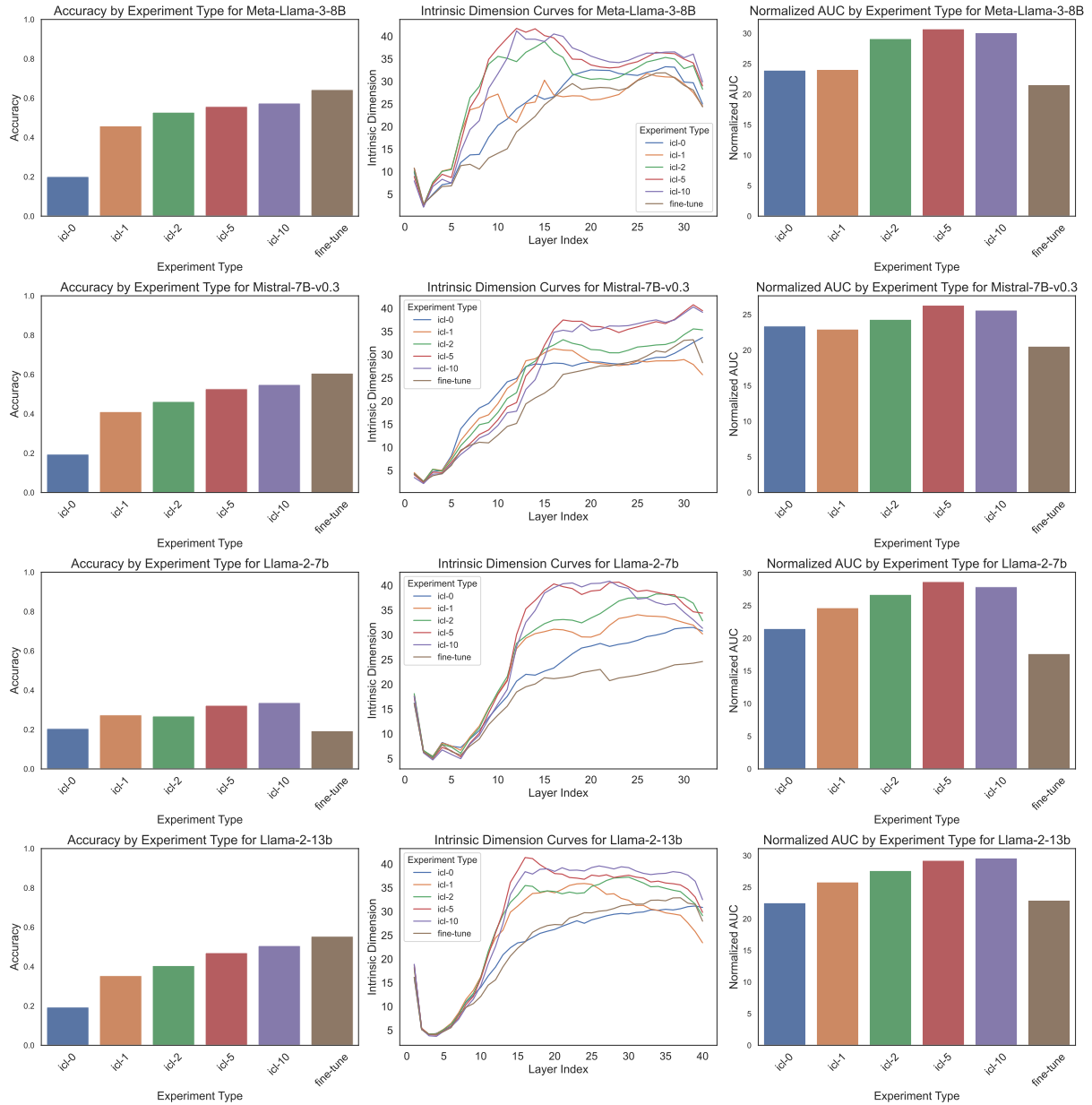


Figure 29: Comparison of Experimental Results for Commonsense QA

Dataset: MMLU

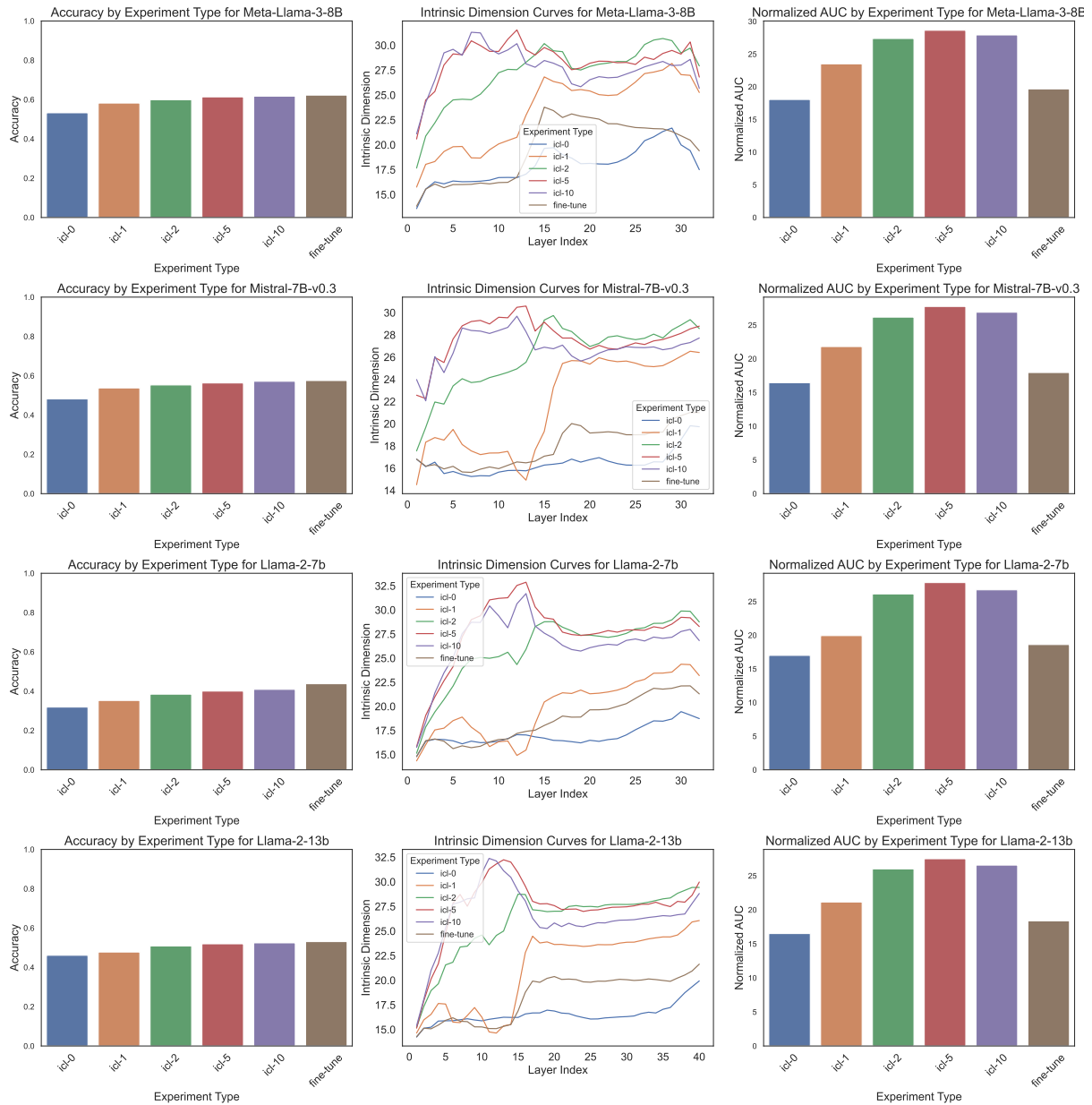


Figure 30: Comparison of Experimental Results for MMLU

Dataset: MNL

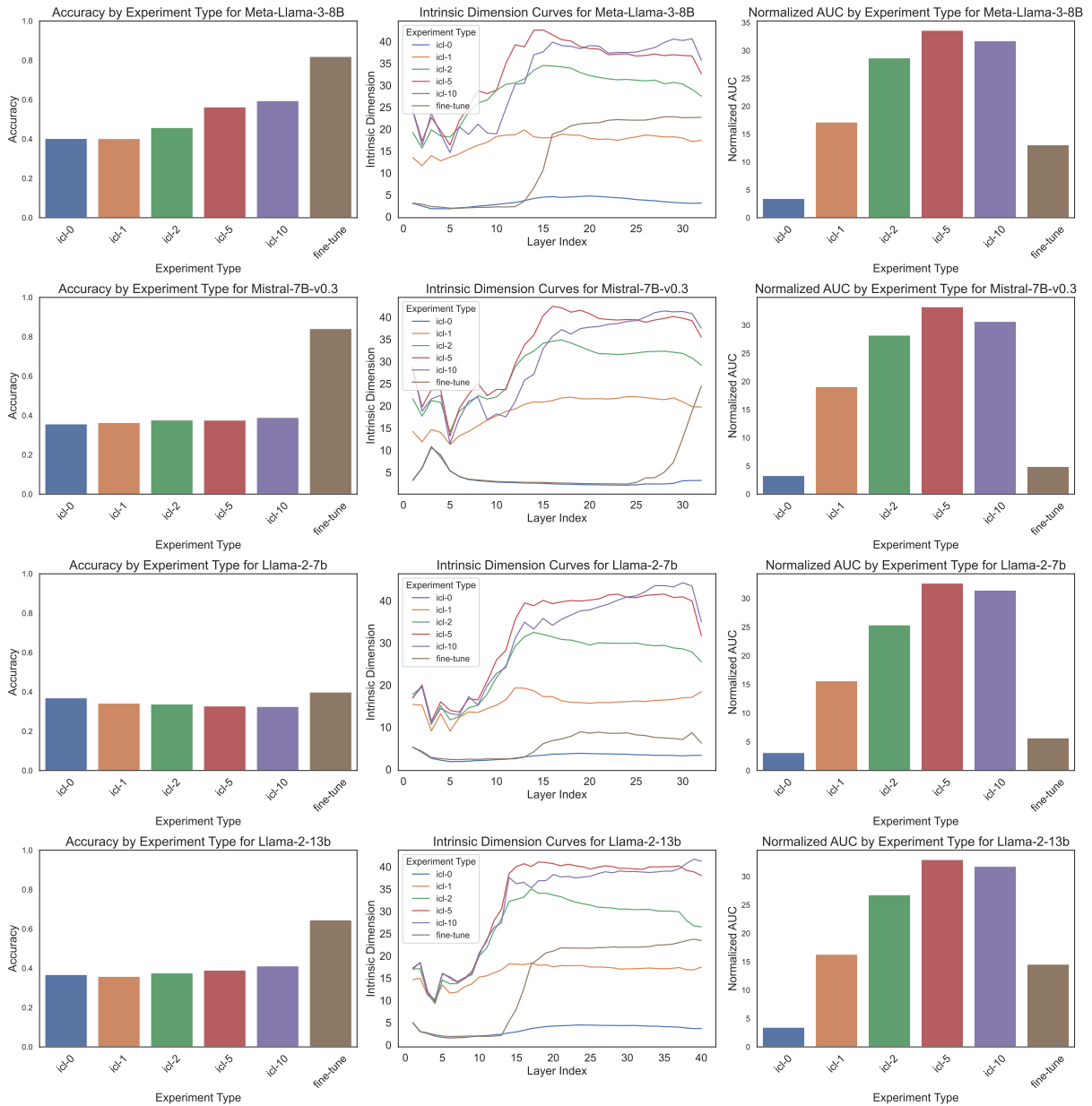


