

Transformer Evaluation Brief

A compact, reproducible publishing example

Lemmafour Documentation

2026-09-04

Contents

- 1 Evaluation question 3
- 2 Evidence 3
- 3 Conclusion 4
- References 5

1 Evaluation question

This report asks whether a compact transformer can retain validation quality while reducing inference cost. The example is synthetic, but its structure mirrors a real technical report: a cited claim, a numbered equation, a figure, a table, a diagram, a footnote, and cross-references.

Self-attention lets each token combine evidence from the complete sequence (Vaswani et al., 2017). For a query matrix Q , key matrix K , value matrix V , and key dimension d_k , the operation used here is shown in Equation 1.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

2 Evidence

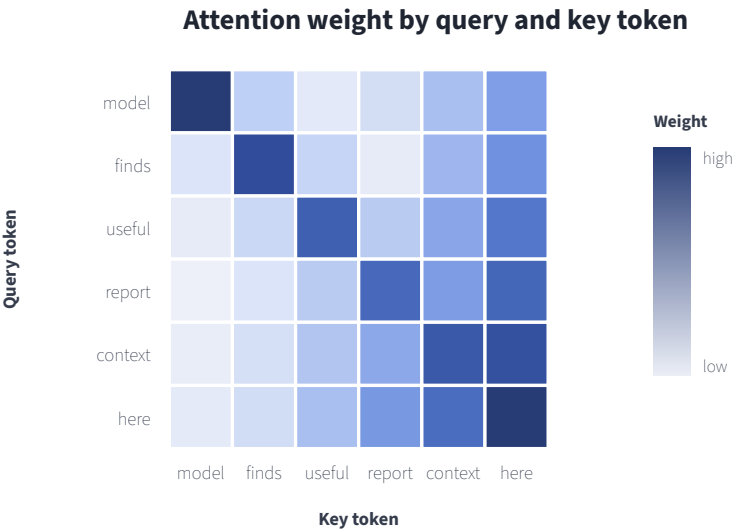


Figure 1: Synthetic attention weights for six report tokens

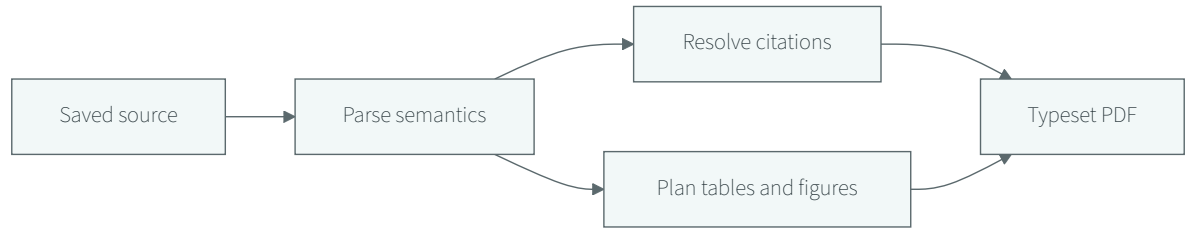
The pattern in Figure 1 is illustrative rather than a trained model result. It exists so image sizing, captions, alternative text, and cross-references can be checked without publishing private research data.

Table 1: Validation metrics by checkpoint

Reading guide: Rows are saved model checkpoints; columns report loss, accuracy, latency, and memory.

Checkpoint	Validation loss	Accuracy	Latency (ms)	Memory (MB)
Baseline	0.82	71%	44	920
Epoch 4	0.54	79%	39	910
Epoch 8	0.39	84%	34	884
Epoch 12	0.31	87%	31	860

The final checkpoint in Table 1 improves the synthetic quality measures while reducing the two resource measures. A production report would also state confidence intervals and test-set provenance.¹



¹All values and labels in this demonstration are synthetic and contain no user or research-subject information.

3 Conclusion

The fixture demonstrates document structure, not a scientific result. It is intentionally small enough to inspect and reuse as a compatibility test.

References

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need. *Advances in Neural Information Processing Systems*, 30.