

Reading the Original Transformer Paper

Gavin · AI-assisted course design

Contents

1	Goal	3
2	Part 1 — Orient before Lesson 2.1	3
	2.1 Find the problem and the proposed change	3
	2.2 Make an orientation sketch	4
	2.3 Stop deliberately	4
3	Part 2 — Return after Lesson 2.3	4
	3.1 Trace rather than recite	4
	3.2 Compare with one documented decoder-only model	5
	3.3 Separate evidence from extrapolation	5
4	Self-check and discussion guide	5
5	Submit your reading note	5
6	More Learning	6

1 Goal

Use a research paper as a recurring reference while your understanding develops. Build an initial map before studying the residual stream, then revisit that map after learning attention and MLPs. You are not expected to understand the entire paper on your first pass.

Primary source: Vaswani and colleagues, *Attention Is All You Need*, originally published in 2017. The linked arXiv version provides accessible HTML; the PDF is an alternative for its figures and equations. Record which version you used.

Prerequisite: Module 1 for Part 1; Lessons 2.1–2.3 for Part 2.

Time: 20–30 minutes now, then about 40–60 minutes after Lesson 2.3.

Requirements: the paper and somewhere to write. No account, API, model download, GPU, paid compute, or programming is required. Keep the paper available locally if you want to do the reading offline.

Expected artifact: one short reading note with dated first-pass and return-pass sections, a simple sketch, source locations, and revised explanations. An incomplete but honest map is more useful than an elaborate summary you cannot explain.

2 Part 1 — Orient before Lesson 2.1

Read only the abstract, introduction, Figure 1 and its caption, and conclusion. Scan the section headings to know where further explanations live. Set a timer if you tend to stop at every unfamiliar term.

Write three kinds of notes separately:

- **What the authors claim:** attach a section or figure number.
- **My current interpretation:** explain it in your own words.
- **A question I cannot answer yet:** leave it open.

This distinction prevents a tentative interpretation from quietly becoming a statement attributed to the paper. Avoid copying sentences into your submission except for an occasional short phrase that genuinely needs quotation.

2.1 Find the problem and the proposed change

Answer these questions in one or two sentences each:

1. What task setting motivates the work? Identify an example from the paper rather than substituting a modern chatbot use case.
2. Which broad architectural approaches are the authors comparing against?
3. What do the authors propose changing about the main sequence-processing mechanism?
4. Which proposed benefits concern computation, and which concern measured task performance?
5. Does the title imply that the network literally contains nothing except attention? Record your initial answer and what in Figure 1 supports it.

You may leave technical details unresolved. The aim is to identify the research question and the shape of the argument.

2.2 Make an orientation sketch

Draw a small, original sketch with two stack-shaped boxes and the major arrows between them. Use your own arrangement rather than tracing or redistributing the paper’s figure. Label where the input enters and where output predictions emerge.

Choose five labels from Figure 1. For each, write one of:

- I can explain this operation.
- I recognize its role but cannot calculate it.
- I do not know what it means yet.

Include residual addition or “Add & Norm” among the labels. Mark the connection whose purpose is least clear to you. You will revisit it after Lesson 2.3.

2.3 Stop deliberately

Write a provisional three-sentence explanation of the architecture. Then list up to five questions you want Module 2 to answer. Good questions name a missing connection: “Why does this addition need equal dimensions?” is more useful than “How do Transformers work?”

Now continue to Lesson 2.1. You do not need to derive attention, reproduce training, or resolve every symbol before moving on.

3 Part 2 — Return after Lesson 2.3

Reopen your initial note without replacing it. Read Sections 3.1–3.3 carefully, referring to Figure 1 and the attention figure. Read Sections 3.4–3.5 for the input and output boundaries. Treat details not yet covered in the course as questions for Lesson 2.4.

3.1 Trace rather than recite

Add to your sketch and answer:

1. Where are the residual additions? What shapes must agree at each one?
2. Where is normalization relative to the addition? How does this differ from Lesson 2.1’s main pre-norm illustration?
3. Which operations allow information to move between positions, and which apply a transformation separately at each position?
4. Which attention operation reads information originating in the other stack? Trace both ends of that arrow.
5. Which positions must be unavailable when the decoder predicts a token, and how does the architecture enforce that restriction?
6. How can the feedforward component use an internal width different from the width of the surrounding stream?

For one attention equation, annotate the meaning and shape of each matrix under the paper’s conventions. Explain the final output shape in words. A shape check is more valuable here than memorizing the equation’s typography.

3.2 Compare with one documented decoder-only model

Use one primary reference, such as the LLaMA report’s architecture section, together with Lesson 2.1. Identify one component that is retained in broad form, one arrangement or operation that changes, and one component of the original encoder–decoder system absent from the chosen decoder-only architecture.

Be specific about the model and source. “Modern Transformers use pre-norm” is too broad: Lesson 2.1 already introduced architectures whose normalization arrangements need more careful description. If an implementation adds detail absent from a report, label which source establishes which fact.

Write one paragraph answering: What would go wrong if I treated Figure 1 as an exact wiring diagram for every current LLM?

3.3 Separate evidence from extrapolation

Choose one result reported in the paper. Record the task, evaluation measure, model variant, and location of the result. You do not need to reproduce it. Explain why that result does not, by itself, establish performance on an unrelated present-day task.

Then revise your original three-sentence explanation. Preserve the earlier version and identify two concrete improvements. Name one question still open after the second pass. Finishing the Lab does not require pretending the paper has become completely transparent.

4 Self-check and discussion guide

Compare only after writing your own answers.

i Orientation checks

The paper studies sequence transduction, especially machine translation. Its proposed architecture replaces recurrence and convolution in its main sequence model with attention-based processing. Figure 1 still contains embeddings, feedforward operations, residual additions, normalization, and output components.

The original model has encoder and decoder stacks; a decoder-only language model does not retain that complete two-stack arrangement. Post-norm placement in the original differs from the chapter’s sequential pre-norm example. Your comparison should identify a particular modern architecture rather than generalize from the word “Transformer.”

For the return pass, a strong answer explains dependencies, checks shapes, and cites locations. If you can name every box but cannot trace what supplies its input, return to that connection. If you can calculate an equation but cannot explain why a mask is needed, revisit the prediction setting.

5 Submit your reading note

Include your first-pass map, unresolved questions, annotated second-pass sketch, one model comparison, one bounded result interpretation, and revised explanation. Clearly distinguish the paper’s reported results from experiments you ran; this reading Lab requires no executed model experiment.

6 More Learning

- The original paper remains the main source. Rereading a specific section with a sharper question often helps more than opening another general overview.
- LLaMA architecture provides one concrete comparison, including normalization and positional choices.
- Trace a Residual Stream turns one part of the wiring diagram into arithmetic. Return to your sketch afterward and mark exactly which operations that toy omits.