Figure 31: Comparison of Experimental Results for MNL

Dataset: QNLI

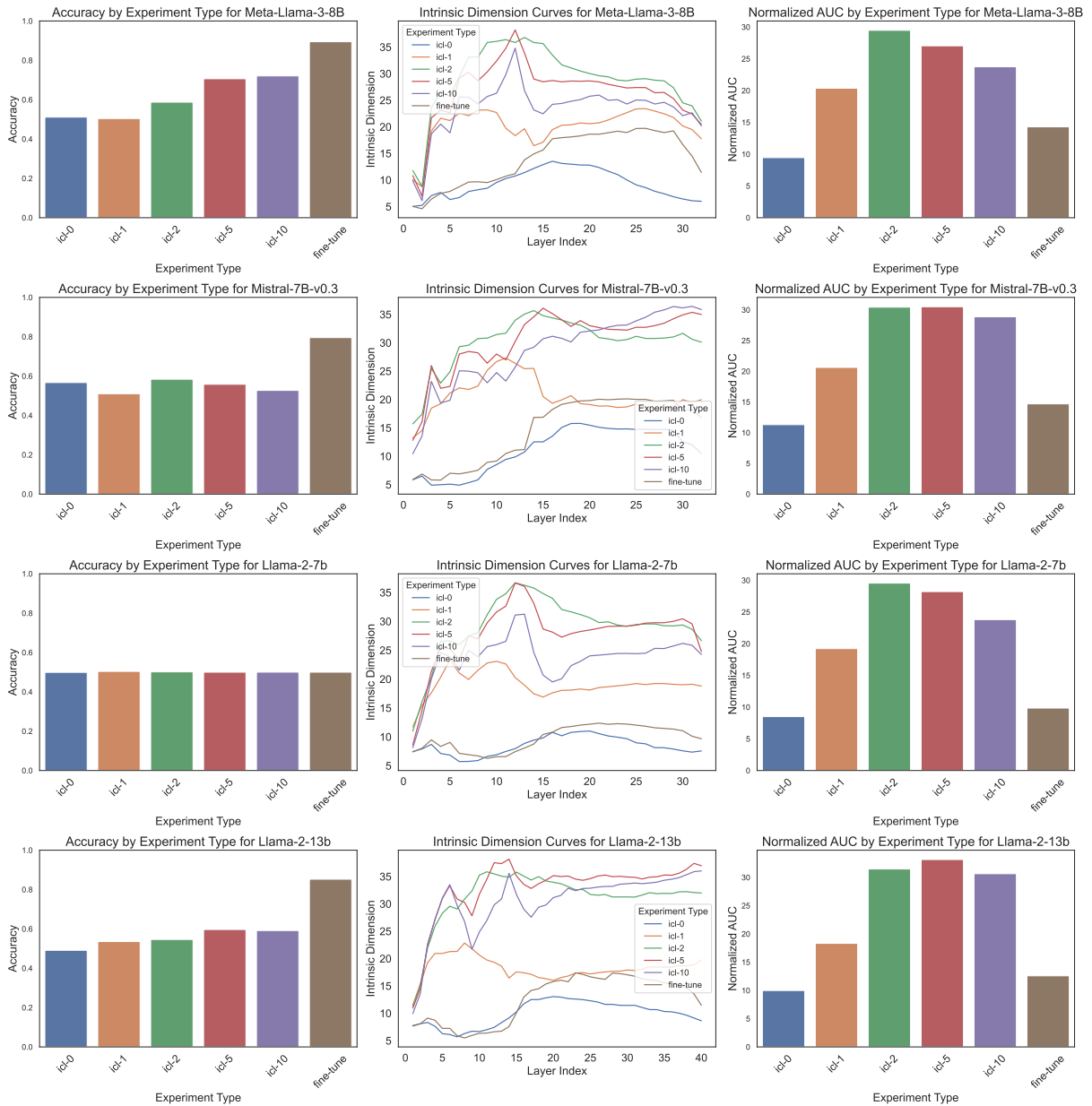


Figure 32: Comparison of Experimental Results for QNLI

Dataset: QQP

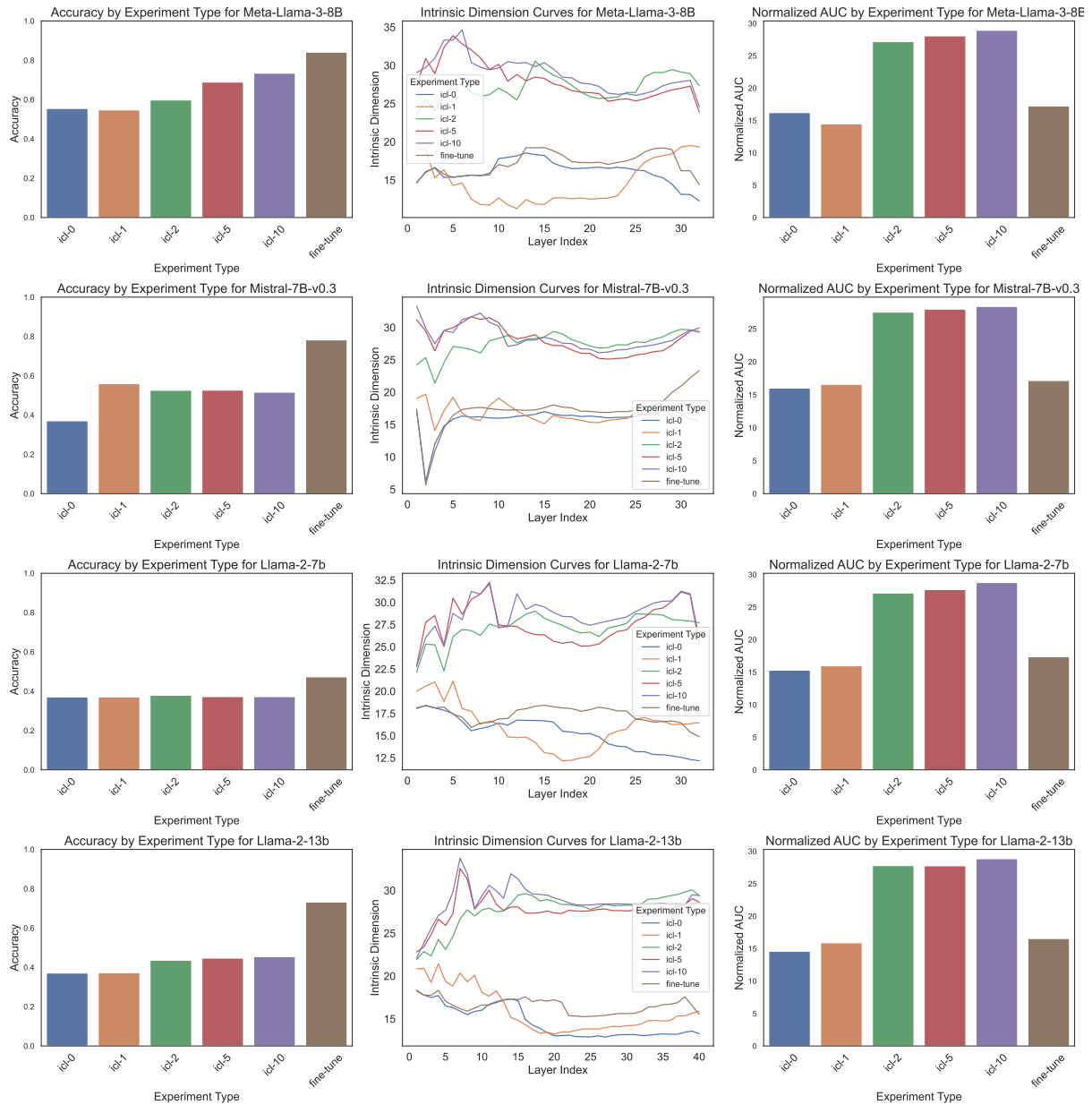


Figure 33: Comparison of Experimental Results for QQP

