# In-Context Meta Learning Induces Multi-Phase Circuit Emergence

松尾・岩澤研究室
MATSUO-IWASAWA LAB UTOKYO

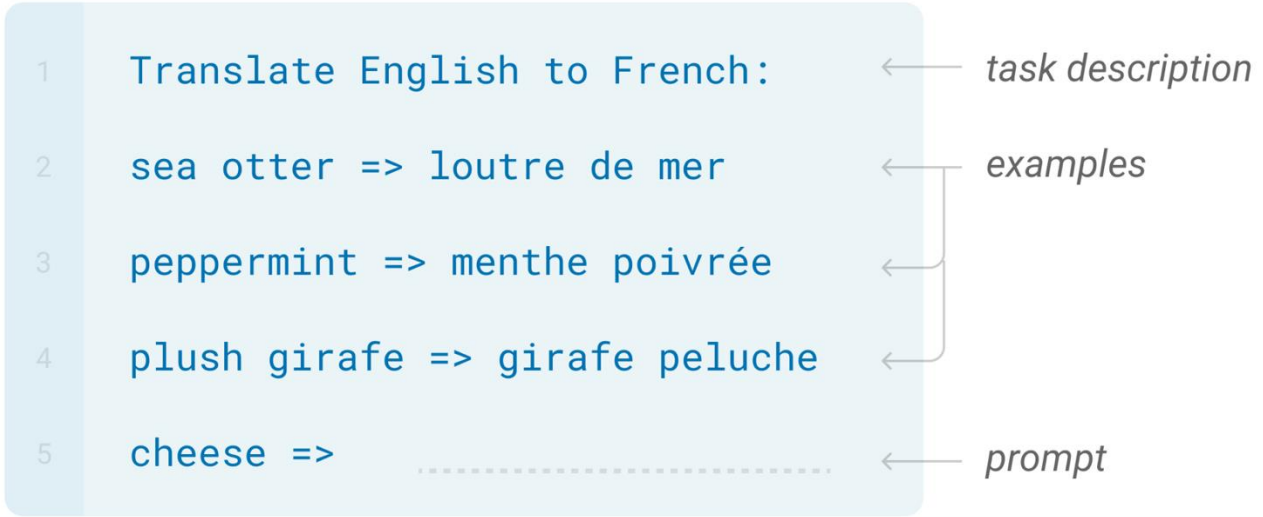
**Gouki Minegishi**

paper

UTokyo

# In Context Learning (ICL)

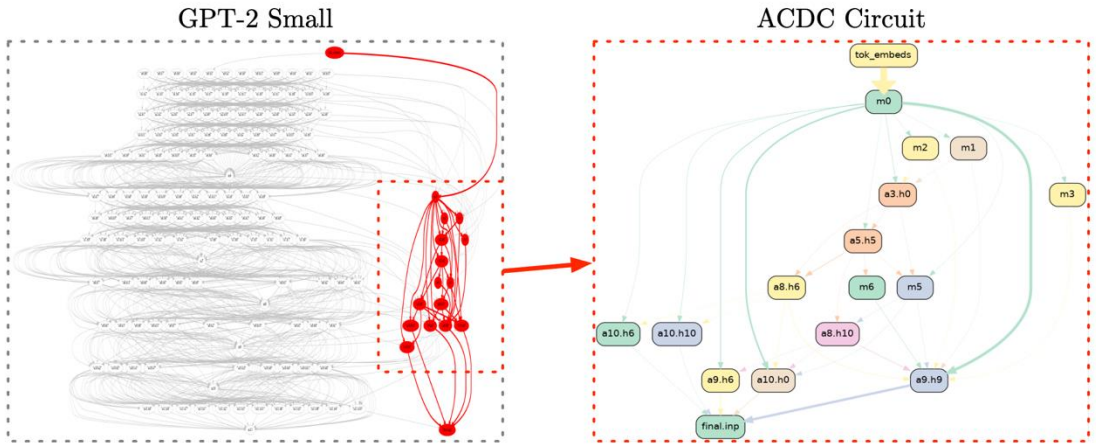
- *In-context learning* means **inferring the task** from the prompt examples and predicting the answer accordingly.
- Why large language models can perform in-context learning so effectively remains an open research question.

