Towards Better Adaptation of Foundation Models

Zhuoyan Xu

Advisors: Prof. Yin Li, Prof. Yingyu Liang, Prof. Yiqiao Zhong

Department of Statistics

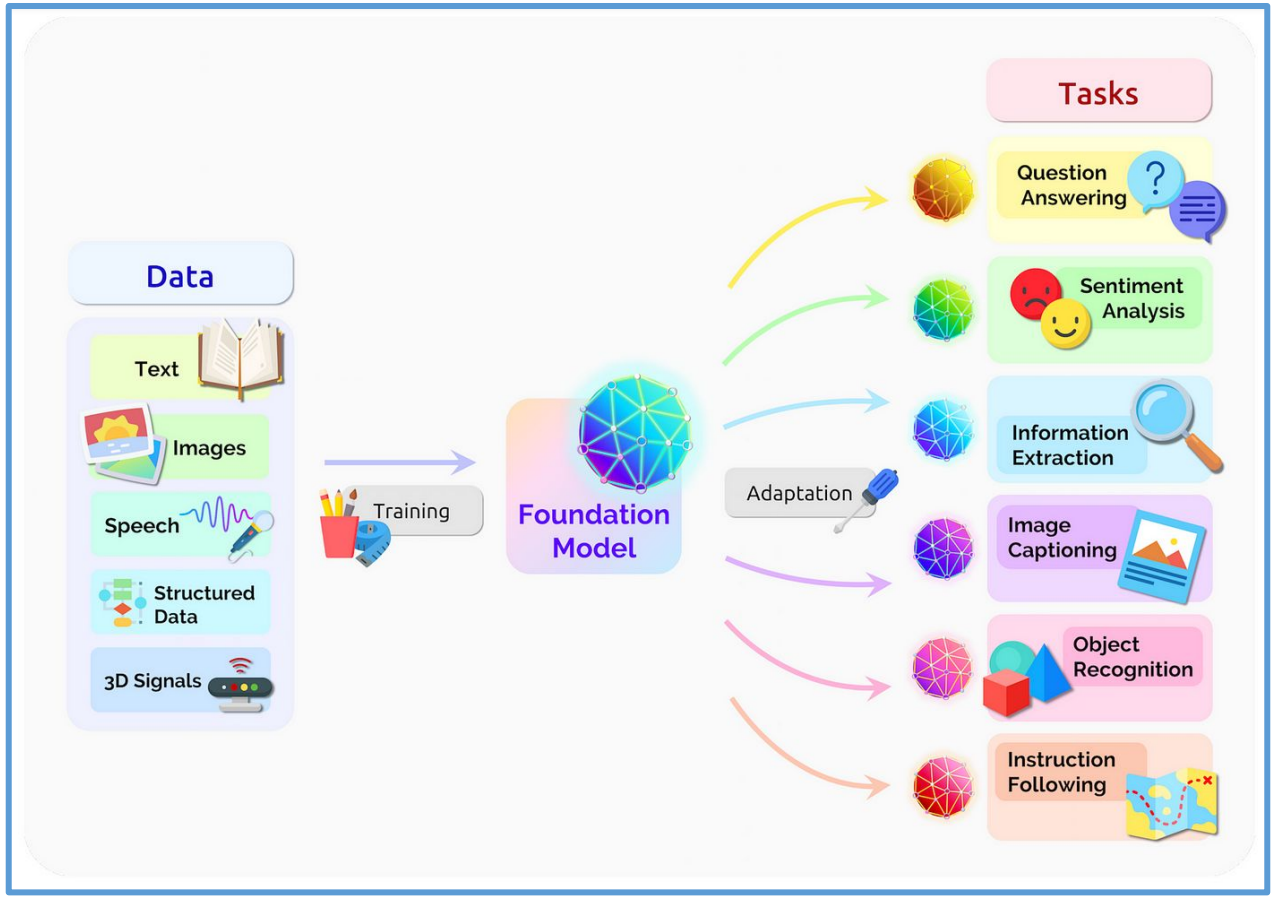
University of Wisconsin–Madison

Ph.D. Final Defense

Dec 16, 2025

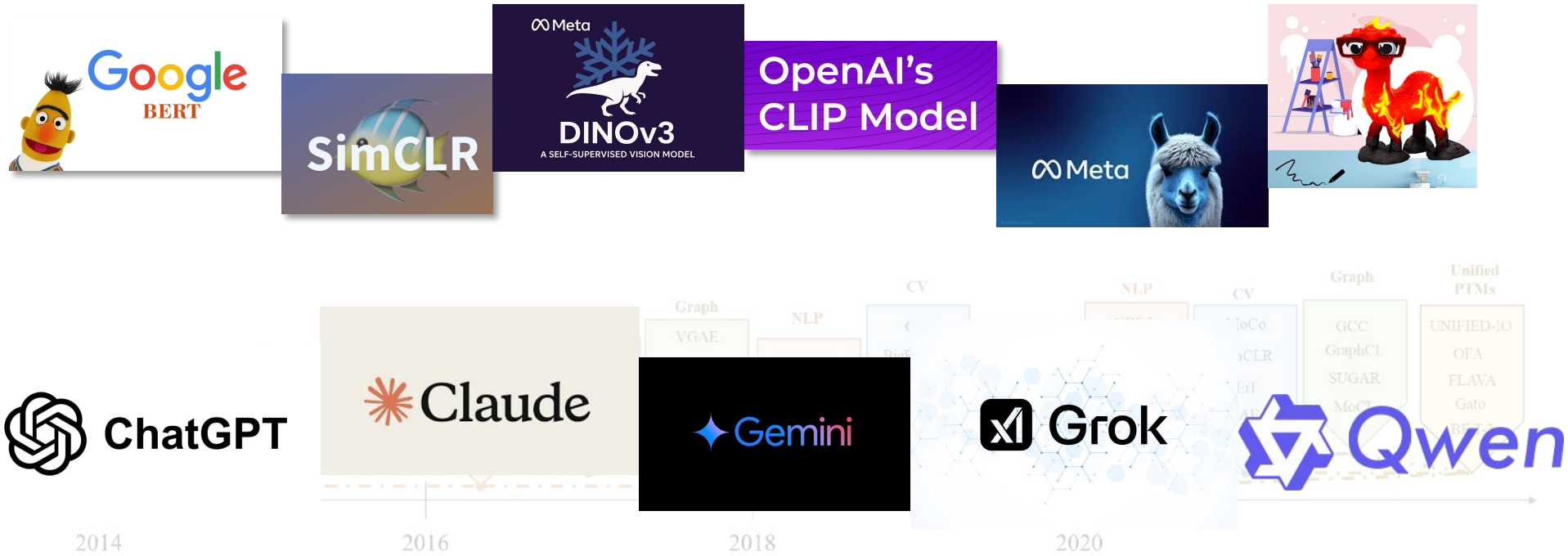
Foundation Models

A foundation model (FM) is a model trained on broad massive data that can be adapted (e.g., finetuned) to a wide range of downstream tasks.



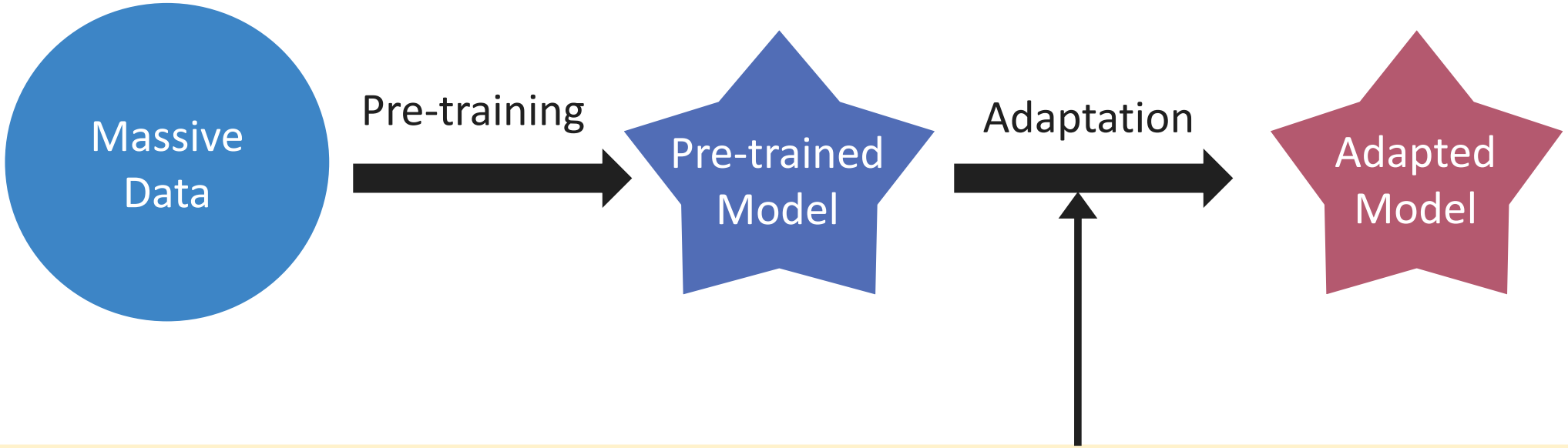
Figures from: *On the opportunities and risks of foundation models, 2021.*

Evolution of Foundation Models



The history and evolution of foundation models

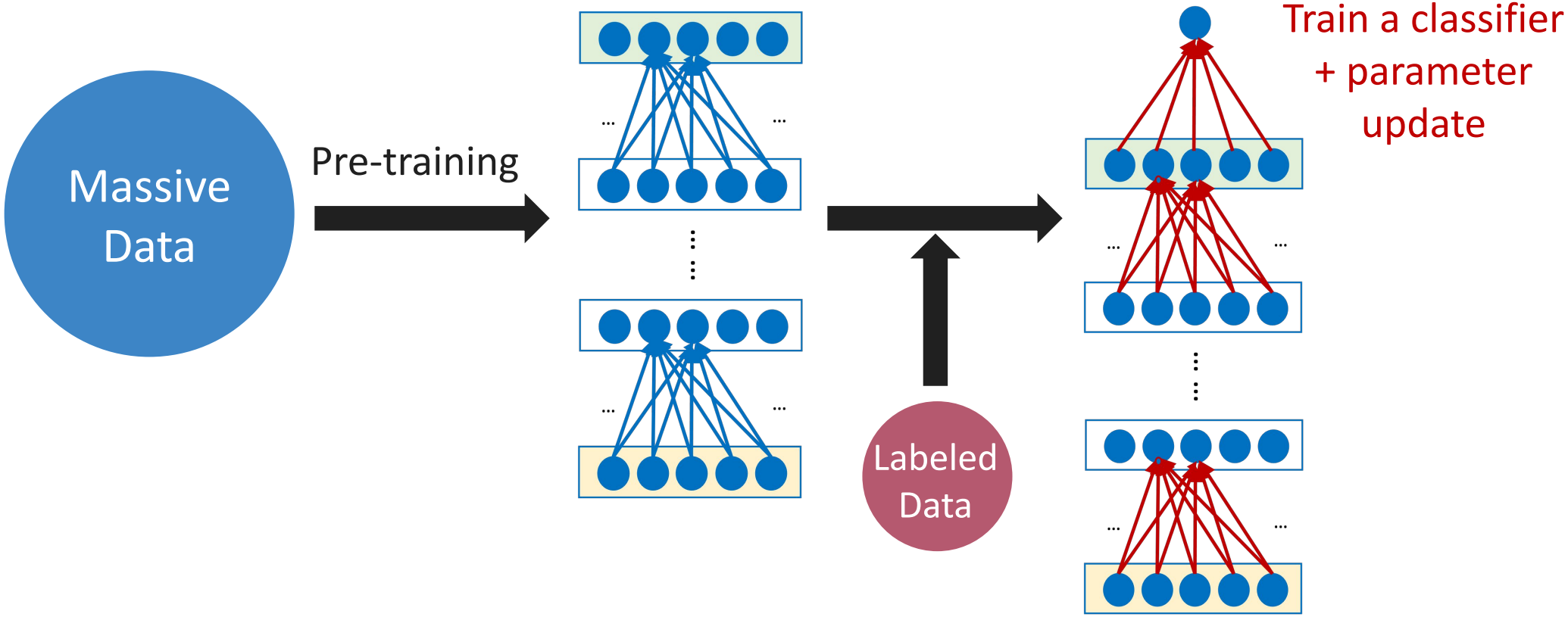
Adaptation of Foundation Models



Parameter & Architectural based: e.g. finetuning, adapters, distillation

Context based: e.g. ICL, CoT, RAG

Adaptation of Foundation Models: Finetuning



Adaptation: In-Context Learning (ICL)