Dataset: SST-2

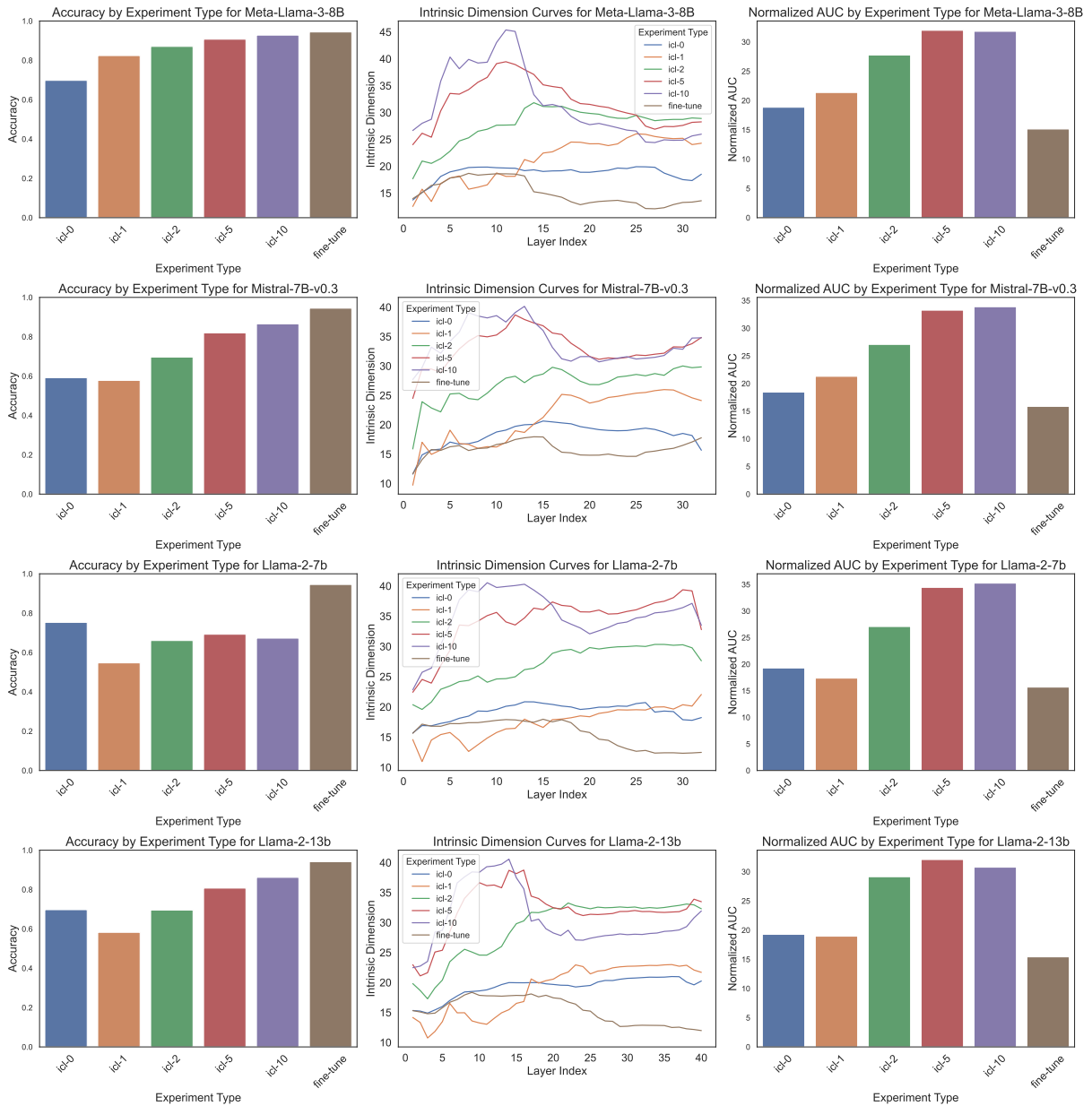


Figure 34: Comparison of Experimental Results for SST-2

Dataset: CoLA

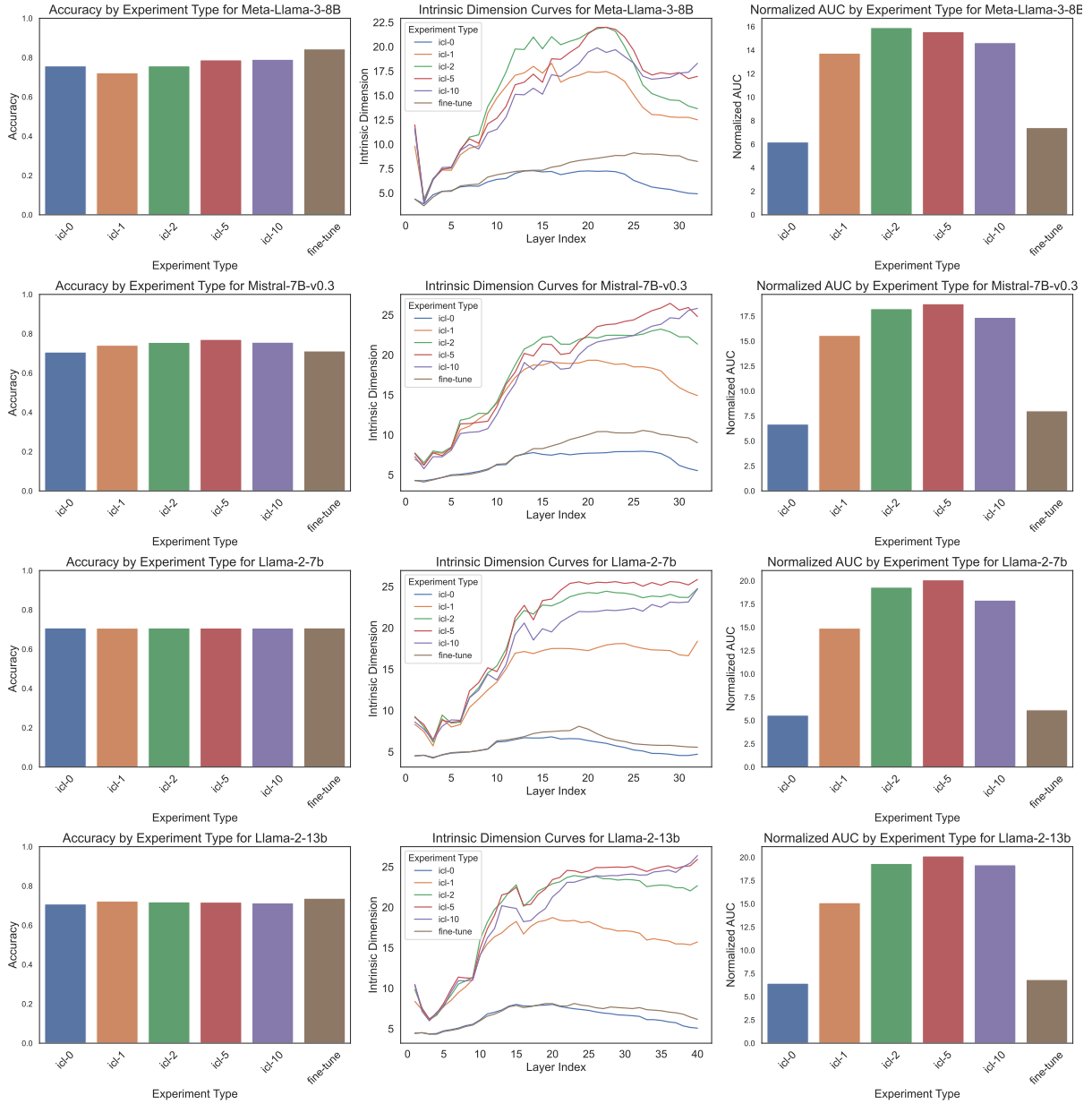


Figure 35: Comparison of Experimental Results for CoLA

Dataset: AG News

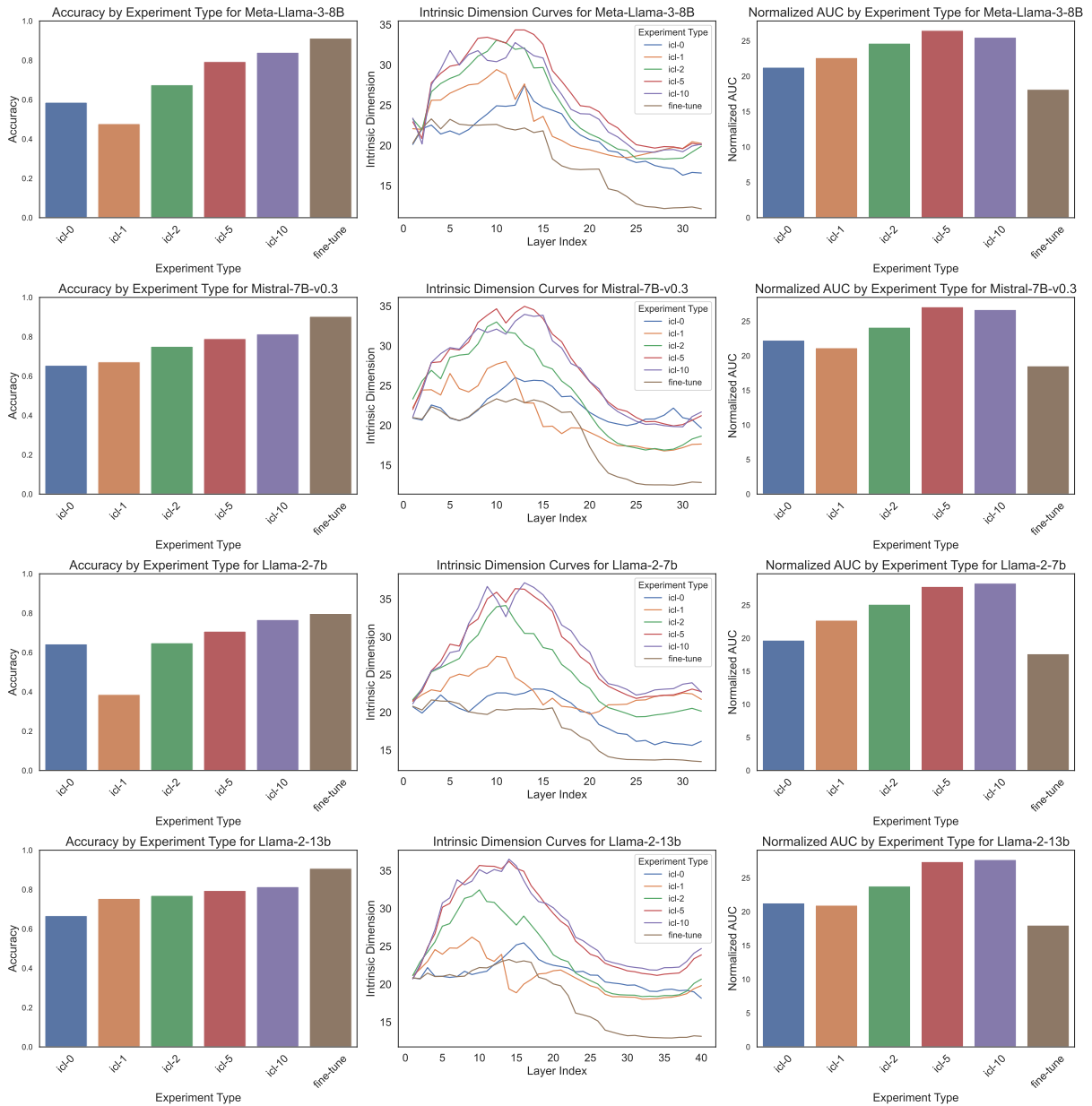