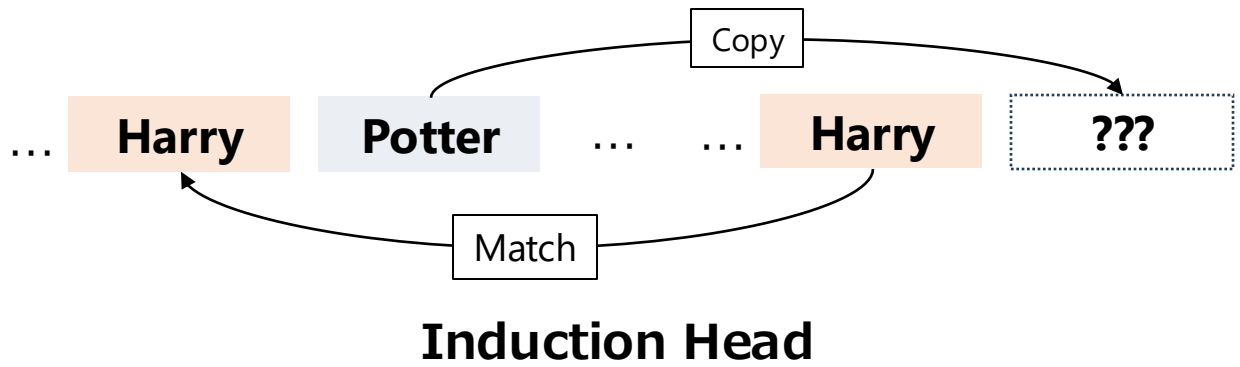



Brown et al., "Language Models are Few-Shot Learners"

# Mechanistic Interpretability

- Reverse-engineering methods that **causally** uncover a model's internal computation, beyond surface I/O or attribution analyses
- **Circuits**: functional units inside the network that implements a specific capability
  - **Induction Head** – the canonical *in-context-learning* circuit
    - Operation: locate a matching token earlier in the prompt and copy the *next* token, letting the model predict the correct continuation




**Induction Head**

Conmy et al., "Towards Automated Circuit Discovery for Mechanistic Interpretability"

# Research Questions

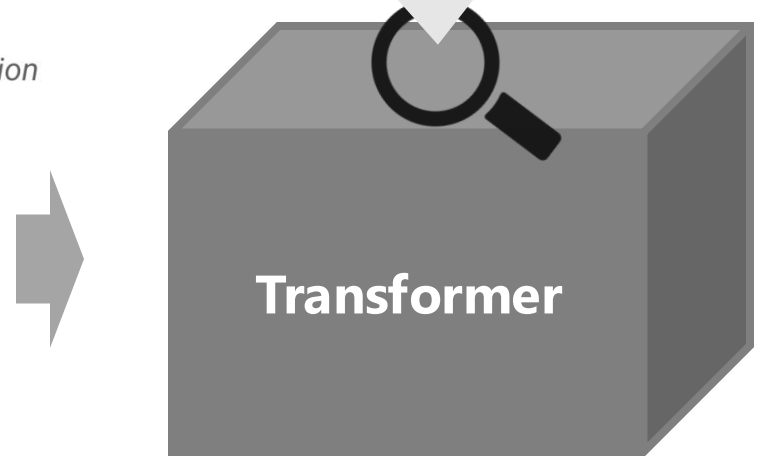
Induction heads is the simple match-and-copy circuit.
We still don't know what kind of circuits emerge **in real few-shot prompts**.

What kind of circuit works?

How does the circuit grow?

**Few-shots prompts**

```
1  Translate English to French:        ← task description
2  sea otter => loutre de mer          ← examples
3  peppermint => menthe poivrée        ← examples
4  plush girafe => girafe peluche      ← examples
5  cheese =>                           ← prompt
```



