

Large Language Models

Week 12 · November 10 & 12, 2026

Harun Pirim

Table of contents

1	Learning Objectives	1
2	From Transformers to Language Models	2
3	Tokenization and Byte-Pair Encoding	2
4	Pretraining Objectives	2
5	Scaling Laws	3
6	Rotary Position Embeddings	3
7	Inference: KV Caching and Sampling	3
8	Fine-Tuning, Instruction Tuning, and LoRA	3
9	Capabilities and Limitations	3
10	Summary	3
11	Exercises	3
12	Further Reading	4

Reading Prince, *Understanding Deep Learning*, Chapter 12 and supplementary notes · **Due** Project progress update, Thursday, November 12

 These notes are still being written. The reading above is authoritative until this banner disappears.

1 Learning Objectives

By the end of this week you should be able to:

1. Describe the next-token-prediction objective and how pretraining data is assembled.
2. Summarize what scaling laws predict and where they break down.
3. Explain subword tokenization and byte-pair encoding.
4. Describe rotary position embeddings and how they differ from sinusoidal encodings.
5. Explain KV caching and the cost structure of autoregressive inference.
6. Outline instruction tuning, RLHF, and parameter-efficient fine-tuning.

```
import numpy as np
import matplotlib.pyplot as plt
import torch

torch.manual_seed(774)
np.random.seed(774)

plt.rcParams.update({
    "figure.figsize": (7, 4),
    "figure.dpi": 130,
    "axes.spines.top": False,
    "axes.spines.right": False,
    "axes.grid": True,
    "grid.alpha": 0.25,
    "font.size": 10,
    "legend.frameon": False,
})

NDSU_GREEN = "#005643"
NDSU_GOLD = "#FFC72C"
ACCENT = "#C1121F"
```

2 From Transformers to Language Models

TODO

3 Tokenization and Byte-Pair Encoding

TODO

4 Pretraining Objectives

TODO

5 Scaling Laws

TODO

6 Rotary Position Embeddings

TODO

7 Inference: KV Caching and Sampling

TODO

8 Fine-Tuning, Instruction Tuning, and LoRA

TODO

9 Capabilities and Limitations

TODO

10 Summary

TODO

11 Exercises

Problem numbers refer to Prince, Chapter 12 and supplementary notes.

1. TODO
 2. TODO
 3. Computational. TODO
-

12 Further Reading

- Prince, *Understanding Deep Learning*, Chapter 12 and supplementary notes.
- TODO