Figure 36: Comparison of Experimental Results for AG News



## D ICL Experiment Results with Unique Demonstrations

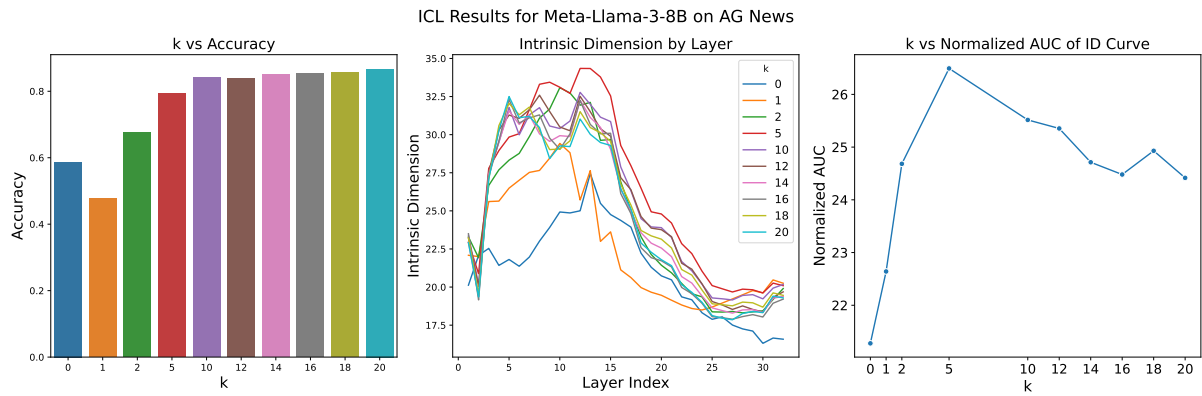


Figure 37: ICL Experiment Results with Unique Demonstrations on AGNews Dataset

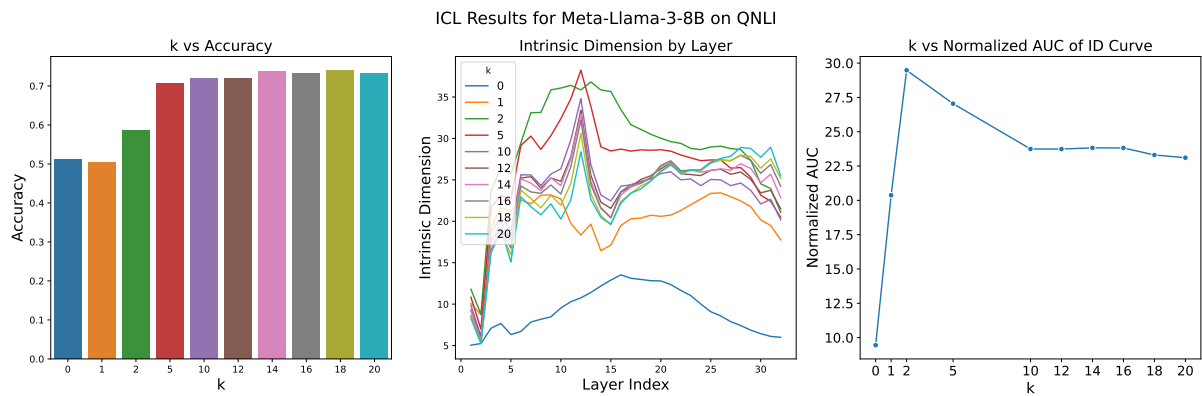