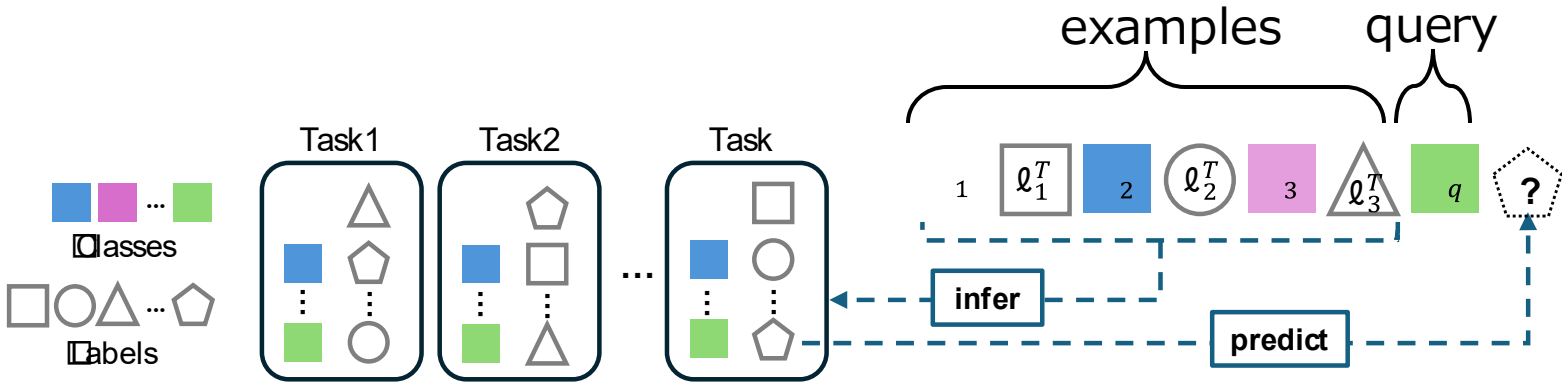

**Prediction**

fromage

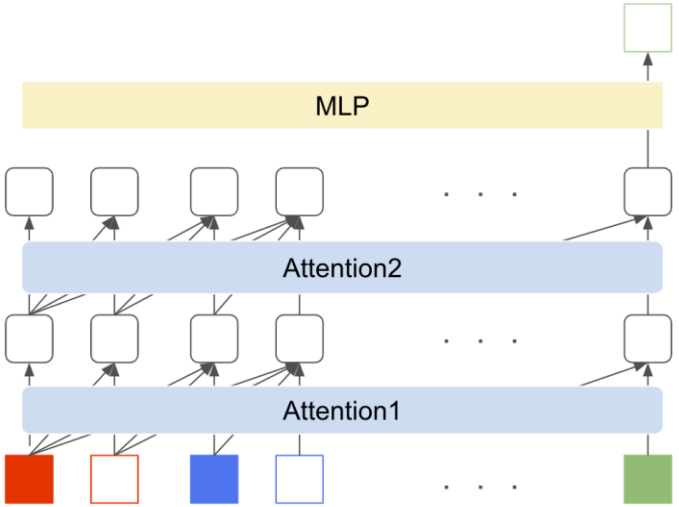
# Toy Experiment Setup

- Problem Setup
  - We designed a task involving 64 classes and 32 labels, with context-specific class–label pairings.
  - To answer accurately, the model must infer the underlying task from the context.
- Network Structure
  - 2-layer attention only Transformer + 1-layer MLP (classification)