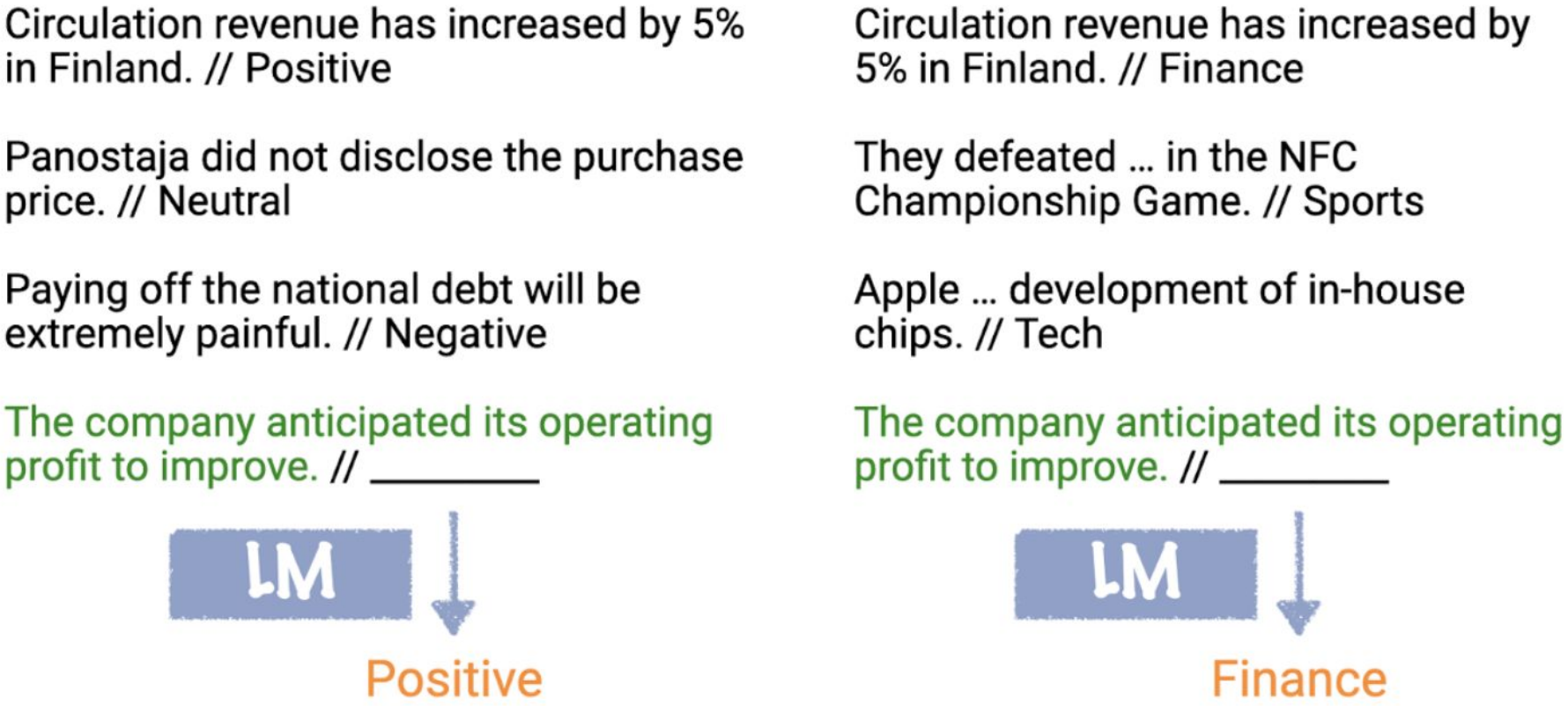


Fig Source: *How does in-context learning work? A framework for understanding the differences from traditional supervised learning, 2022.*

Specialization Gap

Foundation model

Broad capabilities

Acquire: Data efficiency

Lack fine-grained domain knowledge under limited supervision

recognizing rare medical conditions or nuanced scientific categories

Organize: Compositional reasoning

Lacks structured patterns for compositional & multi-step tasks

QA + Translation, Summarize + Extract

Deploy: Adaptivity

Lack adaptivity under strict and fluctuating resource constraints

Edge devices, real-time applications

Specialized Assistant

Deep expertise

Fundamental Question

How can we bridge this specialization gap?

Thesis Statement

Foundation models can be effectively specialized for downstream tasks through

(1) data-centric multitask adaptation,

(2) understanding of compositional reasoning mechanisms,

(3) dynamic inference strategies.

Thesis Outline

How can we bridge this specialization gap?

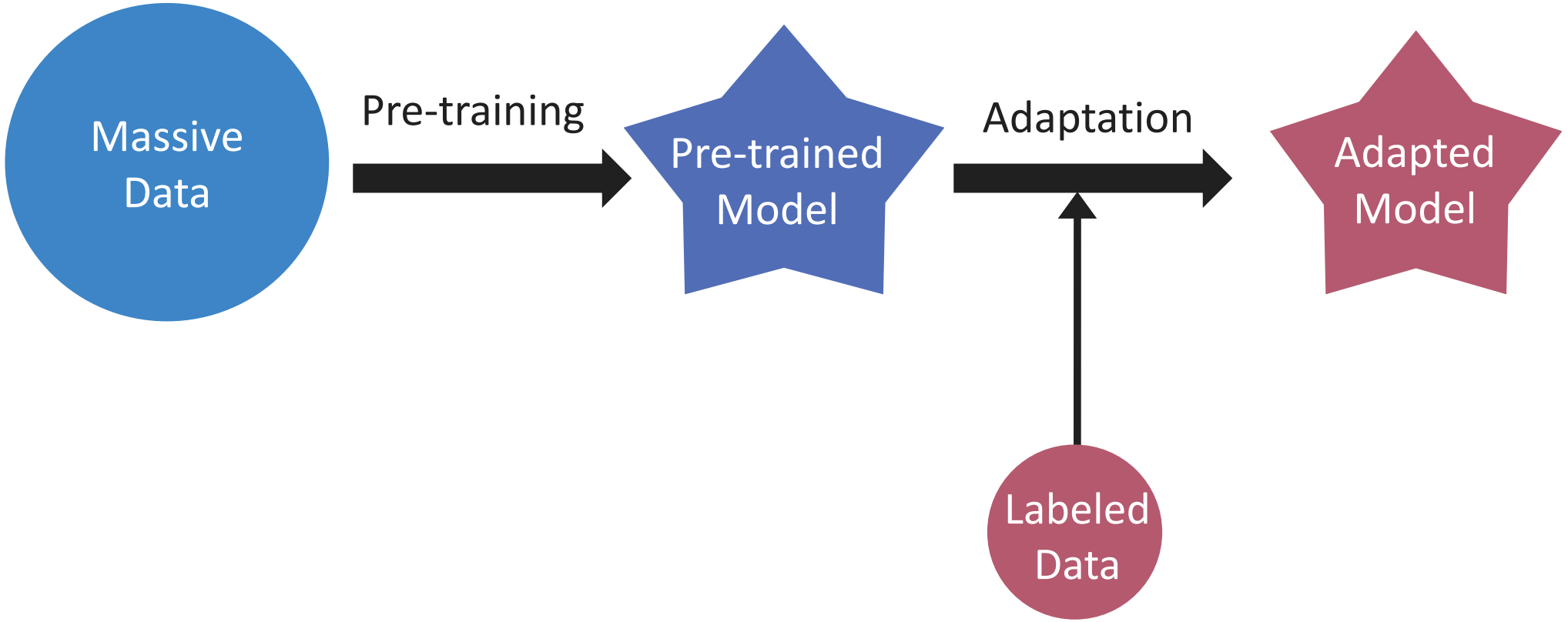
1. **Data-efficient Adaptation:** Enables FMs to better specialize in tasks in different domains (Part 1)
2. **Compositional Ability:** Advances FMs' ability to handle complex problems by combining simple tasks (Part 2)
3. **Adaptive Inference:** Make FMs deployable under varying computational constraints (Part 3)

Thesis Outline

How can we bridge this specialization gap?

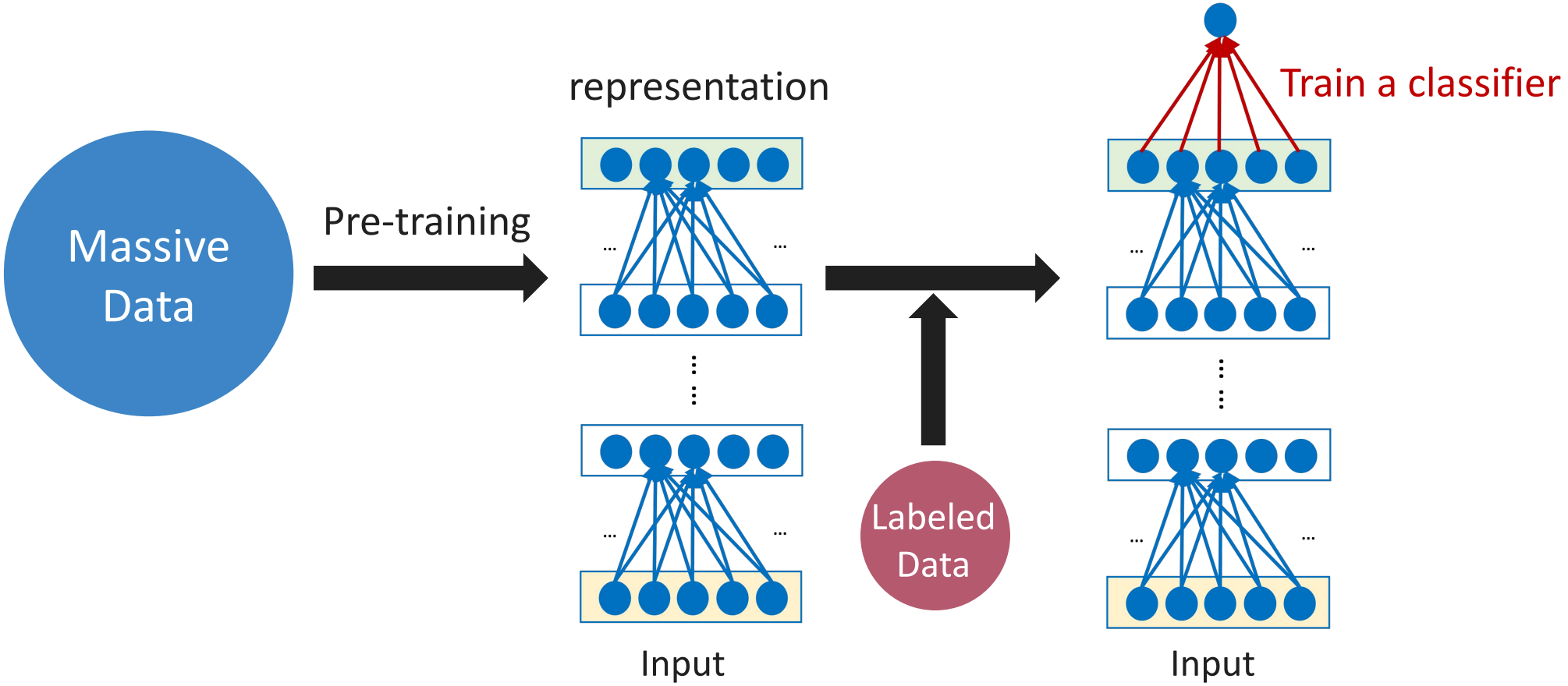
1. **Data-efficient Adaptation:** Enables FMs to better specialize in tasks in different domains (Part 1)
2. **Compositional Ability:** Advances FMs' ability to handle complex problems by combining simple tasks (Part 2)
3. **Adaptive Inference:** Make FMs deployable under varying computational constraints (Part 3)