Figure 38: ICL Experiment Results with Unique Demonstrations on QNLI Dataset

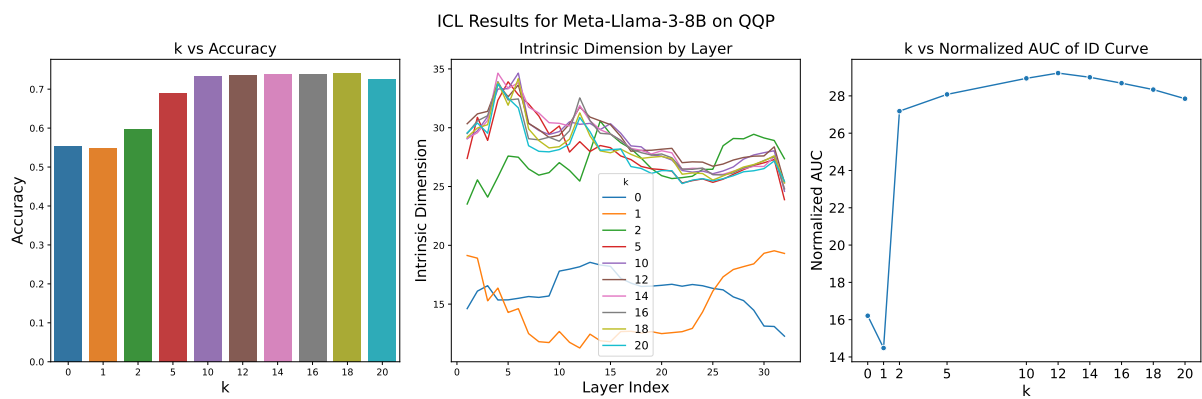


Figure 39: ICL Experiment Results with Unique Demonstrations on QQP Dataset

## E Normalized AUC Boxplot by Model for All Learning