Natural Language Processing

Week 5: Transformers - Predicting the Next Word

Understanding How Transformers Excel at Language Prediction

NLP Course 2025

Predicting the Next Word: The Core Challenge

Why is next word prediction important?

The Problem

- Language models predict: P(next word — context)
- RNNs struggle with long contexts
- Order matters for meaning
- Dependencies span many words

The Solution

- Transformers use attention mechanism
- Access all context equally
- Multiple prediction hypotheses
- Position-aware predictions

Goal: Better next word predictions = Better language understanding

Part 1: The Prediction Problem

What makes next word prediction challenging?

Chart 1: Next Word Probability Distribution

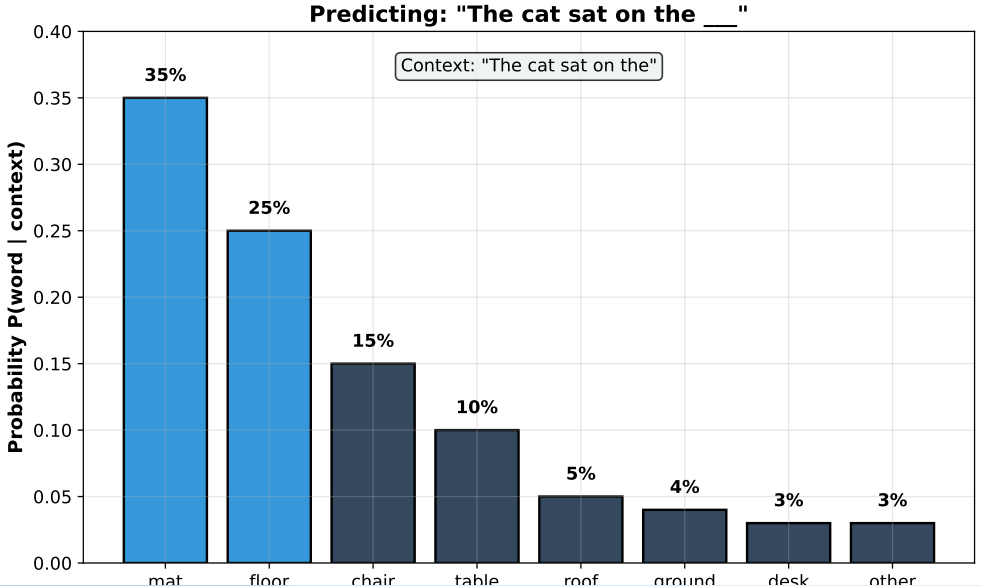
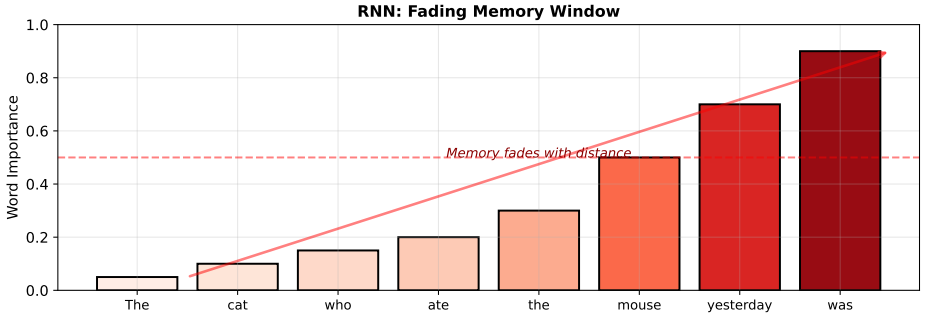
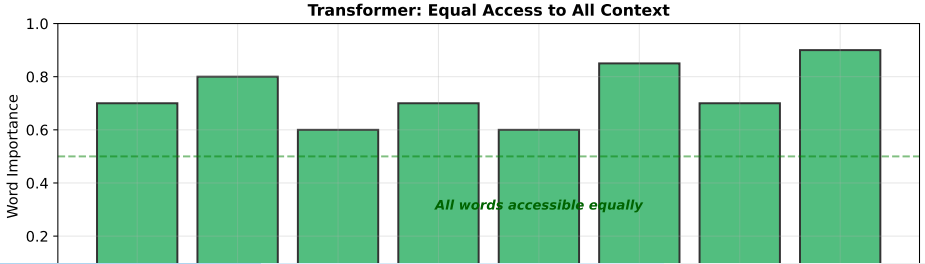
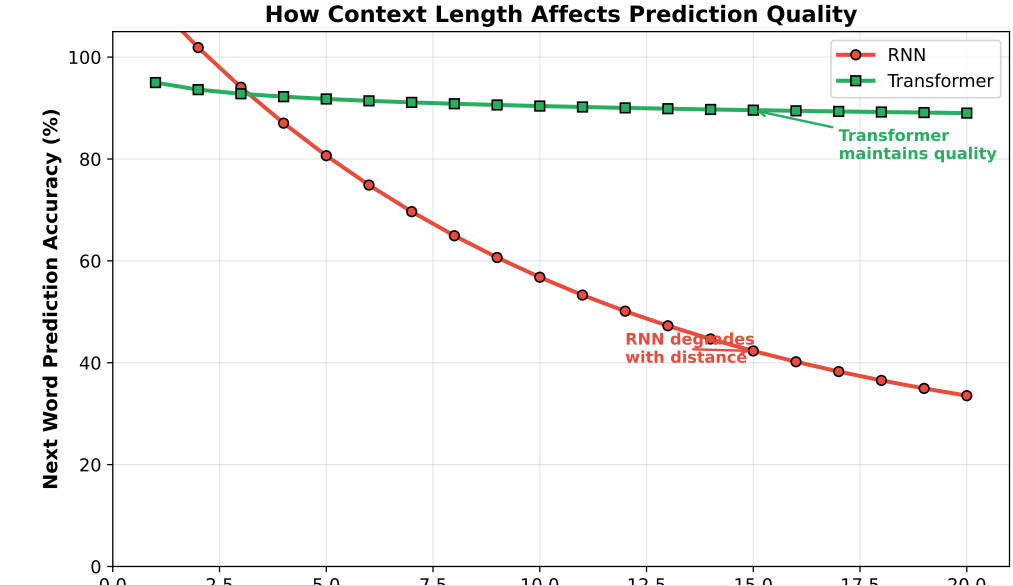