Adaptation of Foundation Models



New Paradigm: Pre-trained Representations

Paradigm shift: supervised learning ⟹ pre-training + adaptation



New Paradigm: Pre-trained Representations

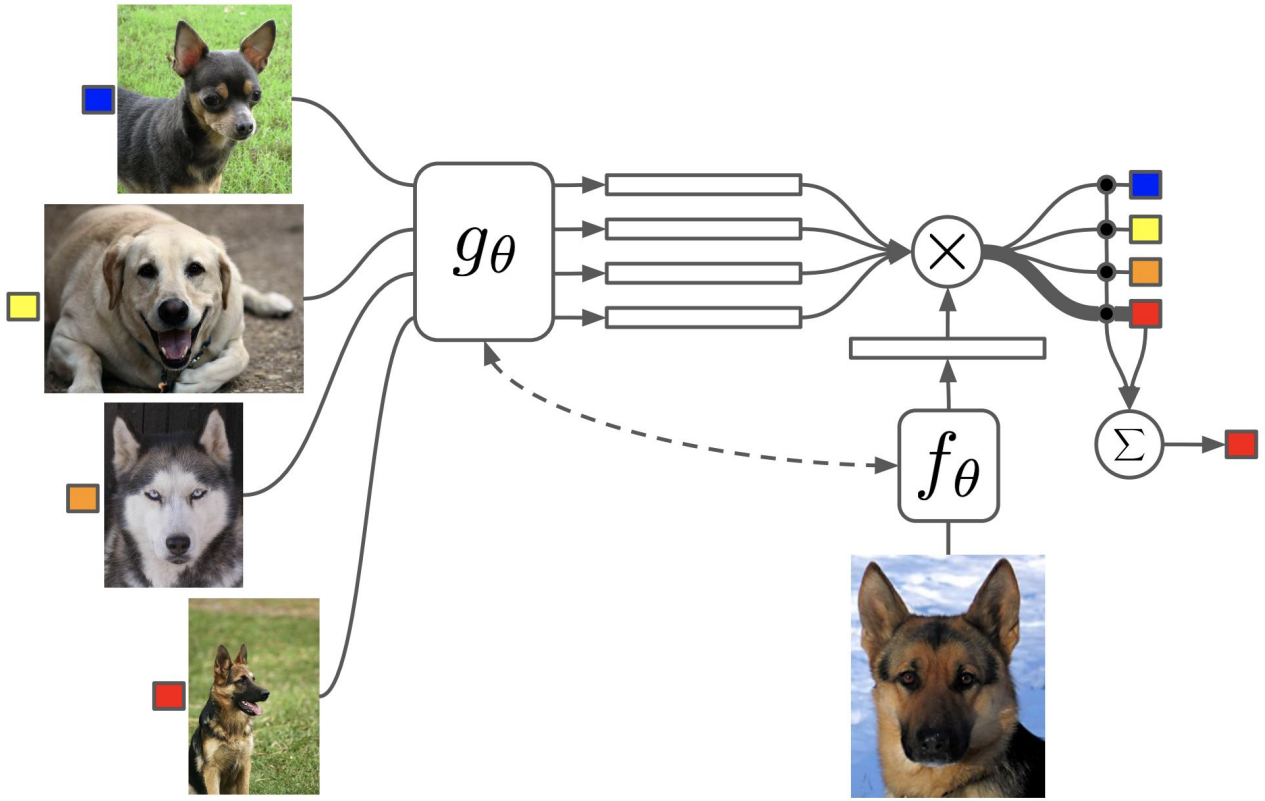
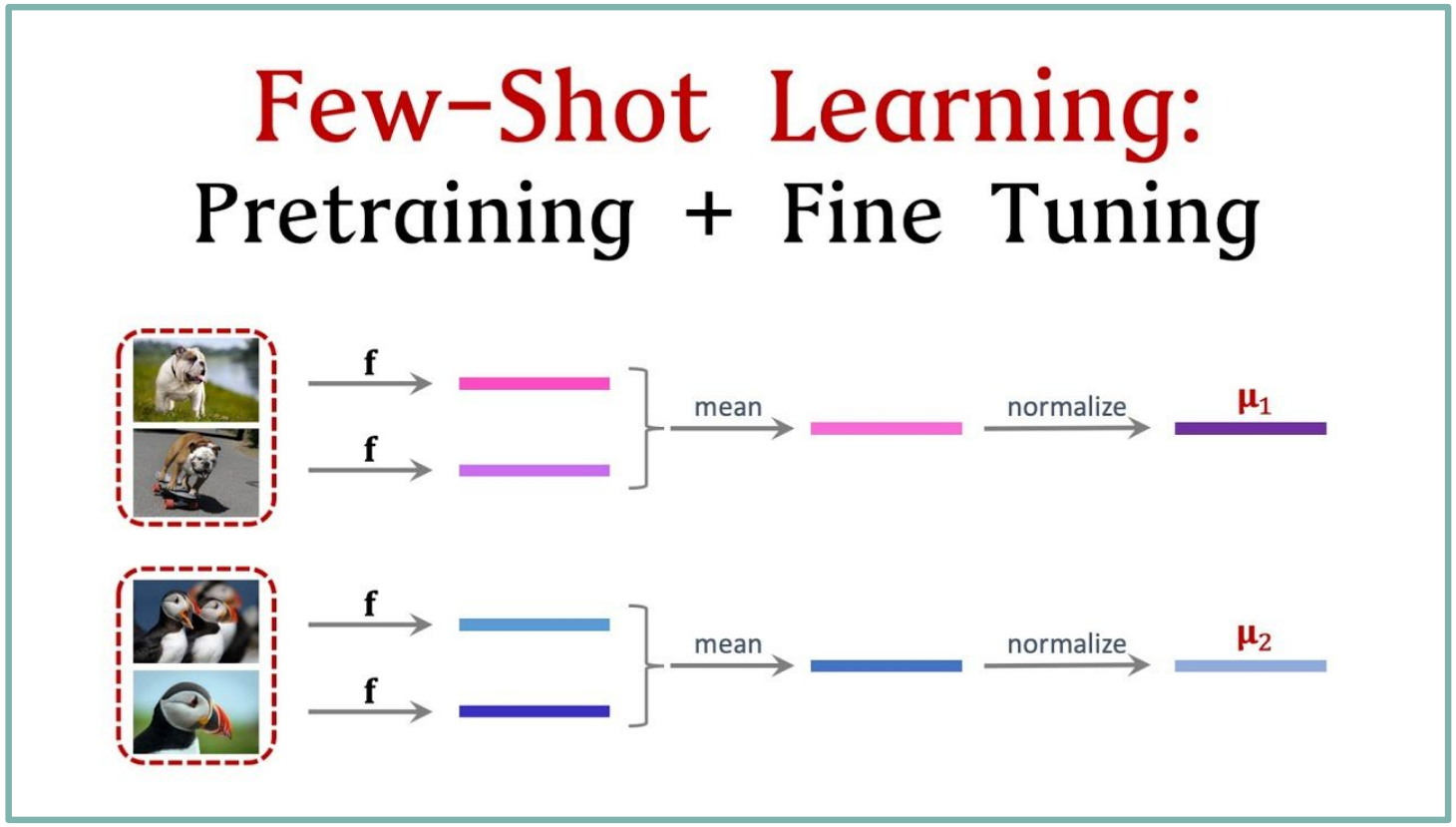


Figure 1: Matching Networks architecture

Adaptation of a pre-trained image encoder

Figures from: *Matching Networks for One Shot Learning, 2017.*

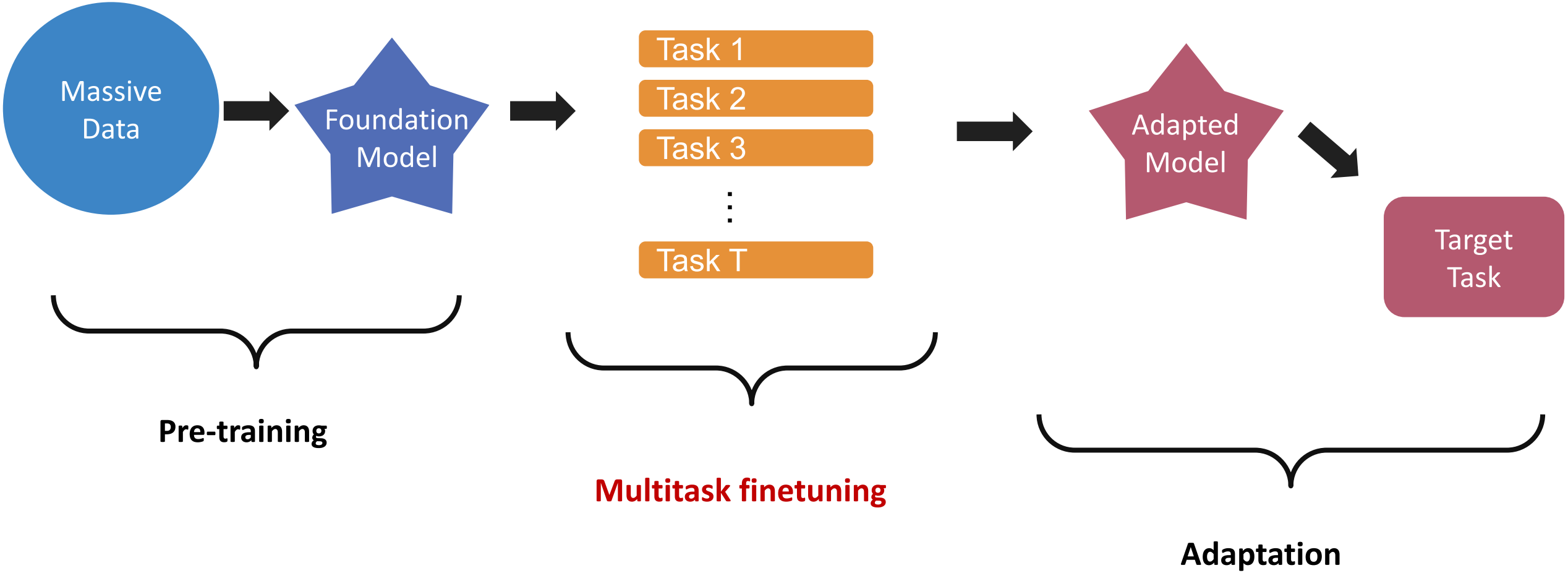
Challenges in Adaptation



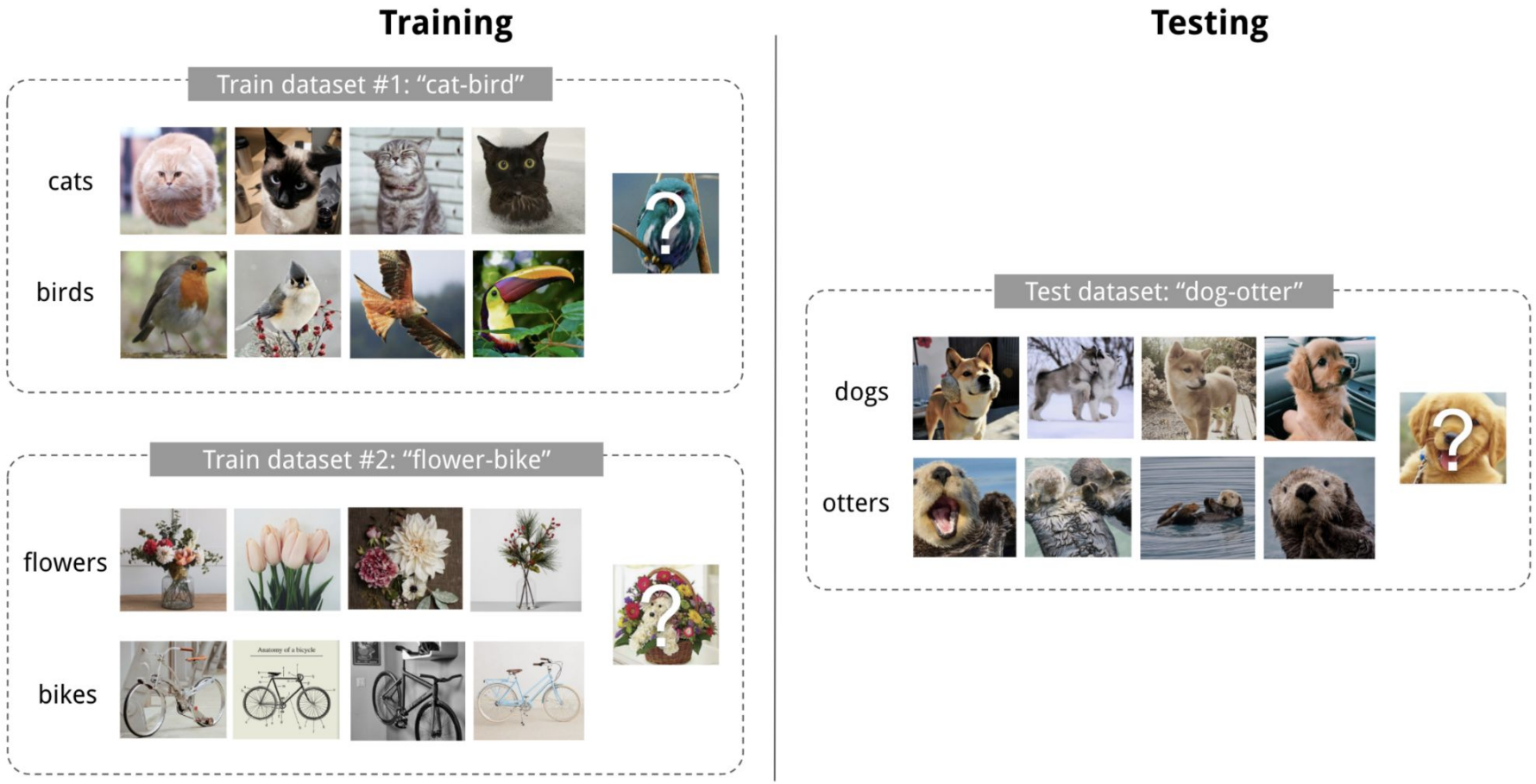
Data Efficiency

Fig Source: *https://www.youtube.com/watch?v=U6uFOIURcD0&ab_channel=ShusenWang, 2020*