Paradigms

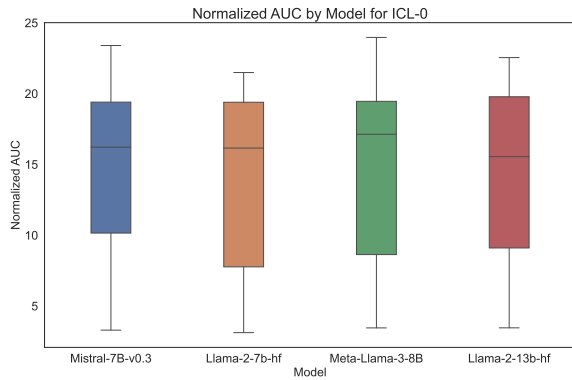


Figure 40: Normalized AUC by Model boxplot for ICL-0 experiments.

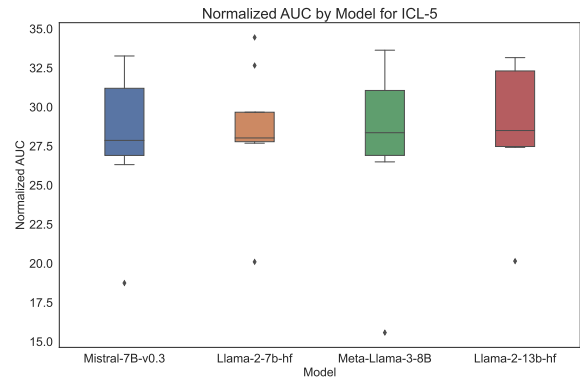


Figure 43: Normalized AUC by Model boxplot for ICL-5 experiments.

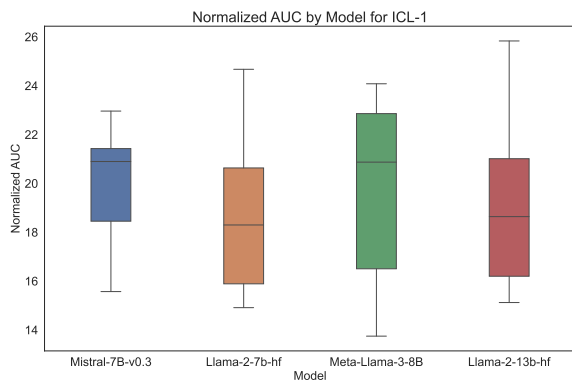