**Problem setup**



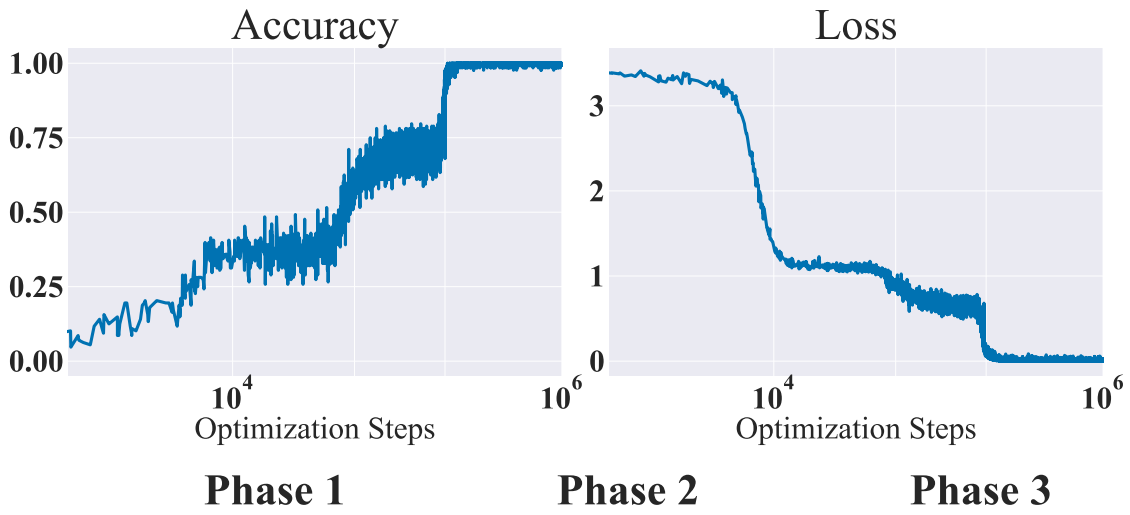

**Network structure**

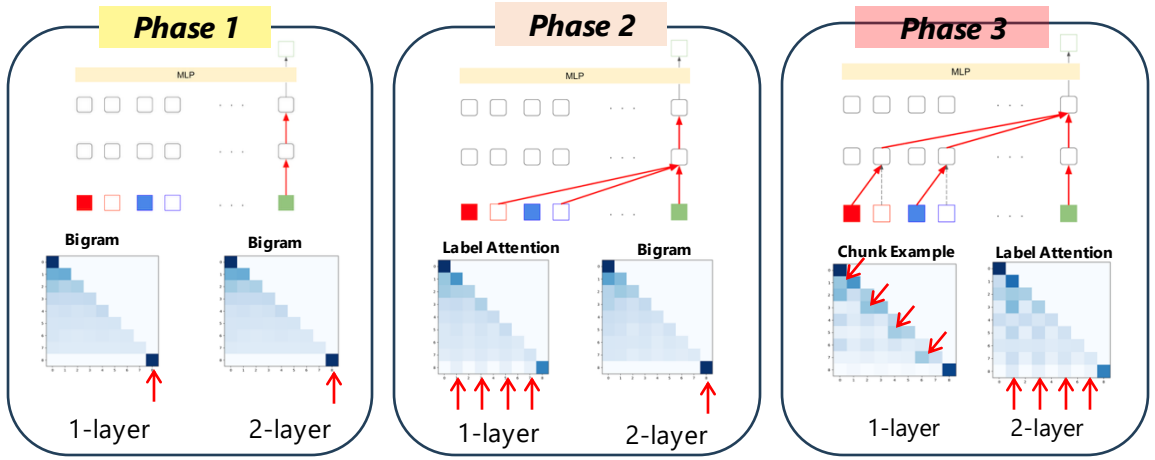


# 3-Phase Transition and 3-Circuit Emergence

- The model arrives at 100% undergoing three accuracy plateaus
- 3-Circuits at each plateau
  - **Non-Context Circuit; NCC (Phase1)**:
    The model Ignore the context and relying solely on the model's weights.
  - **Semi-Context Circuit; SCC (Phase2):**
    The model not only leverages weights memory but also attends to label tokens (i.e., half of the context)
  - **Full-Context Circuit; FCC (Phase3):**
    The model use the entire context.

| Circuit | Accuracy ($T = 3$) | Layer 1 | Layer 2 |
|---|---|---|---|
| NCC | 30–40% | Bigram | Bigram |
| SCC | ≈75% | Label Attention | Bigram |
| FCC | 100% | Chunk Example | Label Attention |



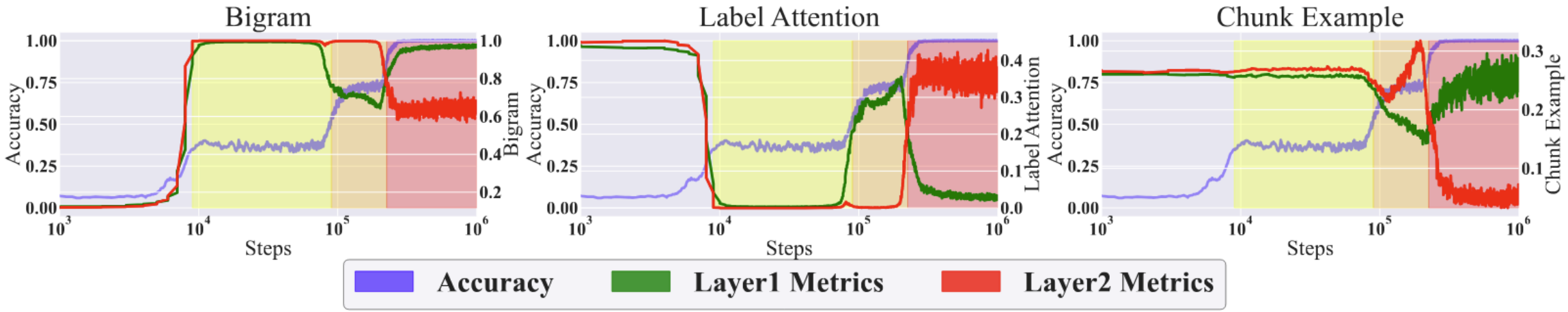
# Circuit Metrics

1. **Bigram Metric:** Attention from a query token to itself.
2. **Label Attention Metric:** Total attention from the query to all label tokens in context.
3. **Chunk Example Metric:** Attention from each example token $x$ to its paired label $\ell$.