Pre-training + Finetuning + Adaptation



Toy Example



An example of 4-shot 2-class image classification

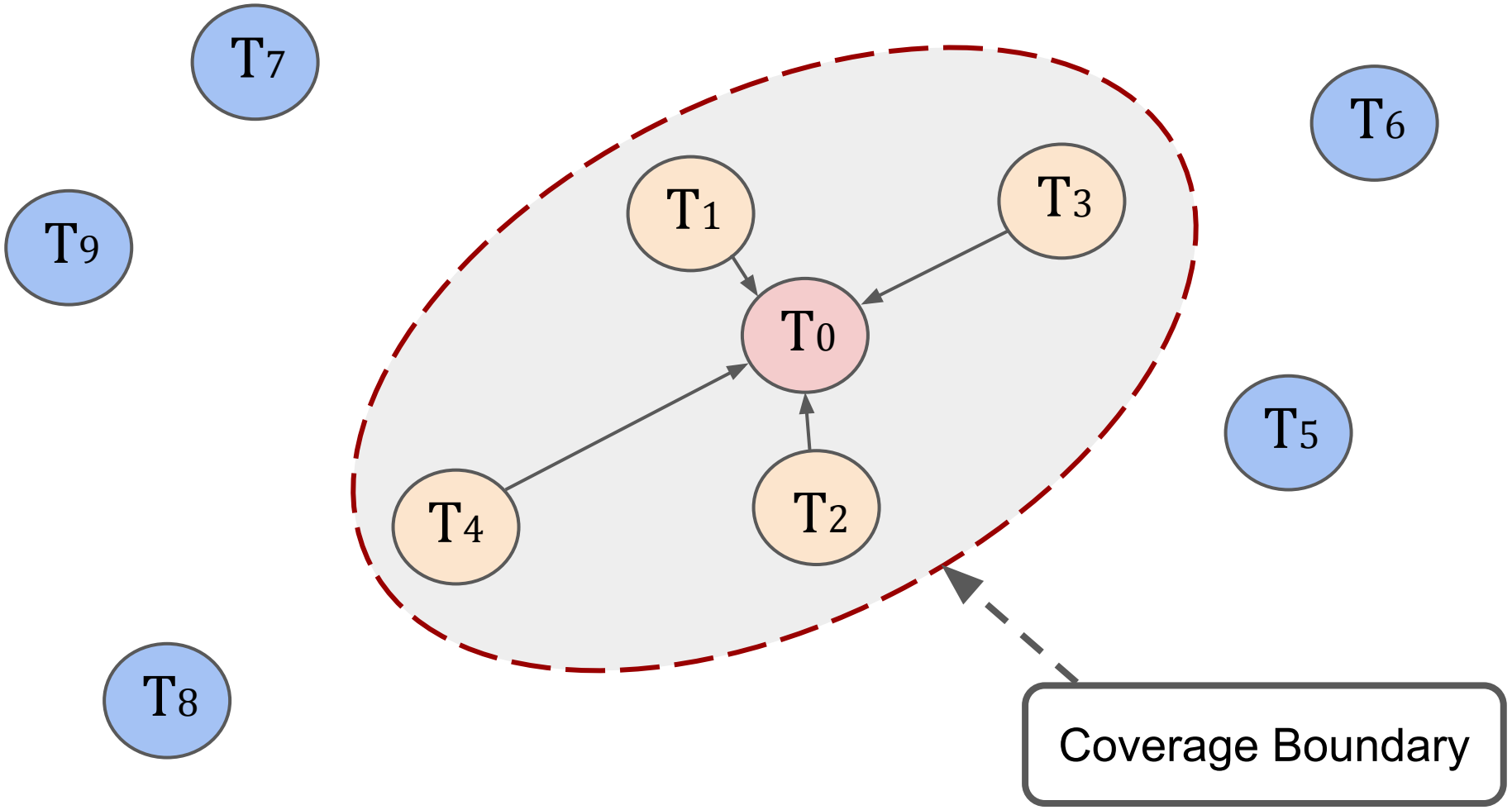
Fig source: *Meta-Learning: Learning to Learn Fast, 2018.*

Definition and Result

Definition 1 (Diversity and Consistency (Informal))
Consider the latent feature space of target task data and finetuning task data. **Consistency** refer to similarity in the feature space. **Diversity** refer to the coverage of the finetuning tasks on the target task in the latent feature space.

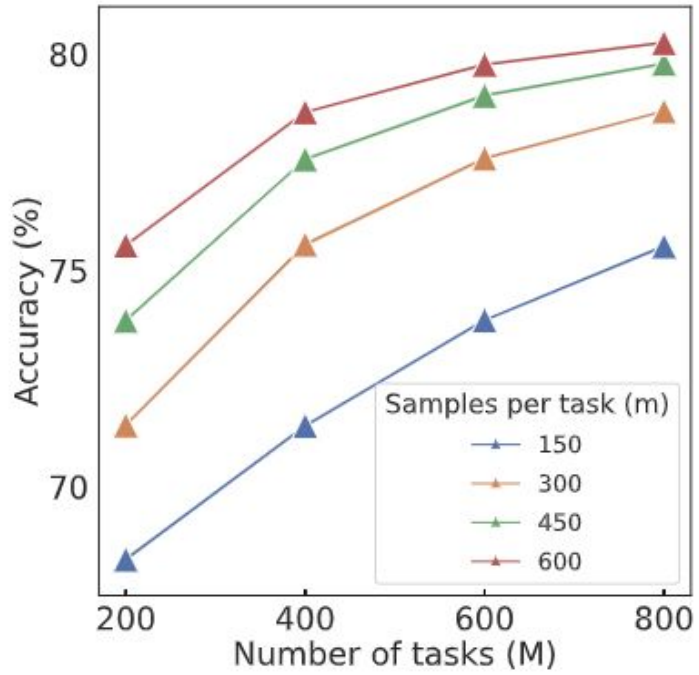
Theorem (Multitask finetuning (Informal))
Suppose pretraining achieves loss ϵ. If finetuning tasks satisfy diversity and consistency properties, then after sufficient multitask finetuning, the target task error reduces to $\alpha\epsilon$ with high probability, where the finetuning sample complexity is proportional to $1/\alpha\epsilon$.

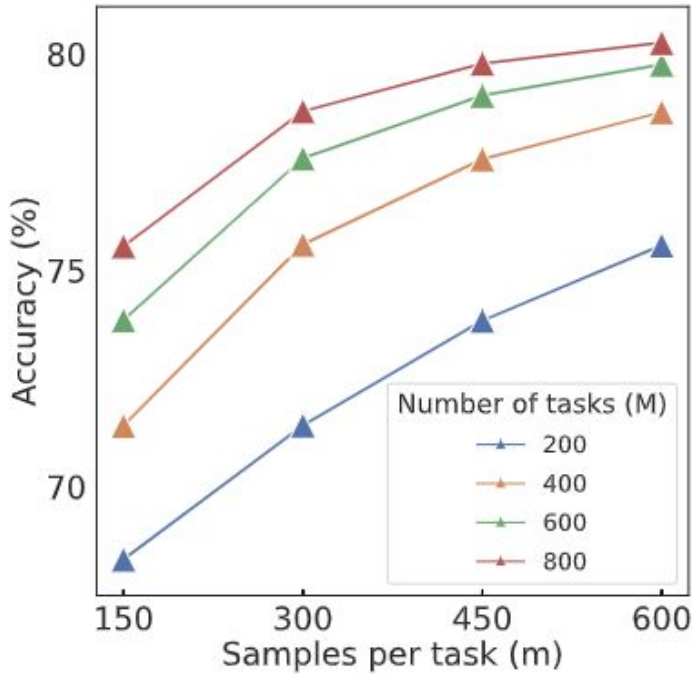
Practical Solution: Task Selection



Experiments: Verification of Theoretical Analysis

- **Models:** Pretrained Vision-Transformer (ViT).
- **Metrics**: Accuracy on few-shot target tasks with novel class.
- ***M***: Number of finetuning tasks. ***m***: Number of samples per task.





Multitask Finetuning for Adaptation

Developed a targeted adaptation framework that:

1. Identify and select relevant data matching target task characteristics.
2. Design specialized multitask finetuning pipeline.
3. Achieve strong performance with limited target data.

Xu, Shi, Wei, Mu, Li, Liang. Towards Few-Shot Adaptation of Foundation Models via Multitask Finetuning. ICLR'24.

Thesis Outline

How can we bridge this specialization gap?

1. **Data-efficient Adaptation:** Enables FMs to better specialize in tasks in different domains (Part 1)
2. **Compositional Ability:** Advances FMs' ability to handle complex problems by combining simple tasks (Part 2)
3. **Adaptive Inference:** Make FMs deployable under varying computational constraints (Part 3)

In-Context Learning (ICL)

Circulation revenue has increased by 5% in Finland. // Positive

Panostaja did not disclose the purchase price. // Neutral

Paying off the national debt will be extremely painful. // Negative

The company anticipated its operating profit to improve. // _____

Circulation revenue has increased by 5% in Finland. // Finance

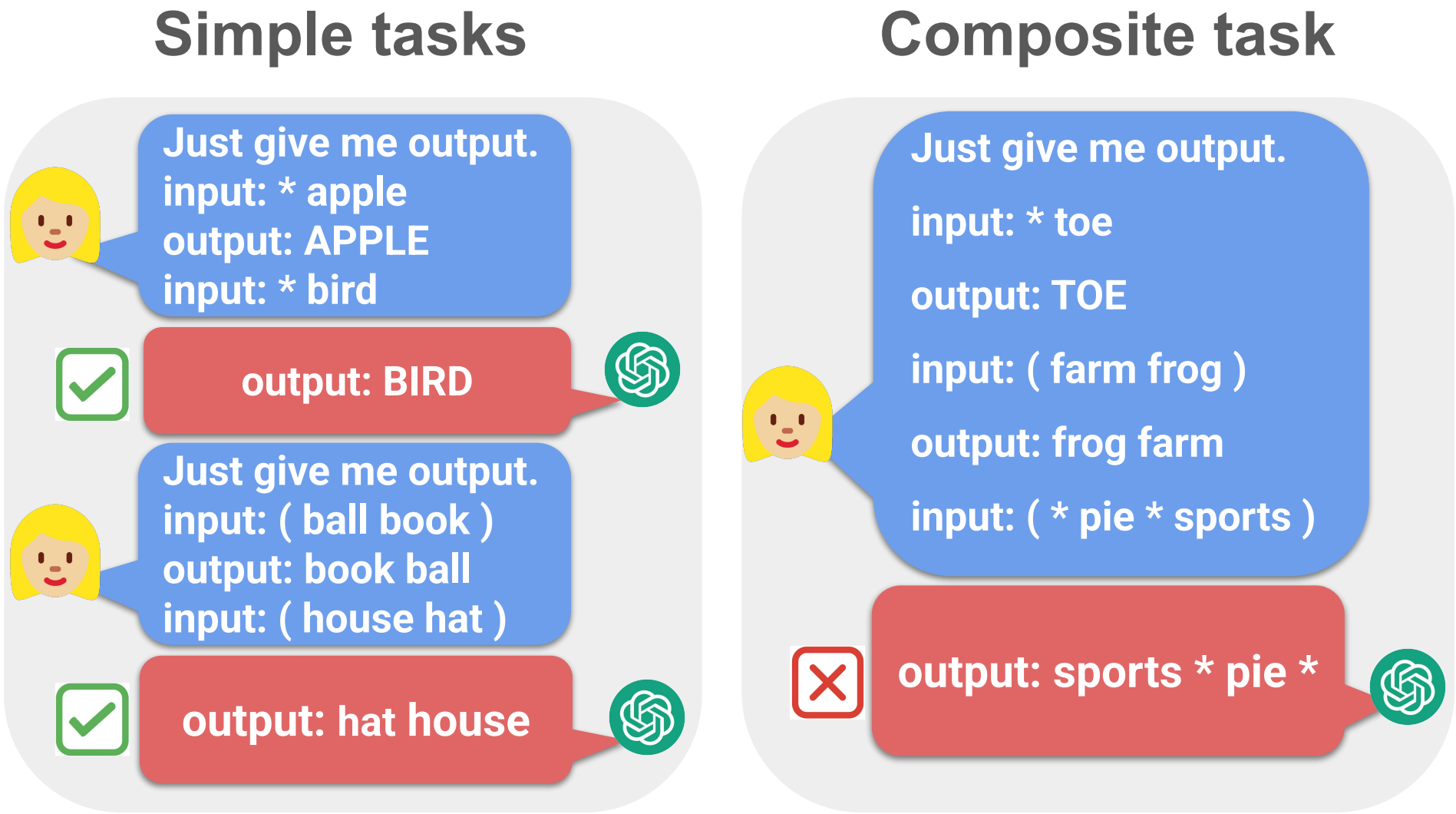
They defeated … in the NFC Championship Game. // Sports