Chart 2: Context Window for Prediction






How Context Length Affects Prediction Quality
Next Word Prediction Accuracy (%)
RNN
Transformer
Transformer maintains quality
RNN degrades with distance
0
20
40
60
80
100
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0

Part 2: Why RNN Predictions Fail

Three fundamental problems with sequential processing

Chart 4: RNNs Forget Important Words

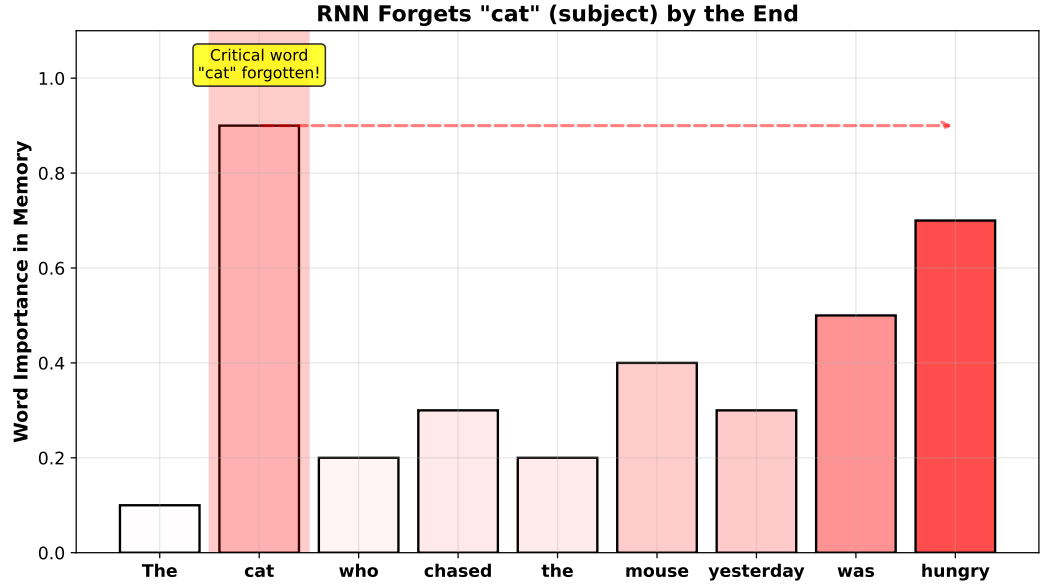
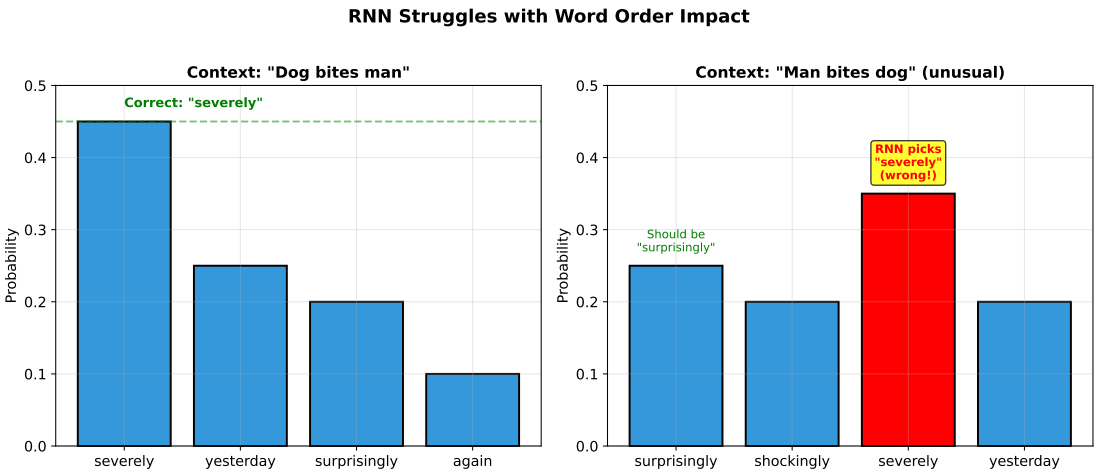


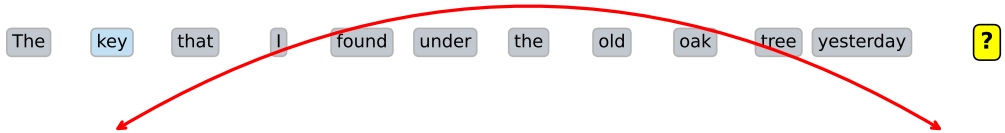
Chart 5: Word Order Confusion



Same words, different order = different meaning. RNNs struggle to maintain order information for correct predictions.

Chart 6: Long-Range Dependency Problem

Long-Range Dependency: "The key ... was/were lost"



10 words distance

RNN: Forgets "key" is singular → predicts "were" (wrong)

Part 3: Transformer Solutions

Three innovations for better predictions

Chart 7: Attention to Relevant Words

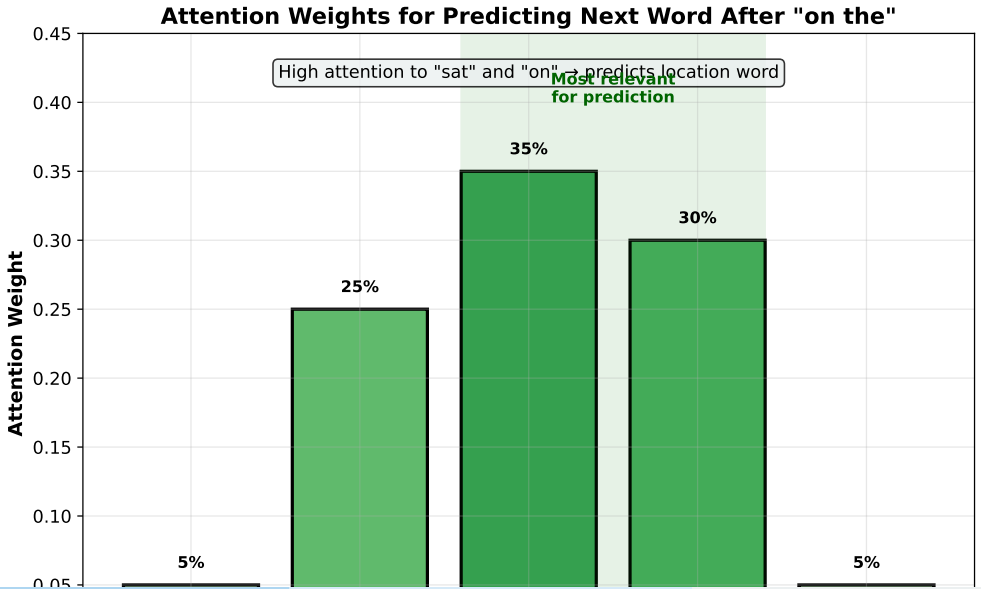
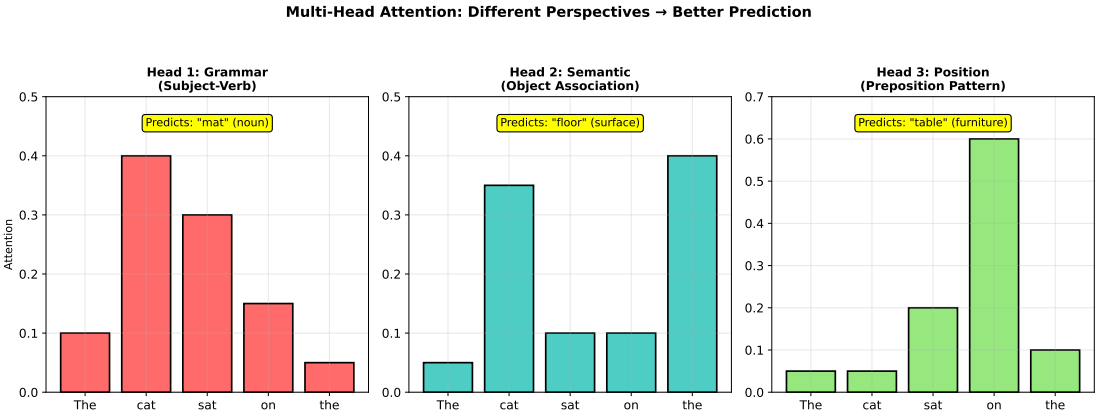
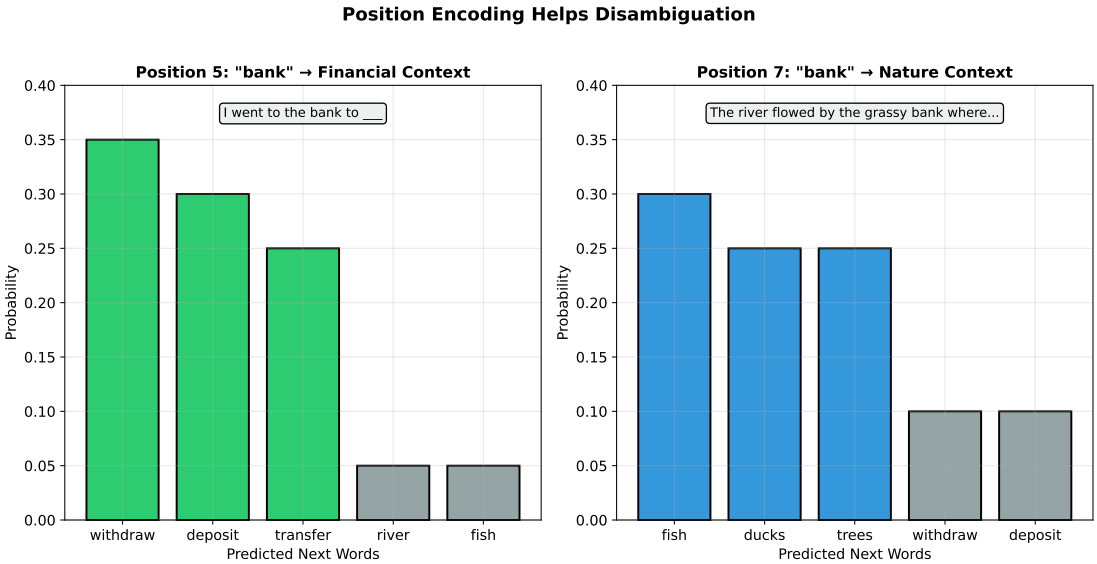


Chart 8: Multiple Prediction Perspectives



Different attention heads focus on grammar, meaning, and position - combining perspectives improves predictions.

Chart 9: Position-Aware Predictions



Part 4: Measuring Prediction Quality

How much better are transformer predictions?

Chart 10: Prediction Uncertainty (Perplexity)

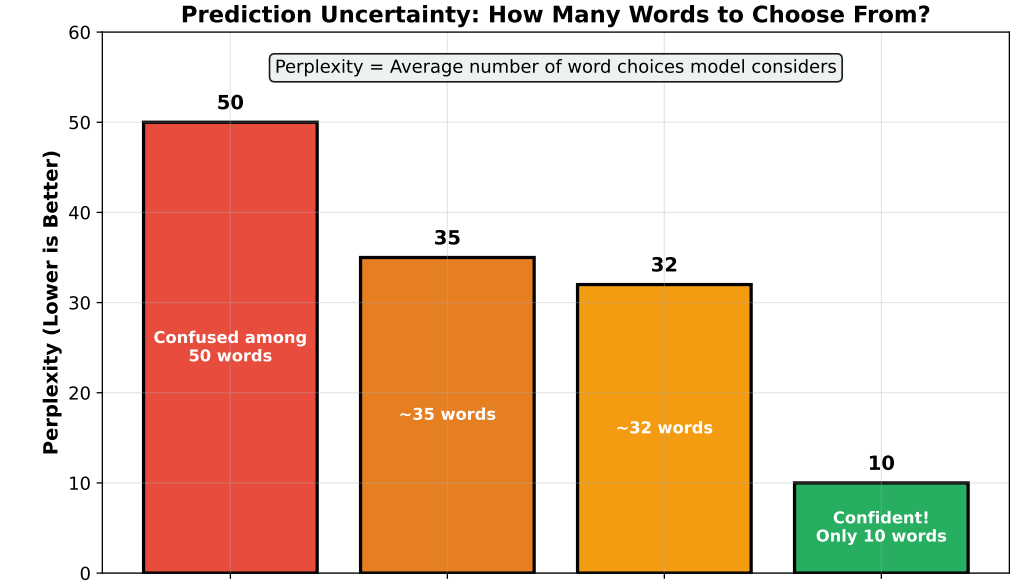


Chart 11: Prediction Accuracy

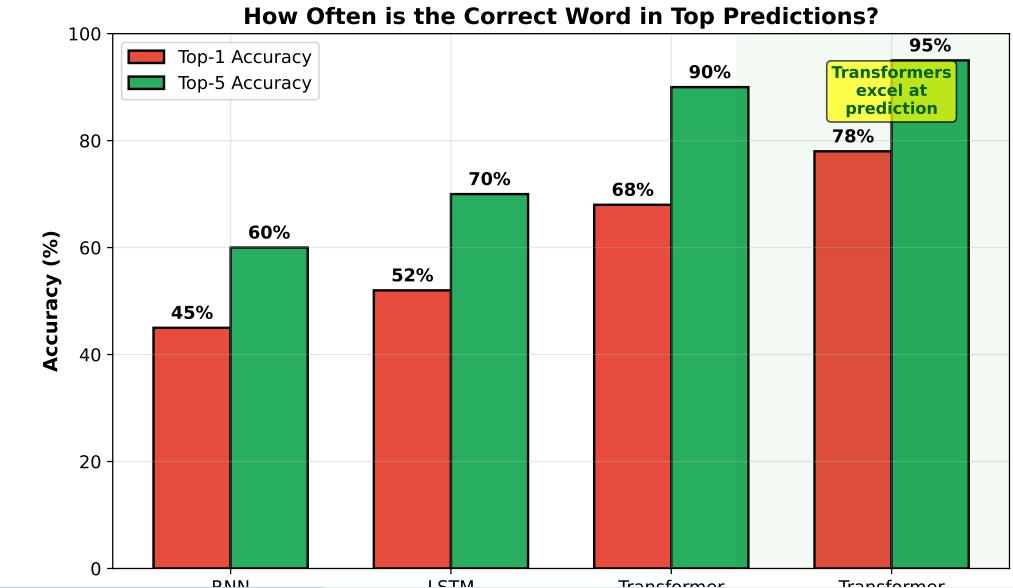
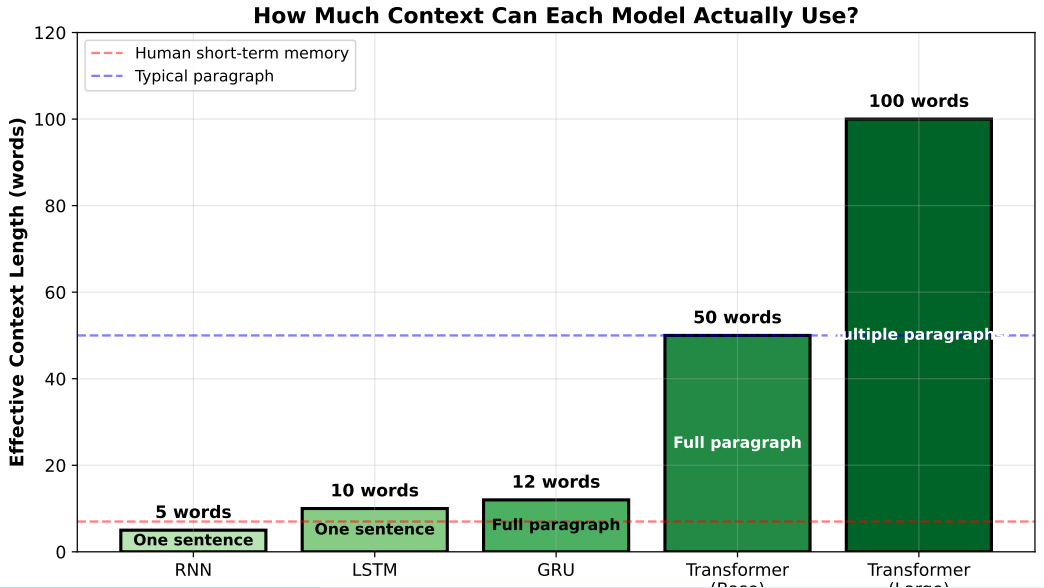


Chart 12: Effective Context Usage



Summary: Why Transformers Predict Better

The Next Word Prediction Revolution

RNN Problems

- Forgets important words
- Loses word order
- Can't handle long dependencies
- Limited to recent context
- High prediction uncertainty

Transformer Solutions

- Attention remembers all words
- Position encoding preserves order
- Direct connections span distance
- Uses full context equally
- Confident, accurate predictions

Result: From 60% to 95% prediction accuracy - enabling ChatGPT, Claude, and modern AI

Better predictions = Better language understanding = Better AI

Part 5: Mathematical Foundations

Understanding the Mathematics Behind Transformers

Chart 13: From Scores to Probabilities - Softmax

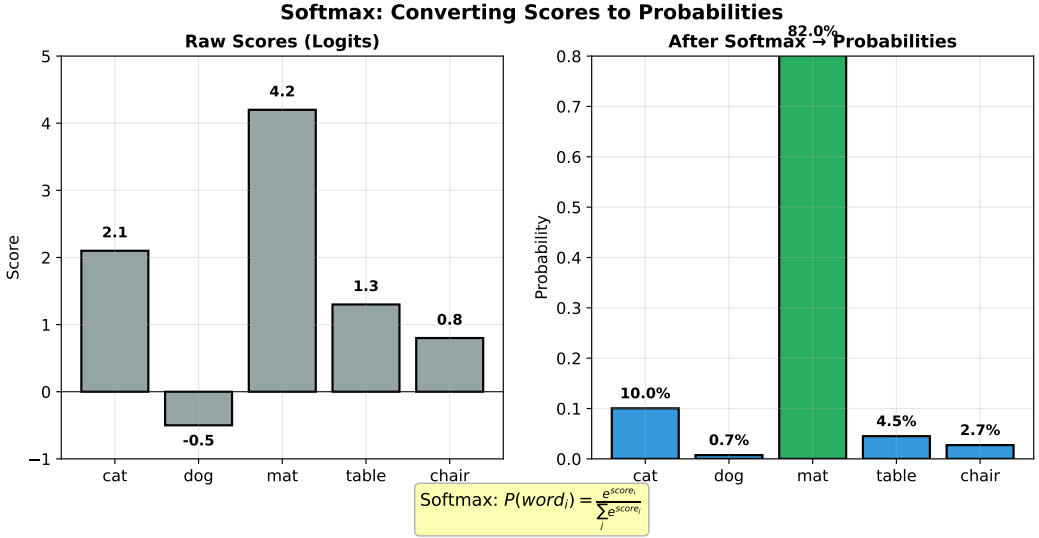
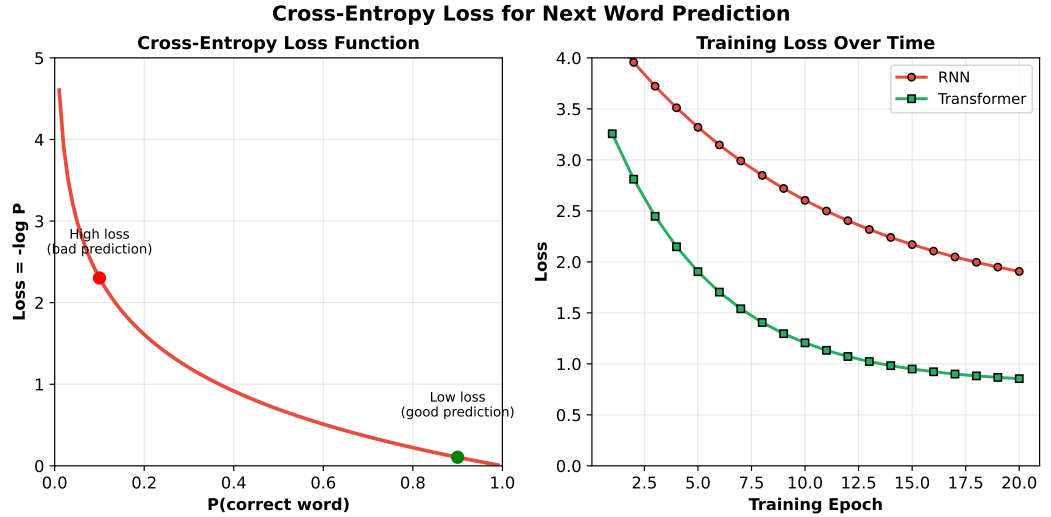


Chart 14: Training Objective - Cross-Entropy Loss



Loss measures how wrong our predictions are. Lower loss = better predictions. Transformers achieve lower

Chart 15: How Attention Actually Works

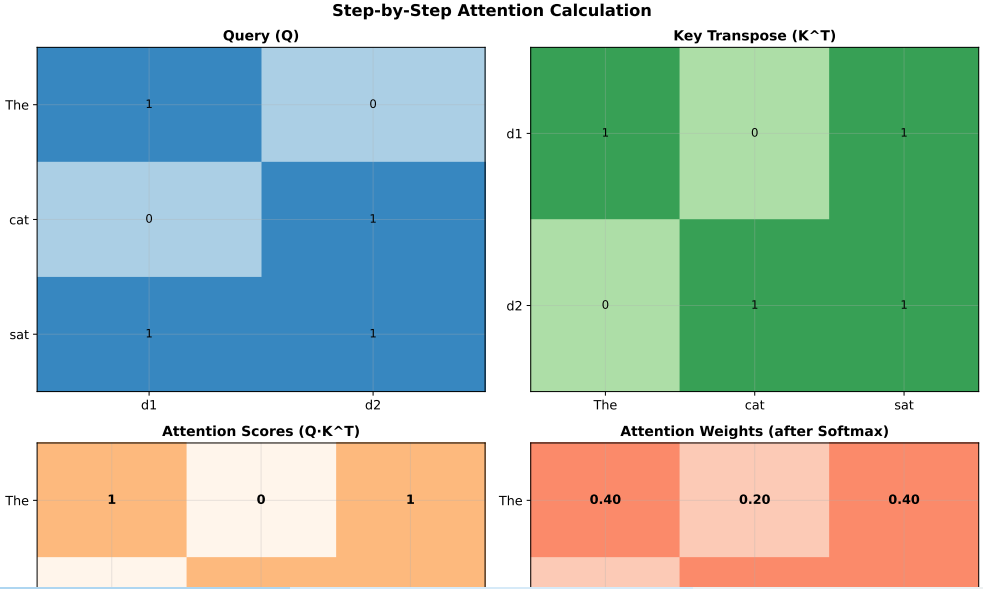


Chart 16: Words as Vectors in Space

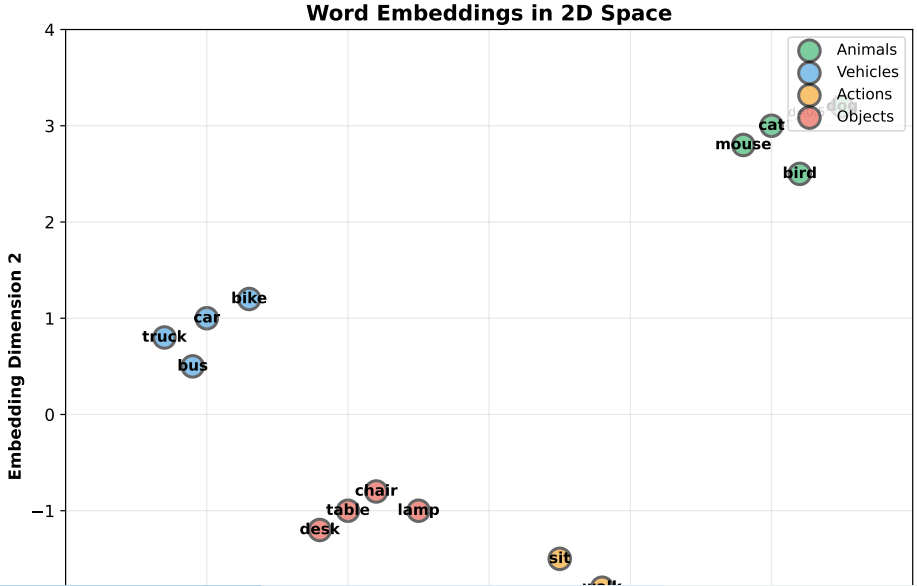
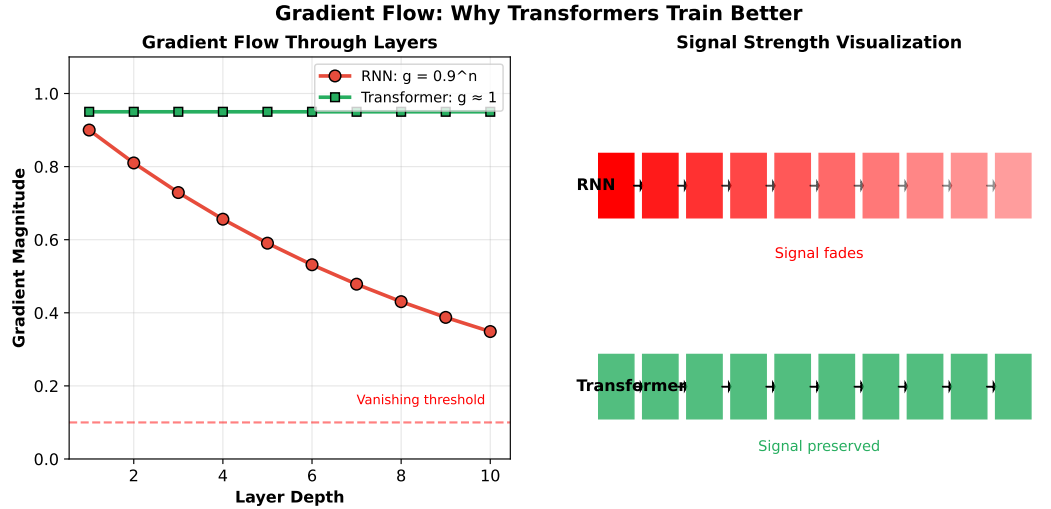
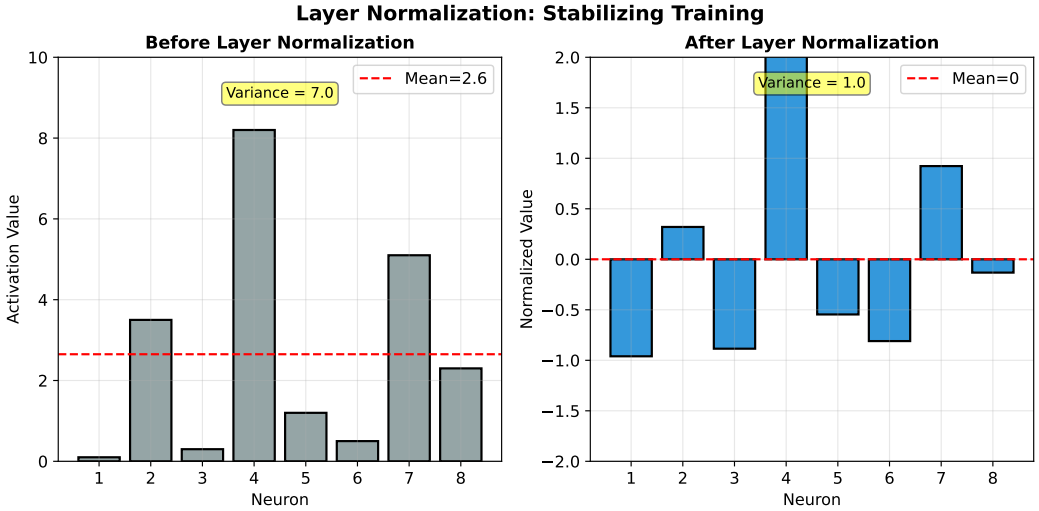


Chart 17: Why Transformers Train Better



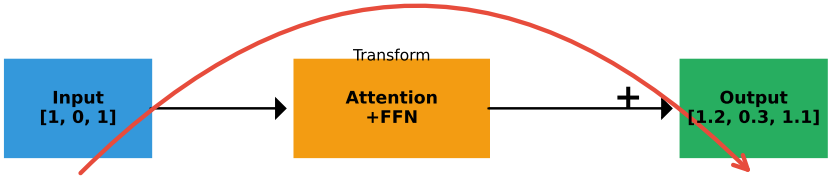
Stable gradient flow enables deep networks. RNNs suffer from vanishing gradients; Transformers maintain signal strength.



Layer normalization keeps values in a stable range (mean=0, variance=1), preventing training instabilities.

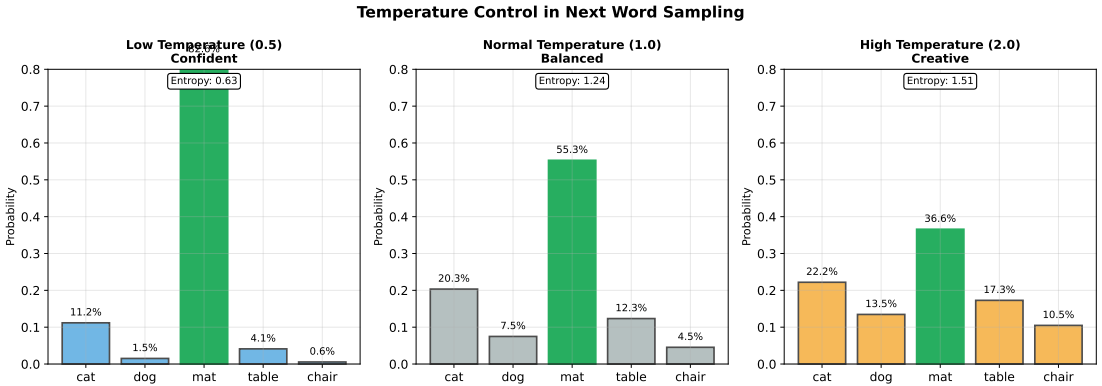
Residual Connections: Preserving Information

Output = Input + Transformation(Input)



Residual (Identity) Connection

Chart 20: Controlling Creativity - Temperature



Temperature controls prediction diversity. Low = confident/repetitive, High = creative/random. Tune for your needs!

Complete Picture: Theory + Mathematics

What We've Learned

Conceptual Understanding

- Next word prediction challenge
- Why RNNs fail (forgetting, order)
- How attention solves it
- Prediction quality metrics

Mathematical Foundation

- Softmax: scores $\rightarrow$ probabilities
- Cross-entropy: training objective
- Attention: $Q \cdot K^T$ mechanism
- Embeddings: semantic space

Theory + Math = Deep Understanding of Modern AI

Now you understand both WHY and HOW transformers revolutionized NLP!