Figure 41: Normalized AUC by Model boxplot for ICL-1 experiments.

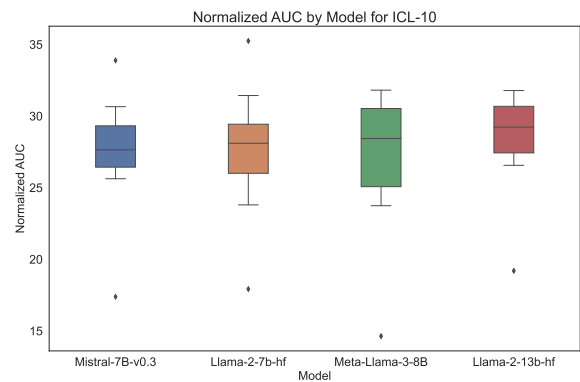


Figure 44: Normalized AUC by Model boxplot for ICL-10 experiments.

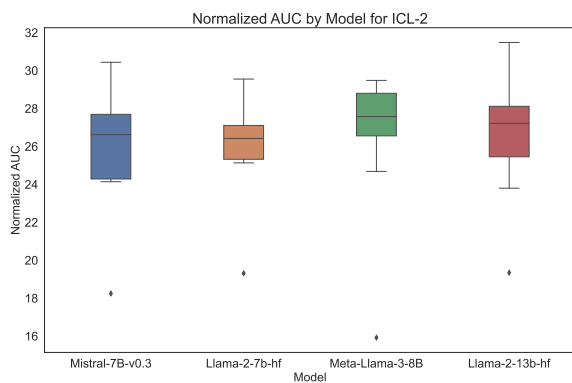


Figure 42: Normalized AUC by Model boxplot for ICL-2 experiments.