Apple … development of in-house chips. // Tech

The company anticipated its operating profit to improve. // _____

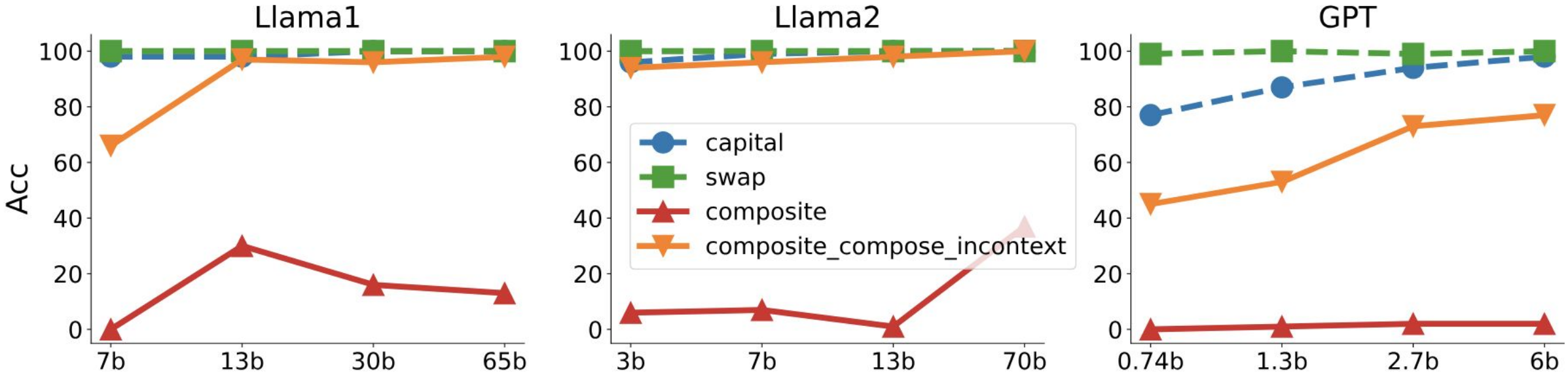
Fig source: *How does in-context learning work? A framework for understanding the differences from traditional supervised learning, 2022.*

Motivation


Simple tasks
Just give me output.
input: * apple
output: APPLE
input: * bird
output: BIRD
Just give me output.
input: (ball book)
output: book ball
input: (house hat)
output: hat house
Composite task
Just give me output.
input: * toe
output: TOE
input: (farm frog)
output: frog farm
input: (* pie * sports)
output: sports * pie *

A Failure Case for Composition

	Composite	Composite in-context
Prompt	*input: * apple* *output: APPLE* *input: (farm frog)* *output: frog farm* *input: (* bell * ford)*	*input: (* good * zebra)* *output: ZEBRA GOOD* *input: (* bicycle * add)*
Truth	*output: FORD BELL*	*output: ADD BICYCLE*



Design Experiments to investigate

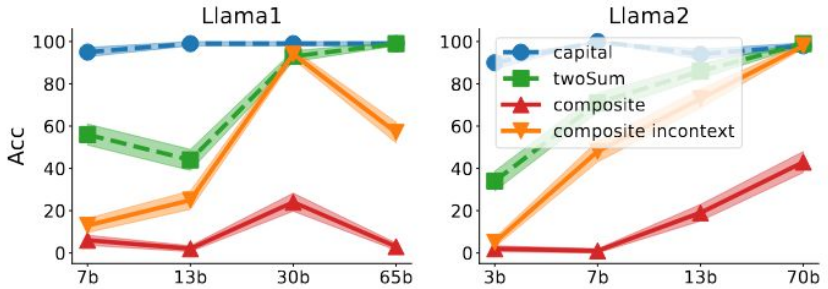
- How do LLMs perform in various tasks?
- Does scaling up the model help in general?
- Is the variability in performance relevant to the nature of tasks?

Compositional Logical Tasks

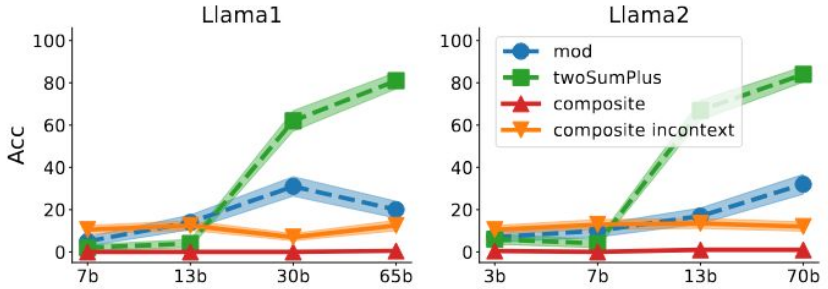
Tasks	Simple Task	Simple Task	Composite
(A) + (B)	input: * apple output: APPLE	input: (farm frog) output: frog farm	input: (* bell * ford) output: FORD BELL
(A) + (C)	input: * (*five*) output: FIVE	input: *twenty @ eleven* output: thirty-one	input: * (*thirty-seven @ sixteen*) output: FIFTY-THREE
(G) + (H)	input: 15 @ 6 output: 3	input: 12 # 5 output: 18	input: 8 # 9 @ 7 Ouput: 4
(A) + (F)	input: 435 output: 436	input: cow output: COW	input: 684 cat output: 685 CAT

Tasks	Simple Task	Simple Task	Composite
(A) + (B)	input: * apple output: APPLE	input: (farm frog) output: frog farm	input: (* bell * ford) output: FORD BELL
(A) + (C)	input: * (*five*) output: FIVE	input: *twenty @ eleven* output: thirty-one	input: * (*thirty-seven @ sixteen*) output: FIFTY-THREE
(G) + (H)	input: 15 @ 6 output: 3	input: 12 # 5 output: 18	input: 8 # 9 @ 7 Ouput: 4
(A) + (F)	input: 435 output: 436	input: cow output: COW	input: 684 cat output: 685 CAT

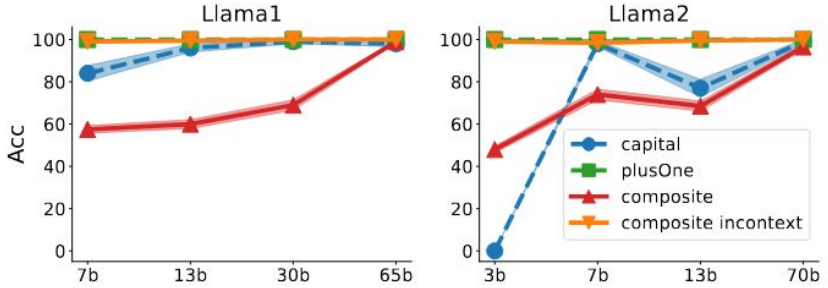
(A) + (C)



(G) + (H)



(A) + (F)

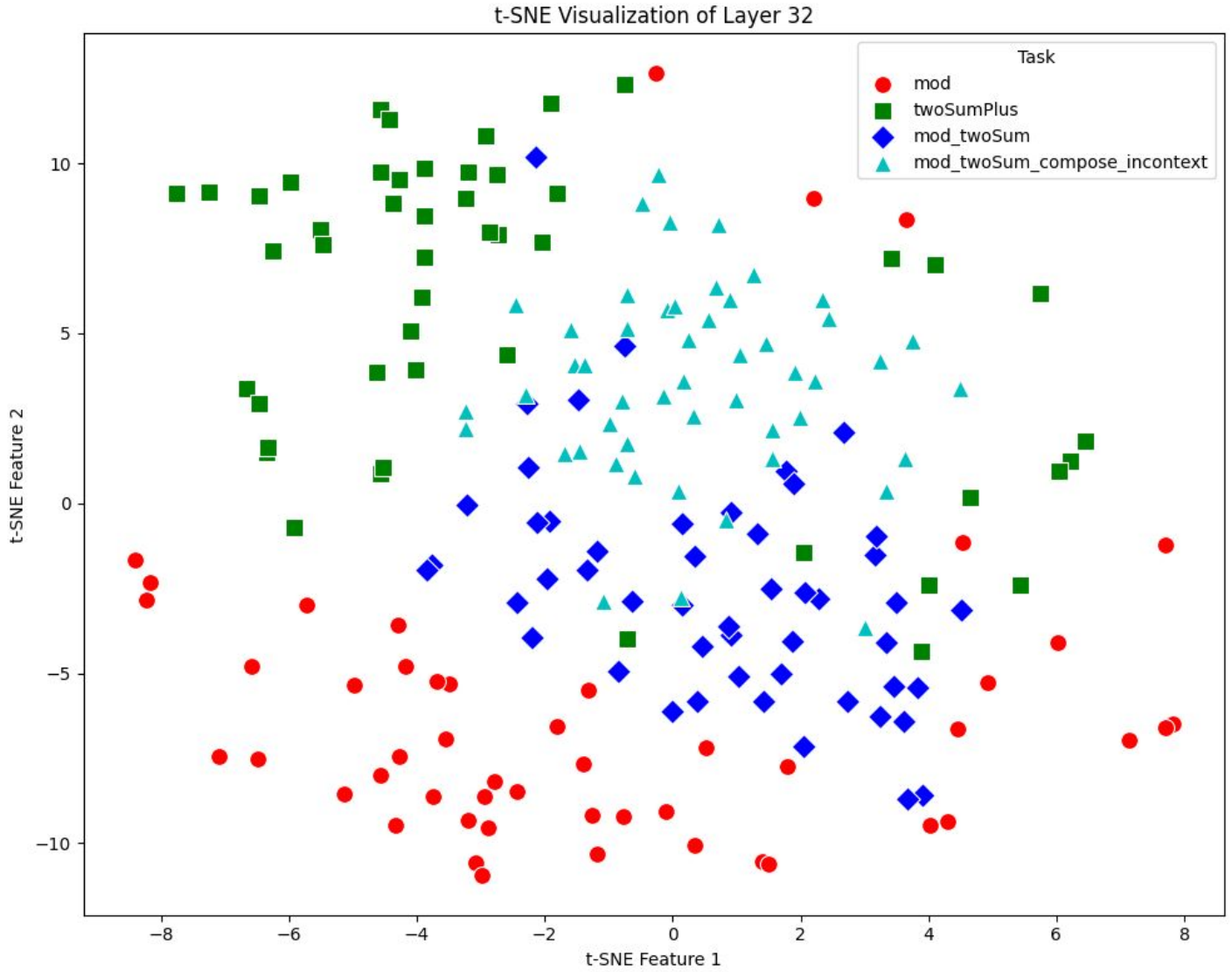


Design Experiments to investigate

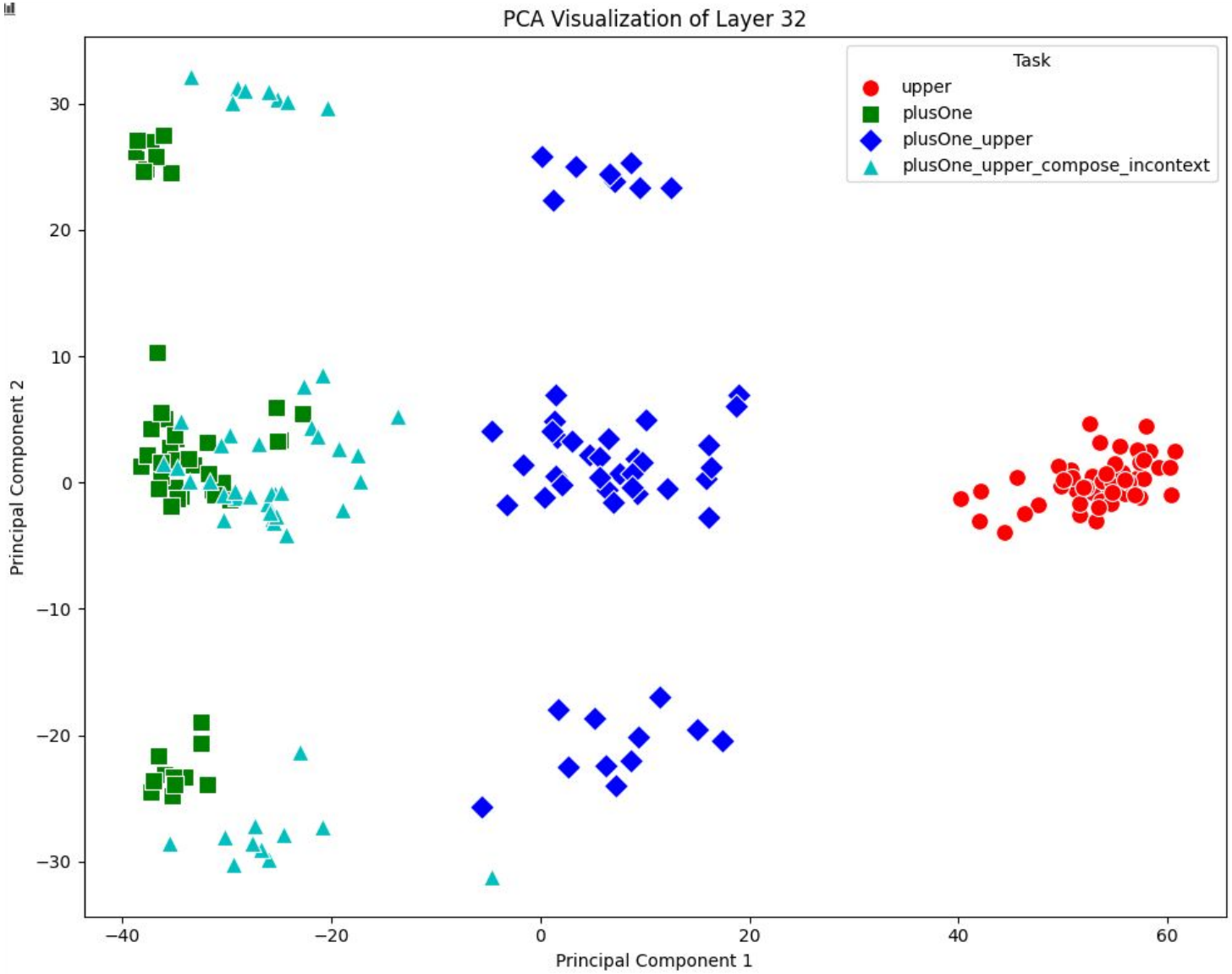
- How do LLMs perform in various tasks?
  - Performance varies significantly by task type.