attention weights $p_{i,j}^{\mu,h}$
(where $\mu$is the layer and $h$ the head),
for a context window of length $2N+1$

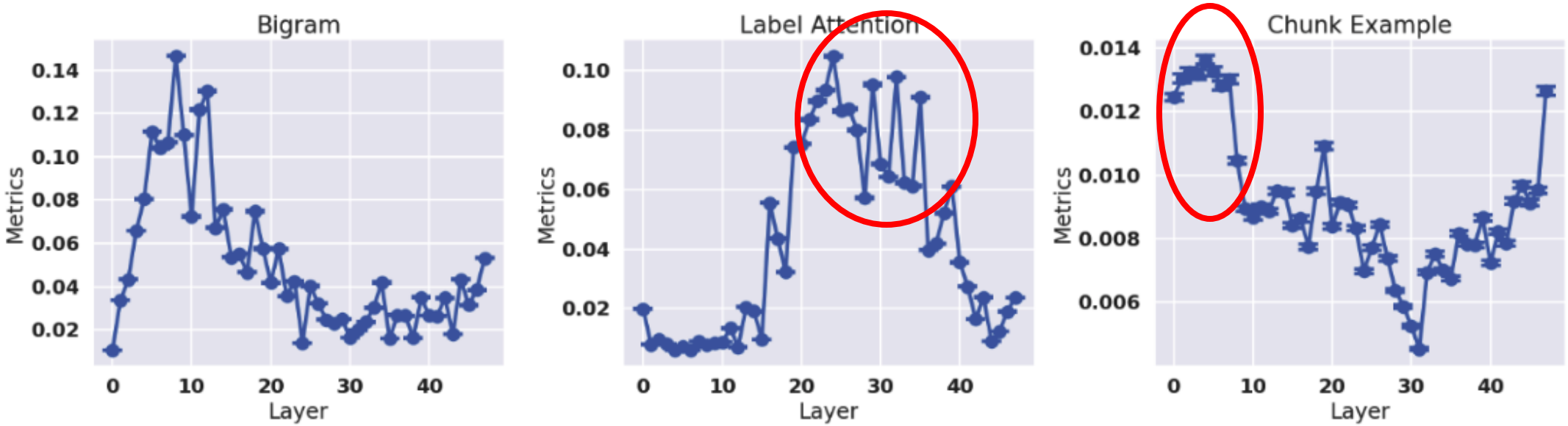
| Metric | Formula |
|---|---|
| Bigram | $p_{2N+1,2N+1}^{\mu,h}$ |
| Label Attention | $\sum_{k=1}^{N} p_{2N+1,2k}^{\mu,h}$ |
| Chunk Example | $\frac{1}{N}\sum_{k=1}^{N} p_{2k,2k-1}^{\mu,h}$ |

**Correlation with Accuracy:**
The timing of these metric transitions closely matches the model's discrete accuracy jumps, validating that our metrics quantitatively capture the internal circuit reconfigurations across all three learning phases



# 3 Circuits in LLM

- Setup
  - Test whether our identified circuits appear in LLMs by evaluating a pretrained GPT2-XL (48 layers) on SST2 (872 samples).
  - 2-shot prompt: two labeled examples (Review: {text}, Sentiment: {label}) and a third, unlabeled query
- Results
  - **Chunk Example** scores peak in earlier layers while **Label Attention** scores are higher in middle or later layers, consistent with the final circuit (FCC) behavior in our 2-layer attention-only model



# Conclusion and Future Direction

- Conclusion
  - We watched what happens inside a model in a *practical* few-shot setting.
  - The model passed through **three distinct circuits**—NCC, SCC, and FCC—before it reached perfect accuracy.
  - Those same circuits also show up inside a pretrained LLM (GPT-2 XL), so the toy findings really scale.

  You'll find extra details in the paper—multi-head results and how the circuits depend on data distribution property.

X

paper

松尾・岩澤研究室

MATSUO-IWASAWA LAB UTOKYO