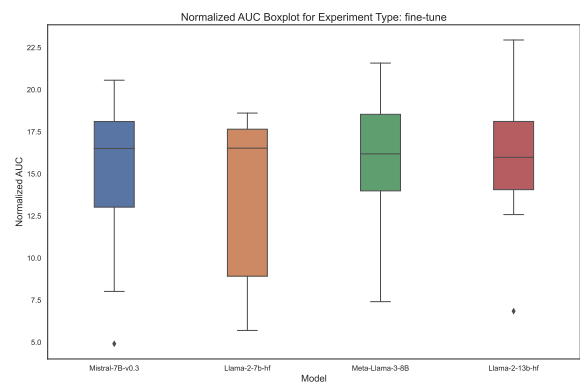


Figure 45: Normalized AUC by Model boxplot for SFT experiments.

## F Validating the TwoNN Estimator with the MLE Estimator

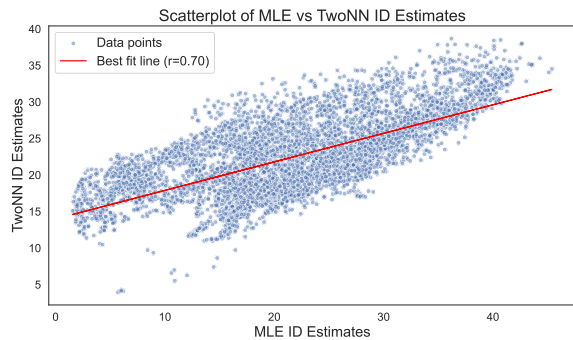


Figure 46: Scatterplot plotting ID estimation results for all experiments using the MLE and TwoNN Estimators.

To assess the validity of our intrinsic dimension estimator, we calculate the intrinsic dimension for different combinations of (learning paradigm, dataset, model, layer) using the TwoNN estimator and Maximum Likelihood Estimator (MLE) introduced by [Levina and Bickel \(2004\)](#). We use a neighborhood of size  $k = 50$  when applying MLE. We find that the estimates from the two estimators are correlated with  $r = 0.7$ . While it is not possible to know the 'true' intrinsic dimensionality of the representations, high correlation between two separate estimators provides a sanity check for our choice of the TwoNN estimator.

## G Dataset Generation Details

### H Dataset Details

We include details about dataset generation below. We get prompts for all datasets except MMLU from the PromptSource library ([Bach et al., 2022](#)).

#### H.1 QNLI

Items for the training and validation splits in our QNLI experiments were taken from the official QNLI 'train' and 'validation' splits respectively.

##### Prompt Template:

```
Does that sentence have all you need to
  ↳ answer the question
  ↳ "{{question}}"?
|||
{{answer_choices[label]}}
```

**Labels:** ['yes', 'no']

#### H.2 CommonsenseQA

Items for both the training and validation splits in our CommonsenseQA experiments were taken

from the official CommonsenseQA 'train' split.

##### Prompt Template:

```
Given the following options, what do
  ↳ you think is the correct answer
  ↳ to the question below:
```

```
{{question}}
```

Options:

```
{% for letter, t in zip(answer_choices,
  ↳ choices.text) %}
- {{letter}}: {{t}}
{% endfor %} |||
{{answerKey}}
{% endif %}
```

**Labels:** ['A', 'B', 'C', 'D']

### H.3 MMLU

Items for both the training and validation splits in our MMLU experiments were taken from the official MMLU 'test' split.

##### Prompt Template:

```
# generate input txt and output txt
letters = ['A', 'B', 'C', 'D']
choices = dataset_element['choices']

input_txt =
  ↳ f"{dataset_element['question']}\n\nA:
  ↳ {choices[0]}\nB:
  ↳ {choices[1]}\nC:
  ↳ {choices[2]}\nD:
  ↳ {choices[3]}\nAnswer:"

output_txt = letters[answer_idx]
combined = input_txt + output_txt
```

### H.4 SST-2

Items for both the training and validation splits in our SST-2 experiments were taken from the official SST-2 'train' split.

##### Prompt Template:

```
{{sentence}}
Question: Was that sentence
  ↳ {"positive"} or
  ↳ {"negative"}? Answer: ||| {{
  ↳ answer_choices[label]}}
```

**Labels:** ['negative', 'positive']

### H.5 CoLA

Items for both the training and validation splits in our CoLA experiments were taken from the official CoLA 'train' split.

##### Prompt Template:

```
Does the following sentence make sense
  ↳ and use correct English? Please
  ↳ answer {"yes"} or {"no"}.
{{sentence}}
|||
{{answer_choices[label]}}
```

**Labels:** ['no', 'yes']

## H.6 AGNews

Items for the training and validation splits in our AGNews experiments were taken from the official AGNews 'train' and 'validation' splits respectively.

### Prompt Template:

```
What label best describes this news
  ↳ article?
{{text}} |||
{{answer_choices[label] }}
```

**Labels:** ['World politics', 'Sports', 'Business', 'Science and technology']

## H.7 MNLI

Items for the training and validation splits in our MNLI experiments were taken from the official MNLI 'train' and 'validation\_matched' splits respectively.

### Prompt Template:

```
{{premise}} Are we justified in saying
  ↳ that "{{hypothesis}}"? Yes, no,
  ↳ or maybe? ||| {{
  ↳ answer_choices[label] }}
```

**Labels:** ['Yes', 'Maybe', 'No']

## H.8 QQP

Items for the training and validation splits in our QQP experiments were taken from the official QQP 'train' and 'validation' splits respectively.

### Prompt Template:

```
I'm an administrator on the website
  ↳ Quora. There are two posts, one
  ↳ that asks "{{question1}}" and
  ↳ another that asks
  ↳ "{{question2}}". I can merge
  ↳ questions if they are asking the
  ↳ same thing. Can I merge these
  ↳ two questions? ||| {{
  ↳ answer_choices[label] }}
```

**Labels:** ['no', 'yes']