- Does scaling up the model help in general?
  - Scaling-up helps when the model exhibits compositional ability for certain tasks but not help when the model initially struggles.

- Is the variability in performance relevant to the nature of tasks?



(G) + (H)	input: 15 @ 6 output: 3	input: 12 # 5 output: 18	input: 8 # 9 @ 7 Ouput: 4



(A) + (F)	input: 435 output: 436	input: cow output: COW	input: 684 cat output: 685 CAT

Compositional Ability

Definition (Compositional Ability)
For a composite task (A)+(B), a model has compositional ability if in-context examples from both (A) and (B) yield higher accuracy than examples from either task alone.

Theorem (Compositional ability under confined support (Informal))

Consider input embedding $x \in \mathbb{R}^d$ of each simple tasks. Consider each sample has a disjoint subset of indices from $1, 2, \ldots, d$. Each simple task only has large values within its corresponding subsets of dimensions of input embeddings. Then with high probability, the model has the compositional ability.

ICL for compositional

Our findings on compositional ability in LLMs reveal:

1. Simple composition (distinct mappings on different inputs): Models perform well and benefit from scaling.

2. Complex composition (multi-step reasoning): Models struggle, with limited gains from scaling.

Xu*, Shi*, Liang. Do Large Language Models Have Compositional Ability? An Investigation into Limitations and Scalability. COLM'24.

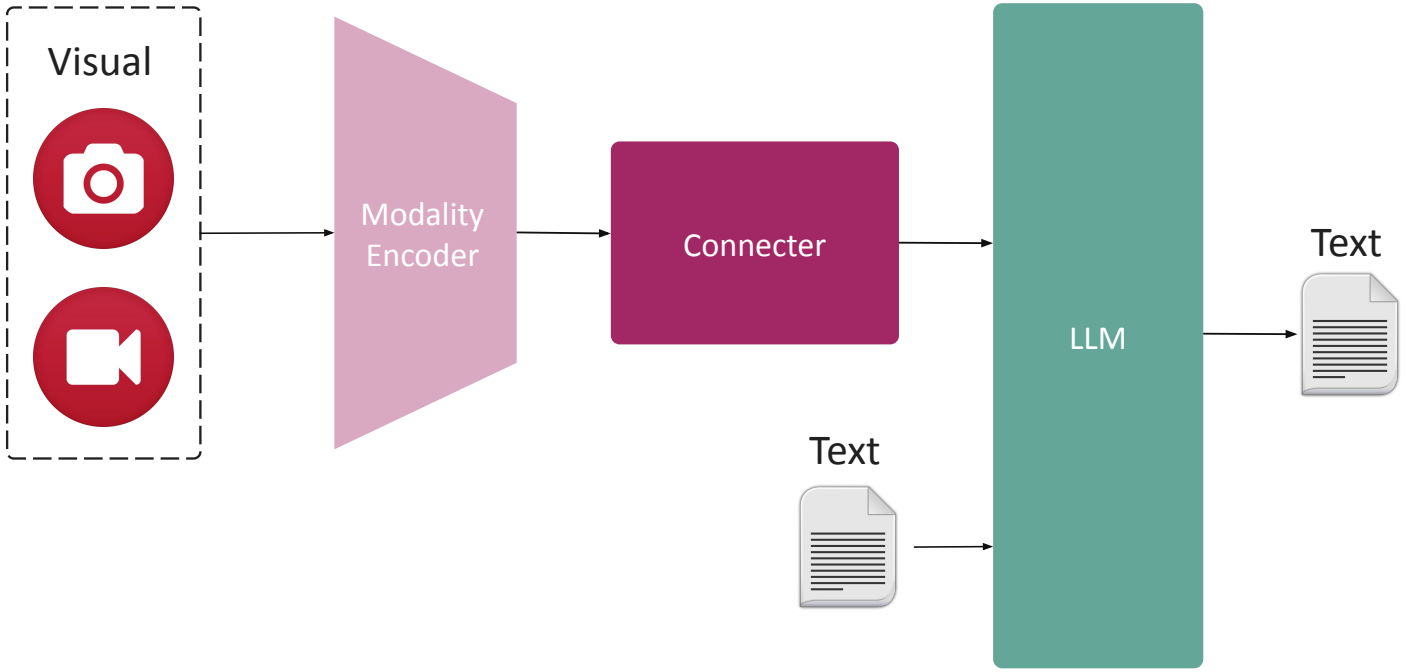
Thesis Outline

How can we bridge this specialization gap?

1. **Data-efficient Adaptation:** Enables FMs to better specialize in tasks in different domains (Part 1)
2. **Compositional Ability:** Advances FMs' ability to handle complex problems by combining simple tasks (Part 2)
3. **Adaptive Inference:** Make FMs deployable under varying computational constraints (Part 3)

Multimodal LLMs

Image Captioning | Visual QA | Video Summarization | •••



Visual Inference with Fixed Compute Footprint

Q: Where is lady liberty in this image?

User A w. 100% compute budget

100% compute budget

MLLM

A: Lady Liberty is lying on the frozen lake, toward the right side of the image.

Fig source: *UW Madison News (https://news.wisc.edu/its-back-40-years-later-lady-liberty-on-mendota/).*

Visual Inference with Fixed Compute Footprint

Fig source: *ChatGPT.*

Q: Where is lady liberty in this image?

User A w. 100% compute budget

Can we use less compute?

100% compute budget

MLLM

A: Lady Liberty is centered in the image. The background is clear blue sky, so there are no surrounding landmarks visible.

Visual Inference with Fixed Compute Footprint

Q: Where is lady liberty in this image?

User B w. **50% compute budget** due to
- compute contention
- tighter latency requirement

Cost Violation

100% compute budget

MLLM

A: Lady Liberty is lying on the frozen lake, toward the right side of the image.

Fig source: *UW Madison News (https://news.wisc.edu/its-back-40-years-later-lady-liberty-on-mendota/).*

Challenges

Conventional MLLM lack **adaptivity** such that:

- Cannot reduce computation for “simpler” inputs
- Stick to compute / energy budget despite varying available resources

Related Approaches Fall Short

1. Model Compression[1]
2. Token Selection[2]
3. Mixture of Experts[3]

Adhere to varying latency and compute

Solution: AdaLLaVA: Latency-aware adaptive inference for MLLMs.

[1] Yao et al. Minicpm-v: A GPT4V level MLLM on your phone. Technical Report. 2024.
[2] Liang et al. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. ECCV, 2024
[3] Lin et al. Moe-llava: Mixture of experts for large vision-language models. PMM 2025.

Adaptive Inference w.r.t Runtime Conditions

Input image

What is the image showing?

LLaVA

The image shows a snowman holding a colorful egg. — 8 TFLOPs

60% compute budget

AdaLLaVA

The image is showing a painting or drawing of a snowy winter scene. — **4.8 TFLOPs**

85% compute budget

The image shows a snowman holding a colorful ball. — **6.8 TFLOPs**

100% compute budget

The image shows a snowman holding a colorful egg. — **8 TFLOPs**

MLLM as a collection of shallower models

Observation: MLLMs have redundant capacity.

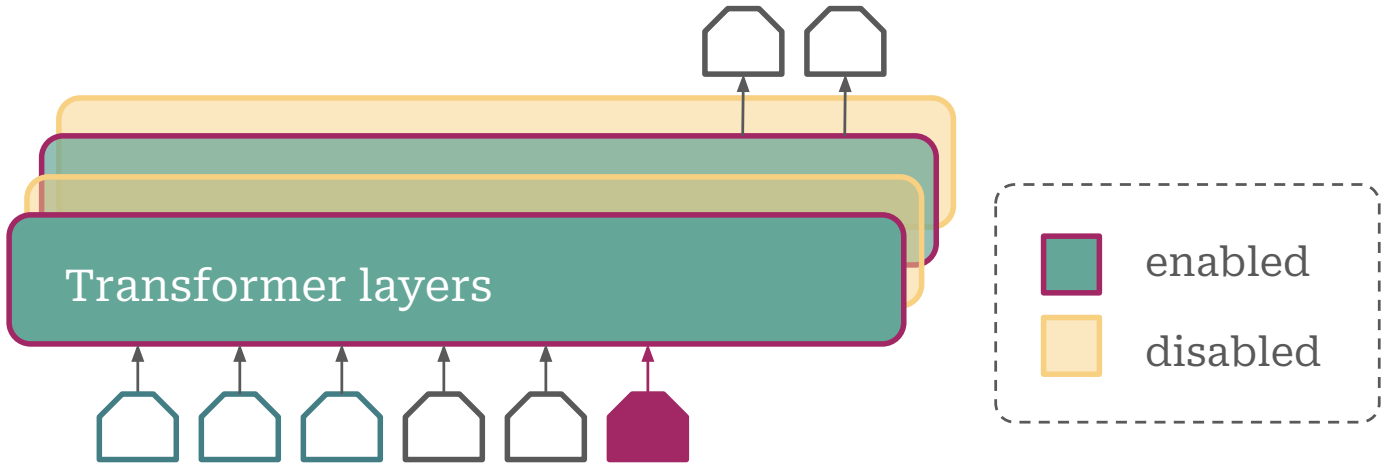
Evidence: Prior works[1,2,3] show many Transformer layers/heads can be skipped with minimal accuracy loss.

Opportunity: We can exploit this redundancy to dynamically reconfigure models at runtime, achieving diverse accuracy-latency trade-offs.

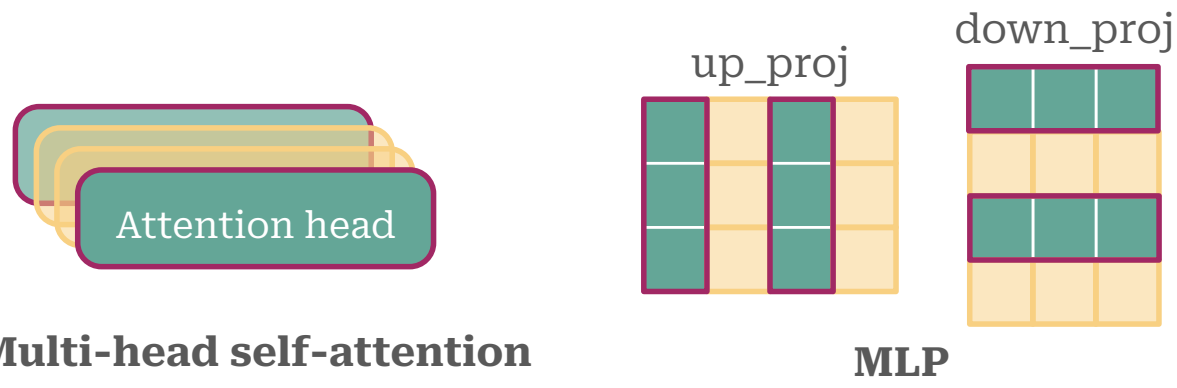
[1] Huang et al. Deep networks with stochastic depth. ECCV, 2016.
[2] Touvron et al. Training data-efficient image transformers & distillation through attention. PMLR 2021.
[3] Wang et al. Going deeper with image transformers. ICCV 2021.

Design of Reconfiguration

AdaLLaVA-L



AdaLLaVA-H



Problem Setup

Conventional MLLM

LLM

visual/text tokens

Answer

$$f_\theta\left(\{\mathbf{z}^v\},\{\mathbf{z}^q\}\right)\rightarrow\mathbf{x}^a$$

Adaptive MLLM

Scheduler

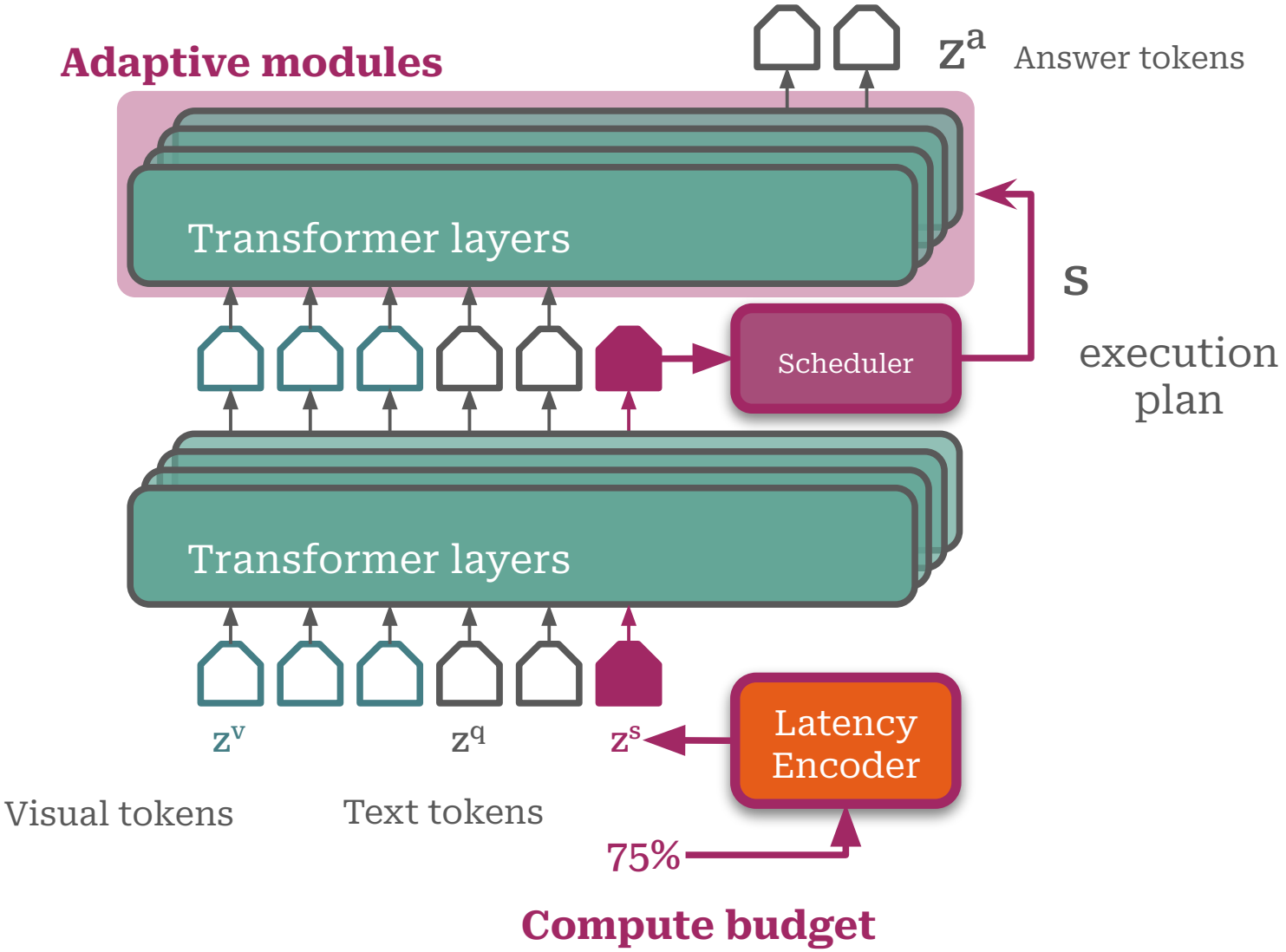
compute budget

$$g_\phi\left(\{\mathbf{z}^v\},\{\mathbf{z}^q\},l\right)\rightarrow\mathbf{s}$$

$$f_\theta\left(\{\mathbf{z}^v\},\{\mathbf{z}^q\};\mathbf{s}\right)\rightarrow\mathbf{x}^a$$

Execution plan

AdaLLaVA Framework

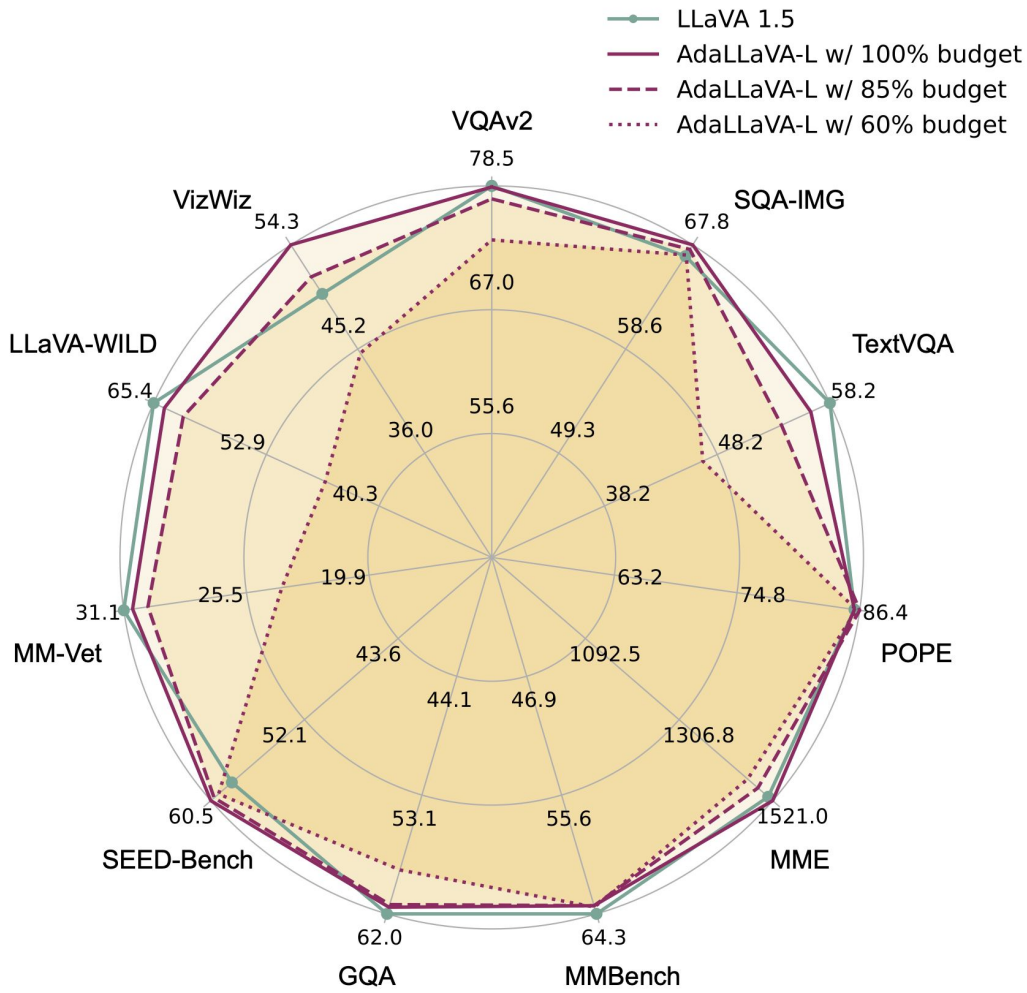


Experimental Setup

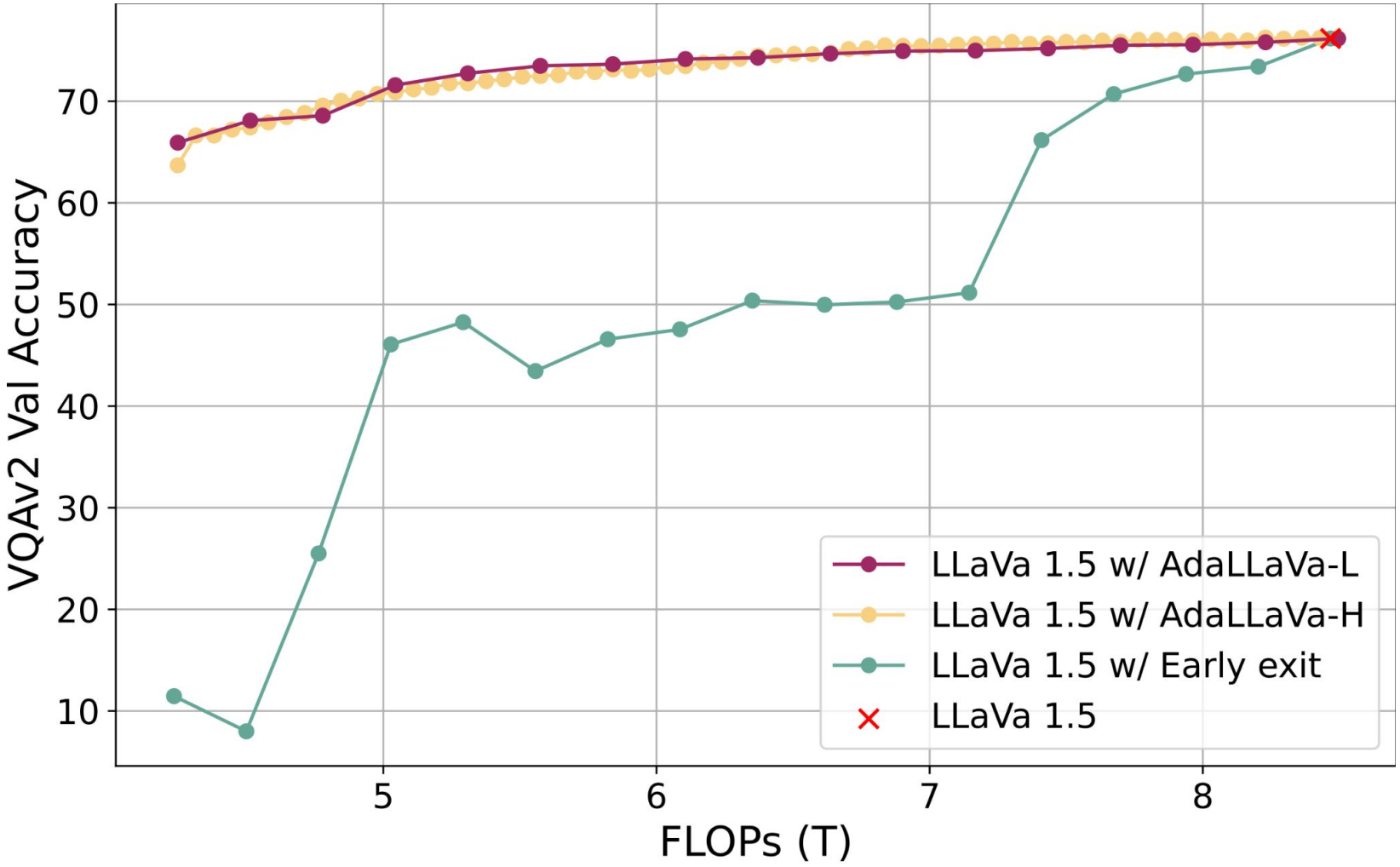
- **Models:** LLaVA 1.5 (7B and 13B), Mipha-3B
- **Training**: Supervised FT with visual instruction data (665K)
- **Visual QA benchmarks**: VQAv2, ScienceQA, TextVQA, POPE…
- **Metrics**: Accuracy under different compute budget, compute adherence

Main Results

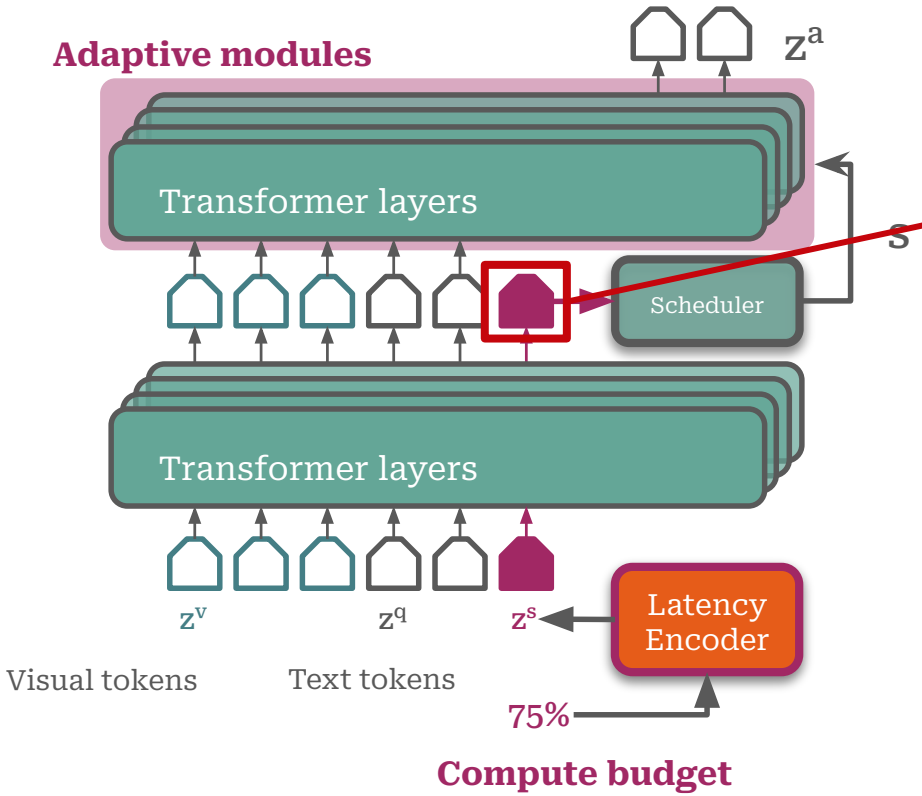
- Baseline: LLaVA 1.5
- Performance:
  Same AdaLLaVA model at 60%, 85%, and 100% compute budgets
- Metrics: 11 VQA benchmarks
- Budget Level: 0 violatation



Adaptivity Analysis: Compute Adaptivity



Adaptivity Analysis: Content Adaptivity

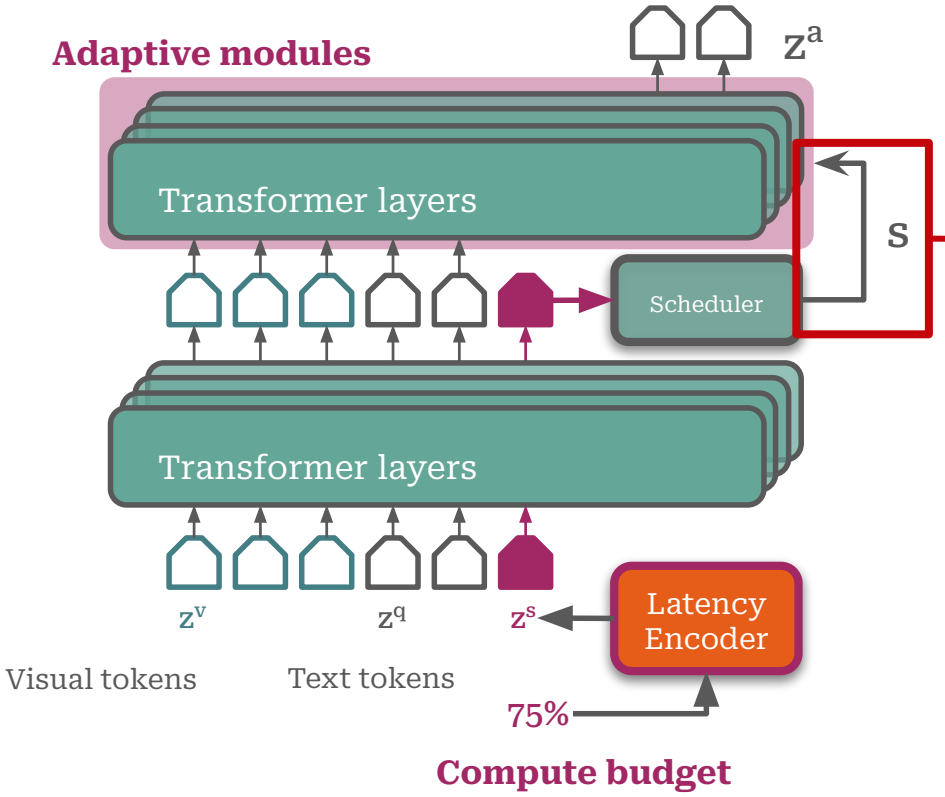


Question:	What is the title of this movie?	Who is the main actor?
Answer:	Yes Man	The name of the main actor in the movie is Jim Carrey.

Question:	What is the name of the main actor?	What activity are they doing?
Answer:	The name of the main actor is Ryan Gosling.	The man and woman are sitting in a boat, likely rowing or paddling it…

Latency token attention mapping

Adaptivity Analysis: Content Adaptivity



Question: Does this artwork exist in the form of painting? Please answer yes or no.

Answer: Yes.

Execution plan:

FLOPs: 6.4T

Question: Does this artwork belong to the type of religious? Please answer yes or no.

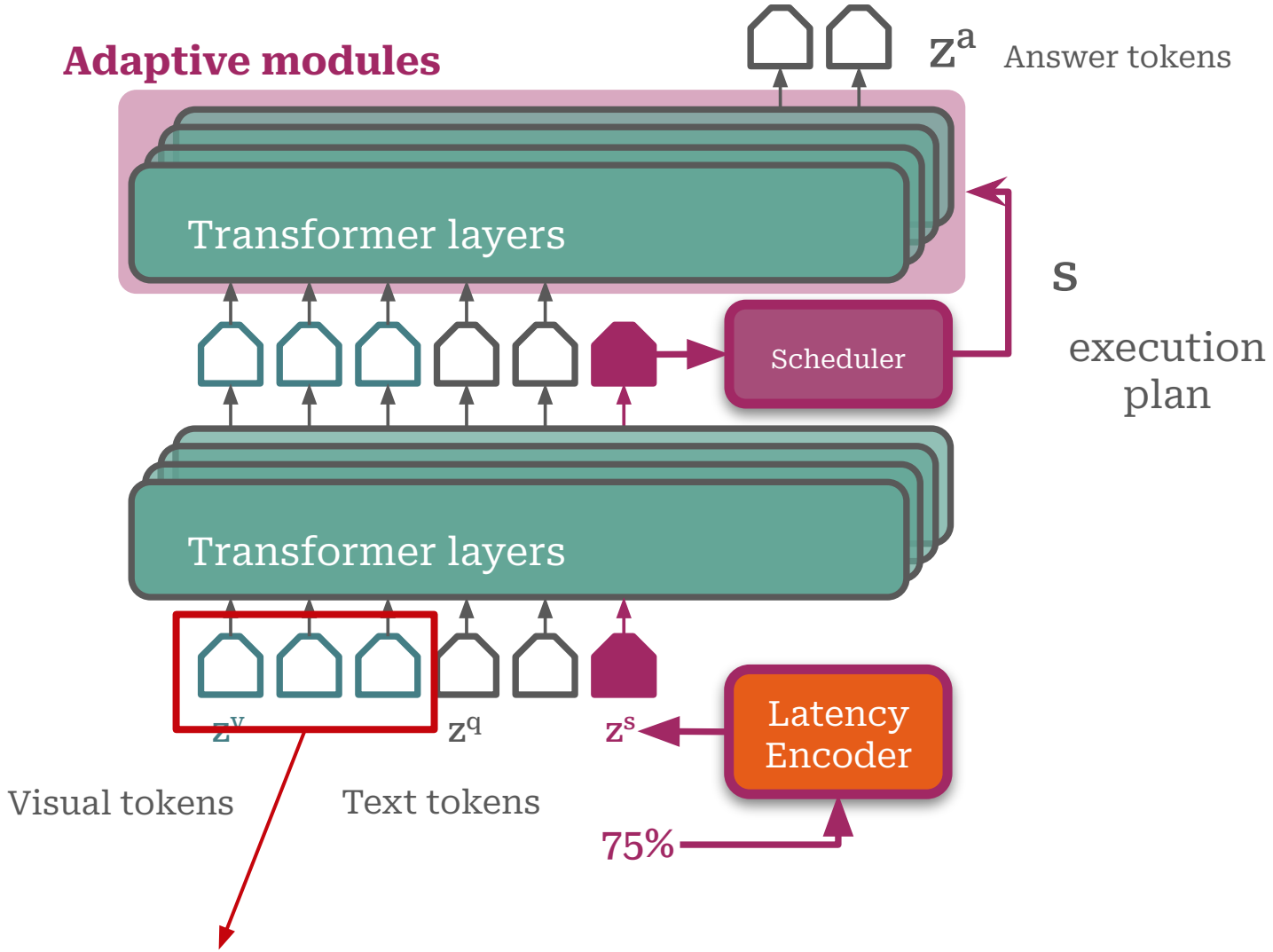
Answer: Yes.

Execution plan:

FLOPs: 6.4T

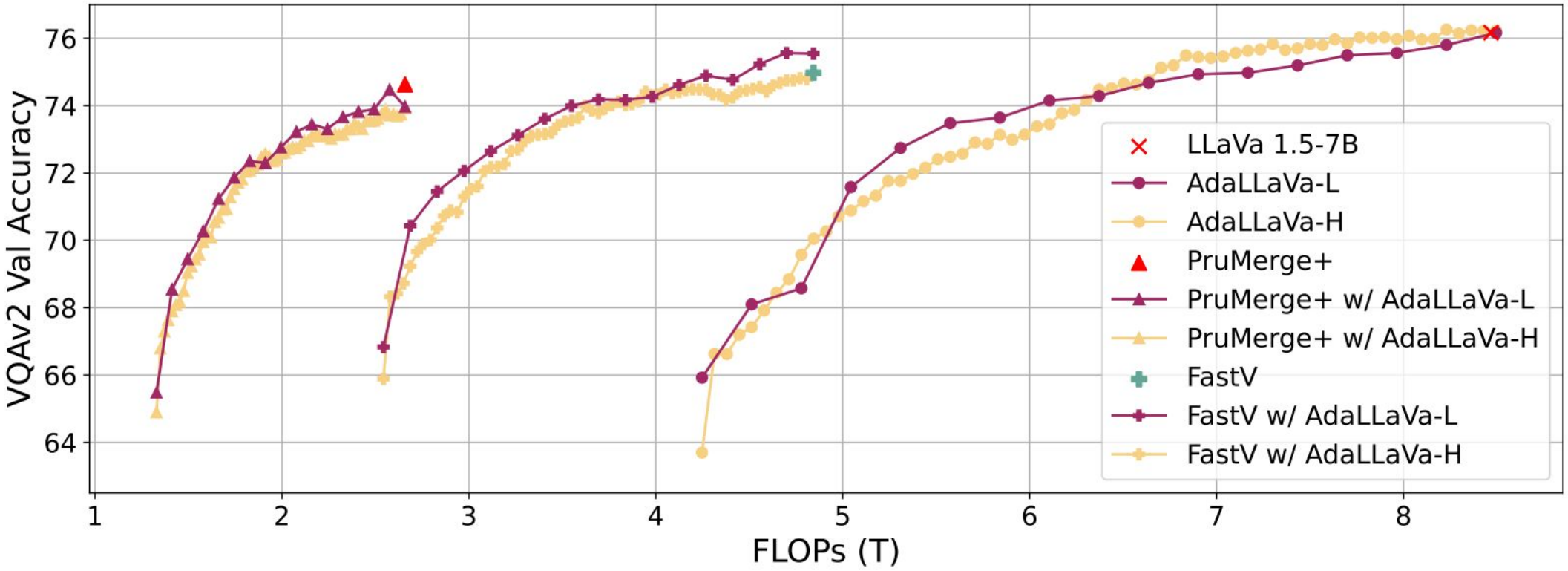
Execution plan under different content

Visual Token selection



- Redundant vision tokens: vision token selection

Adaptivity Analysis: Compute Adaptivity



[1] **Prumerge+**: Shang et al. LLaVA-PruMerge: Adaptive Token Reduction for Efficient Large Multimodal Models. ICCV, 2025.
[2] **FastV**: Chen et al. An Image is Worth 1/2 Tokens After Layer 2: Plug-and-Play Inference Acceleration for Large Vision-Language Models. ECCV 2024.

Adaptive Inference

Our Contribution:

1. First adaptive inference framework for MLLMs.
2. Novel scheduler with compute awareness.
3. Strong empirical results through extensive validation across benchmarks and models.

Xu*, Nguyen*, Mukherjee, Bagchi, Chaterji, Liang, Li. Learning to Inference Adaptively for Multimodal Large Language Models. ICCV'25.

Conclusion

Thesis Statement

Foundation models can be effectively specialized for downstream tasks through

(1) data-centric multitask adaptation (Part 1),

(2) understanding of compositional reasoning mechanisms (Part 2),

(3) dynamic inference strategies (Part 3).

More Contributions

Foundation Model Understanding

- Why Larger Language Models Do In-context Learning Differently? [SW**X**L, ICML'24]
- Out-of-distribution generalization via composition: a lens through induction heads in Transformers. [S**X**Z, PNAS'25]
- Can Language Models Compose Skills In-Context? [L**X**SL, submit to ICLR'26]

Retrieval-augmented generation (RAG)

- TabRAG: Efficient Table Retrieval and Understanding with Multimodal Large Language Models [**X**FHMWZ, AWS Internship Work'24]
- Augmenting Agent Memory With Temporal GraphRAG [**X**DP, AWS Internship Work'25]

Transformer Acceleration

- Conv-Basis: A New Paradigm for Efficient Attention Inference and Gradient Computation in Transformers [*LLSS**X**Y, EMNLP'25 Findings]

Statistical Genomics

- Spatial Transcriptomics Dimensionality Reduction using Wavelet Bases [**X**K, F1000'22]

Acknowledgement
















WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON