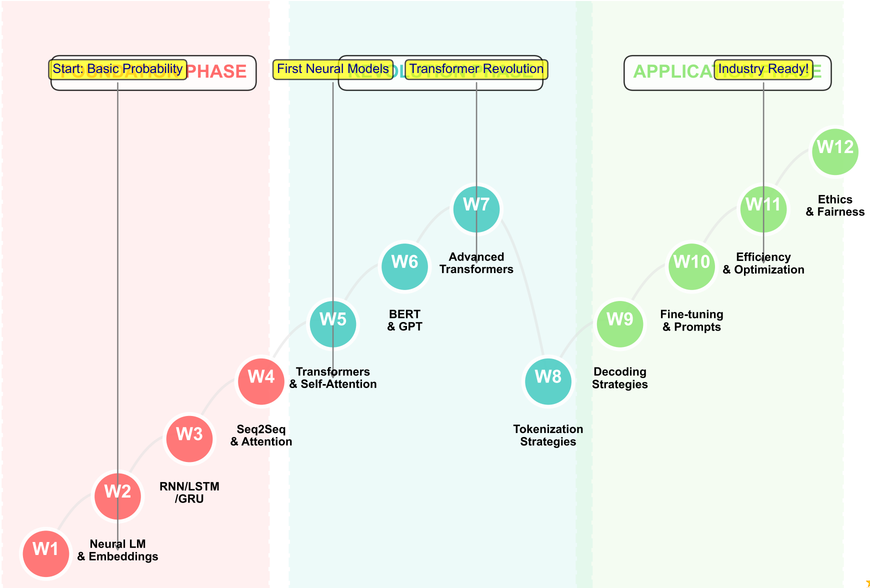
Natural Language Processing: Complete Course Overview

From Foundations to State-of-the-Art

Joerg R. Osterrieder
www.joergosterrieder.com

BSc Computer Science - 12 Week Journey

The NLP Journey: From Counting to Understanding



FOUNDATION PHASE

REVOLUTION PHASE

APPLICATION PHASE

Start: Basic Probability

First Neural Models

Transformer Revolution

Industry Ready!

W1 Neural LM & Embeddings

W2 RNN/LSTM /GRU

W3 Seq2Seq & Attention

W4 Transformers & Self-Attention

W5 BERT & GPT

W6 Advanced Transformers

W7 Tokenization Strategies

W8 Decoding Strategies

W9 Fine-tuning & Prompts

W10 Efficiency & Optimization

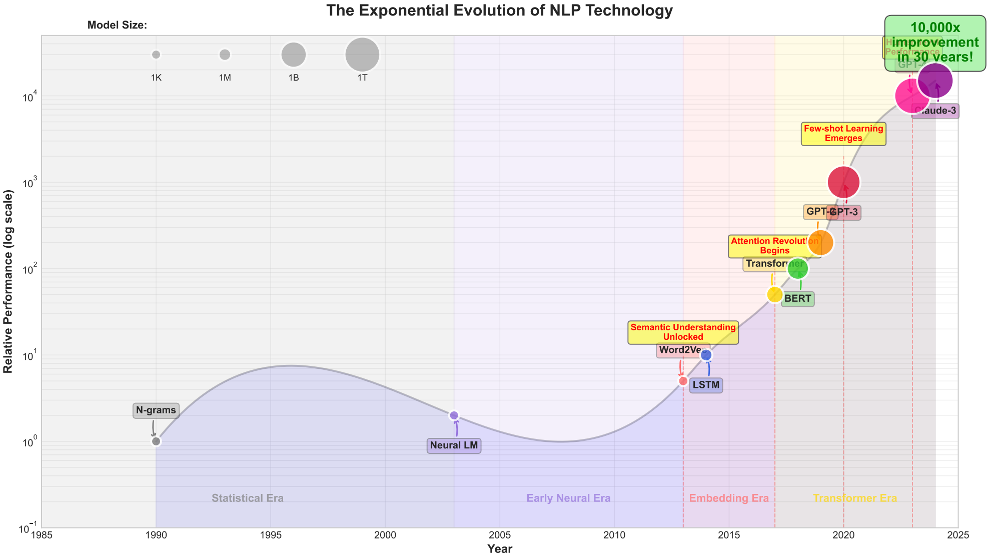
W11 Ethics & Fairness

W12

Complexity:

★ Low

Core Technologies & Breakthroughs

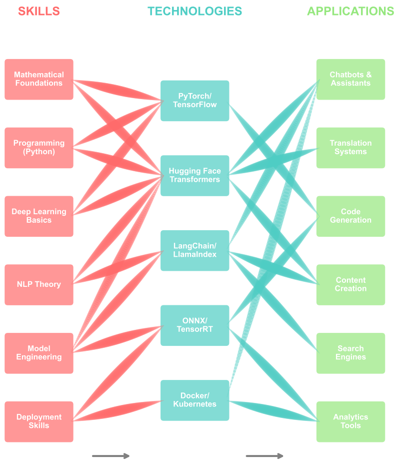


Major Paradigm Shifts:

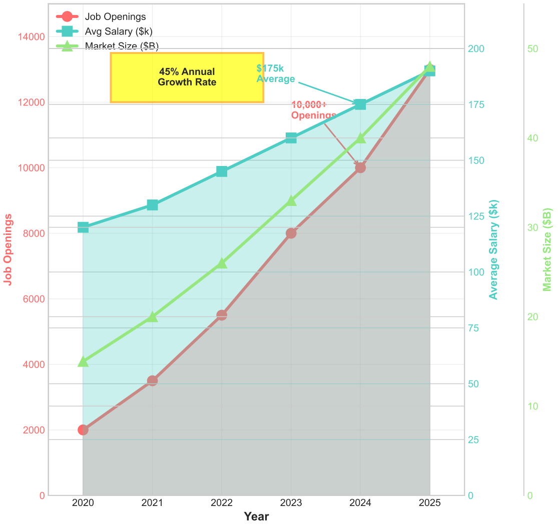
Performance Leaps:

Learning Outcomes & Real-World Impact

From Learning to Real-World Impact



NLP Industry Growth Metrics

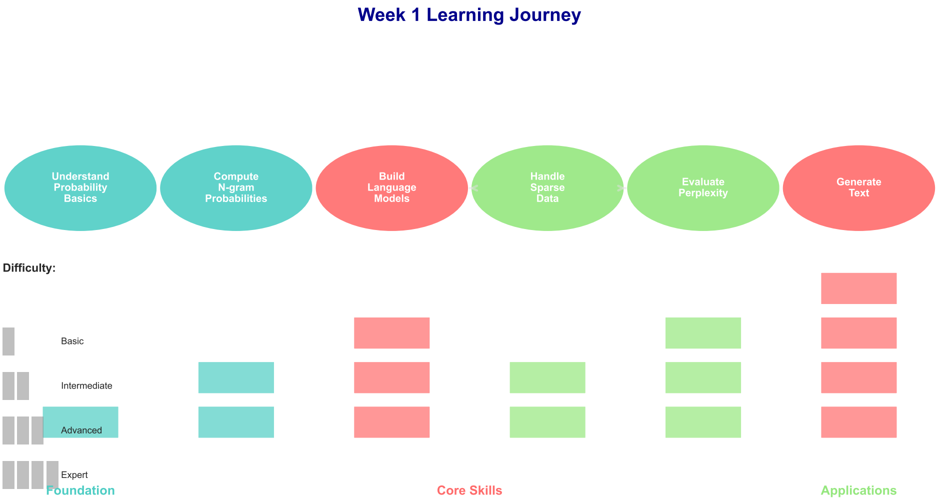


Core Skills

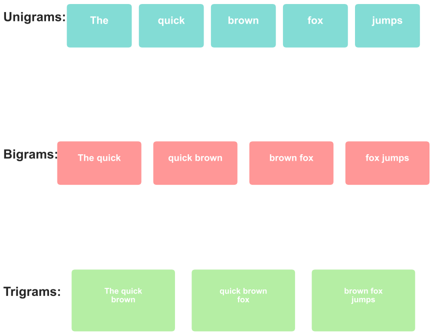
Technologies

Applications

Week 1: What We'll Learn



Week 1: N-gram Language Models

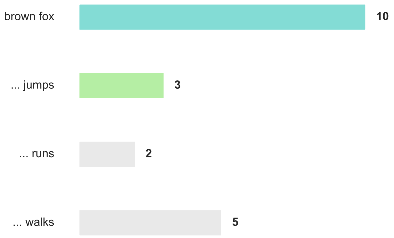


N-gram Extraction

Unigrams: The | quick | brown | fox | jumps

Bigrams: The quick | quick brown | brown fox | fox jumps

Trigrams: The quick brown | quick brown fox | brown fox jumps



Probability Calculation

P(jumps | brown fox)

P(jumps | brown fox) = 3/10 = 0.3

Markov Assumption

- Future depends on recent past
- Makes computation tractable

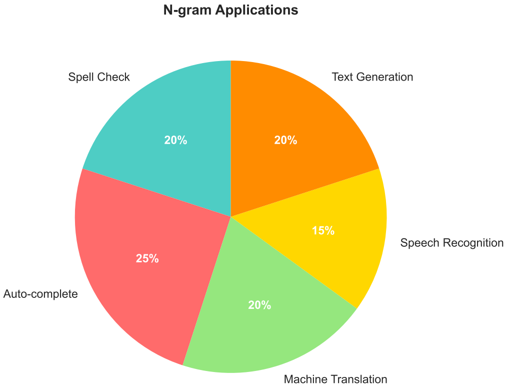
Probability

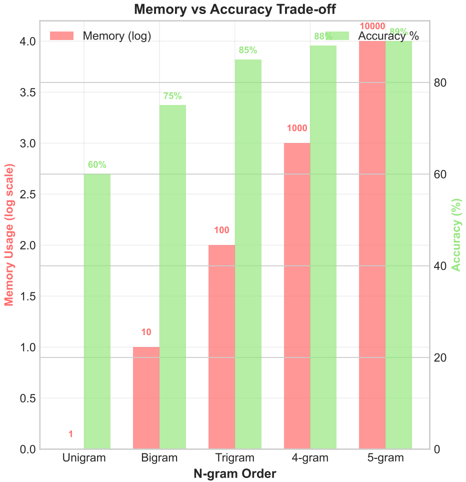
- Count n-grams
- Maximum likelihood

Challenges

- Data sparsity
- Context limitations

Week 1: Applications





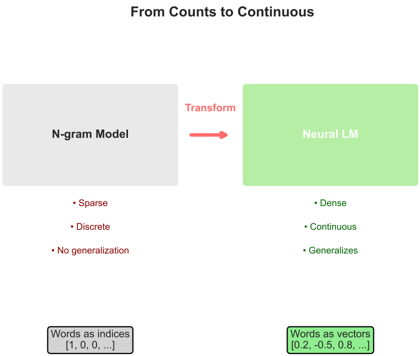
Where N-grams Excel:

- Spell checking
- Auto-complete

Historical Impact:

- Dominated 1980s-2000s
- Still used in hybrid systems

Week 2: Neural Revolution



From Counts to Continuous

N-gram Model

Transform

Neural LM

• Sparse

• Discrete

• No generalization

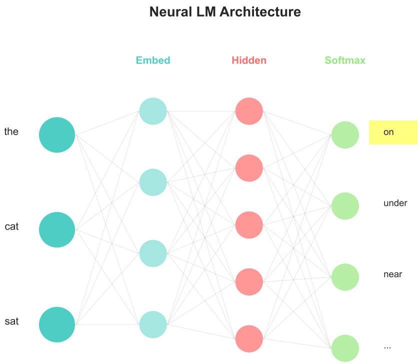
• Dense

• Continuous

• Generalizes

Words as indices
[1, 0, 0, ...]

Words as vectors
[0.2, -0.5, 0.8, ...]



Neural LM Architecture

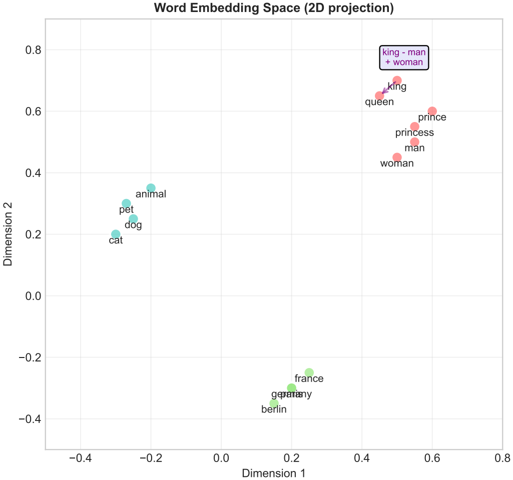
Paradigm Shift:

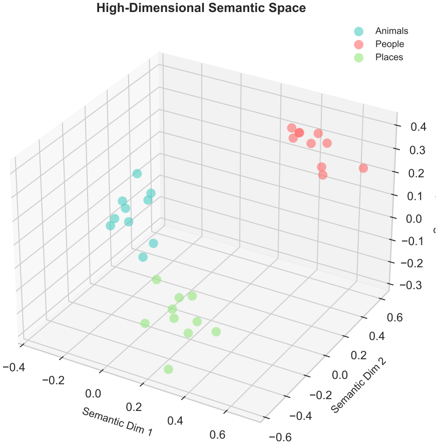
- From counting to learning
- Continuous space

Key Innovations:

- Distributed representations
- Backpropagation training

Week 2: Word Embeddings


Word Embedding Space (2D projection)


High-Dimensional Semantic Space

Properties

- Similar words cluster
- Analogies work

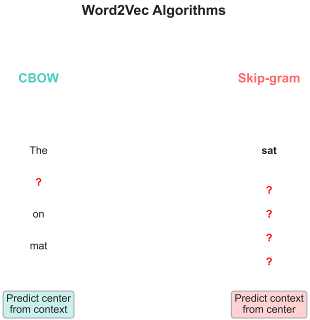
Methods

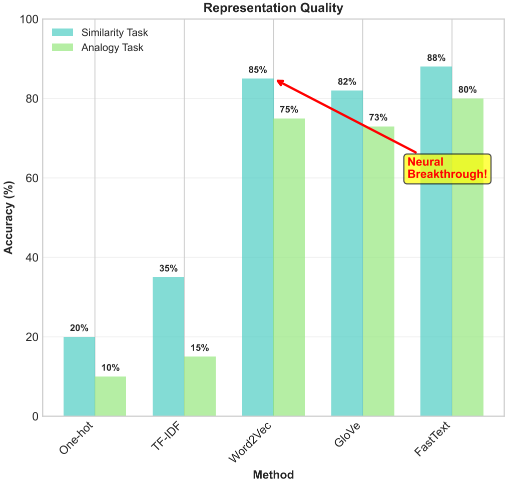
- Word2Vec
- GloVe, FastText

Dimensions

- 50-300 dims typical
- Semantic features

Week 2: Impact





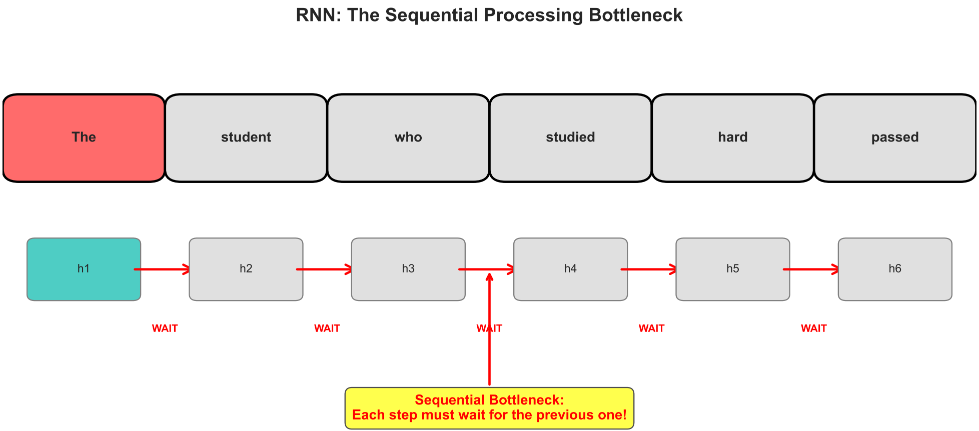
Algorithms:

- CBOW: Context → center
- Skip-gram: Center → context

Applications:

- Sentiment analysis
- Named entity recognition

Week 3: Sequential Processing



RNN: The Sequential Processing Bottleneck

Sequential Bottleneck:
Each step must wait for the previous one!

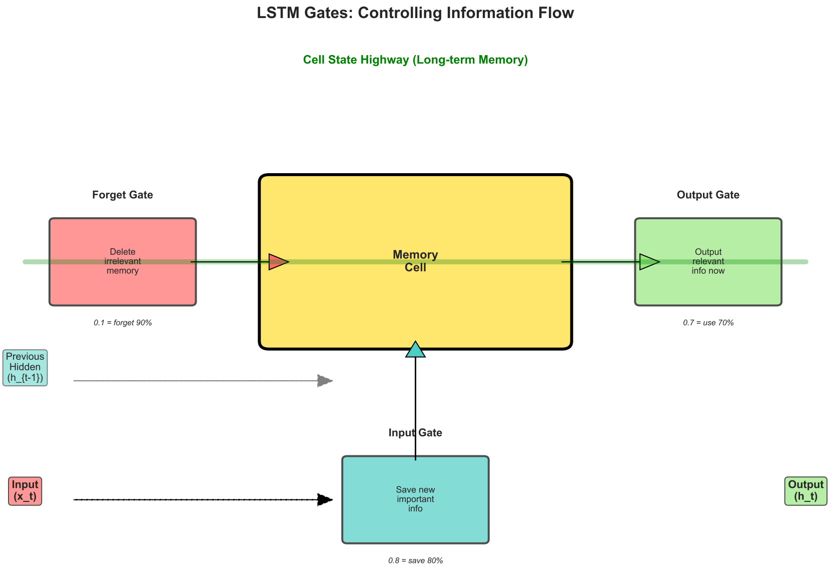
What We'll Learn:

- RNN architecture
- LSTM gates and memory

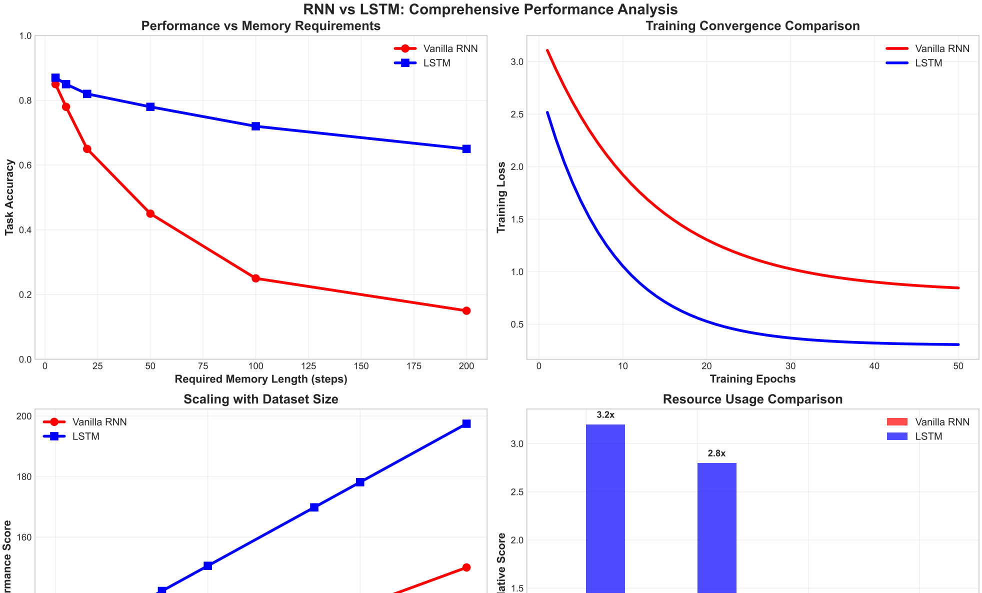
Core Concepts:

- **Vanishing Gradients**
  - Problem with long sequences
  - LSTM/GRU solutions

Week 3: LSTM Architecture



Week 3: Applications



RNN vs LSTM: Comprehensive Performance Analysis

Performance vs Memory Requirements

Training Convergence Comparison

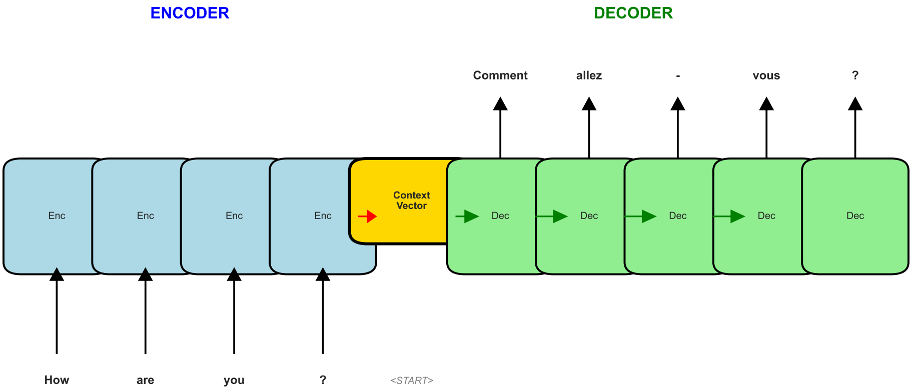
Scaling with Dataset Size

Resource Usage Comparison

Week 4: Encoder-Decoder Architecture

Sequence-to-Sequence Architecture

Variable input length → Fixed context → Variable output length



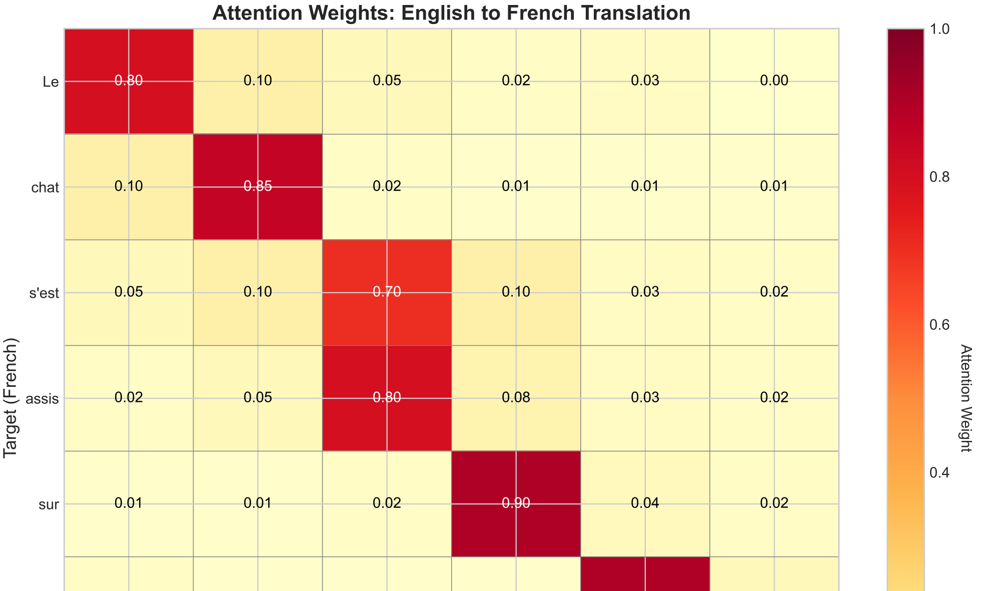
Process entire input

Generate output step-by-step

What We'll Learn:

Key Concepts:

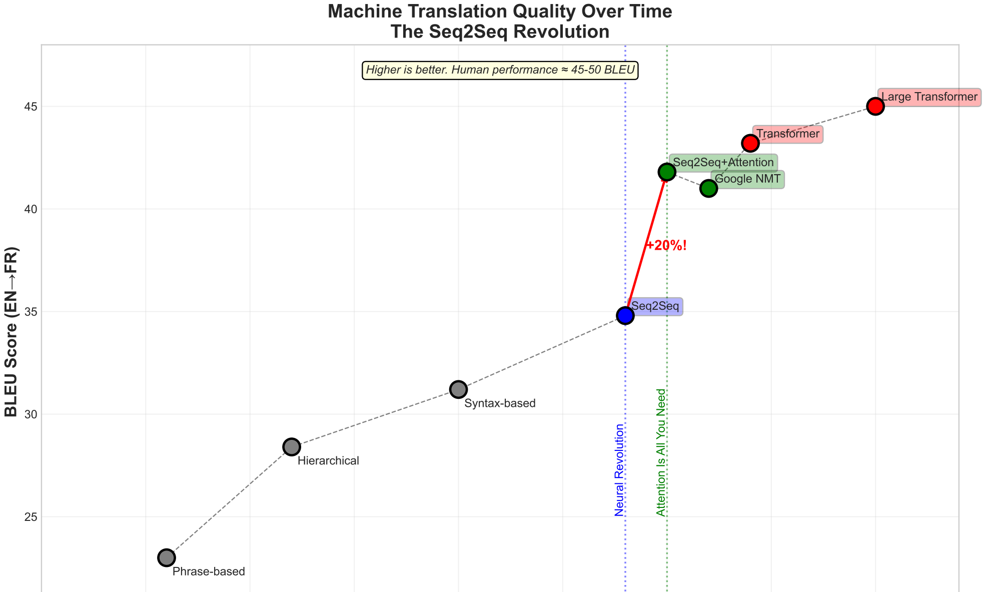
Week 4: Attention Mechanism



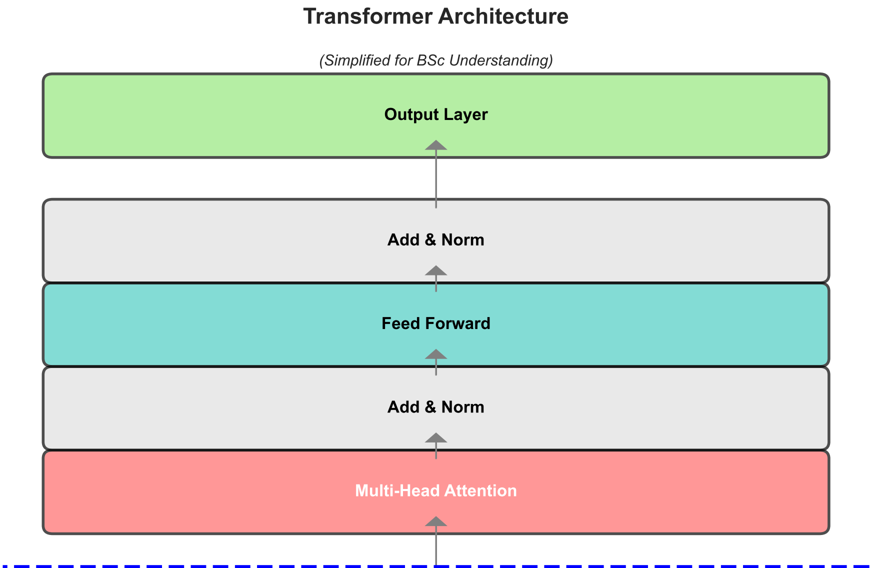
Attention Weights: English to French Translation

Target (French)						
Le	0.80	0.10	0.05	0.02	0.03	0.00
chat	0.10	0.85	0.02	0.01	0.01	0.01
s'est	0.05	0.10	0.70	0.10	0.03	0.02
assis	0.02	0.05	0.80	0.08	0.03	0.02
sur	0.01	0.01	0.02	0.90	0.04	0.02

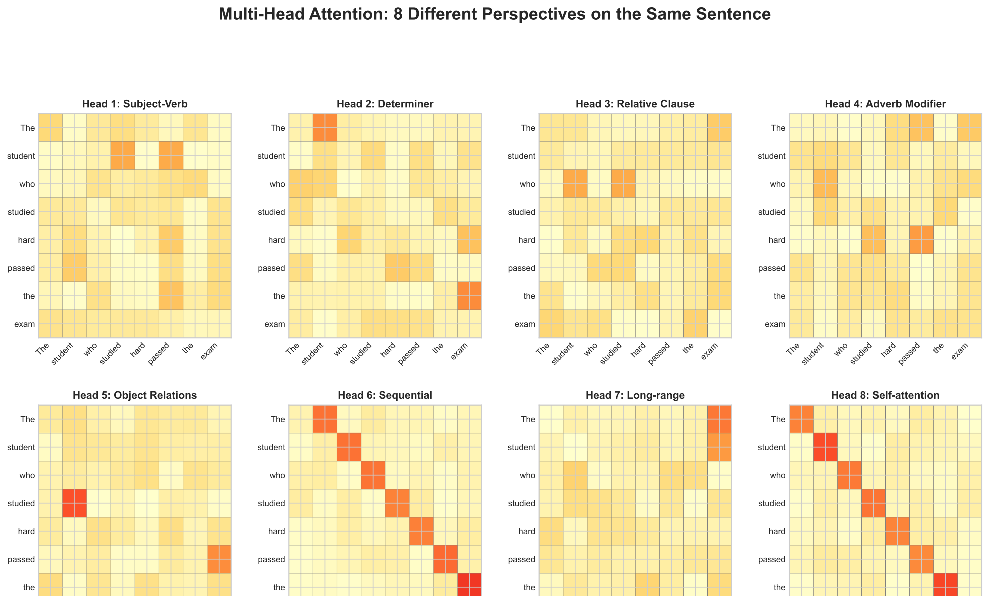
Week 4: Applications



Week 5: Attention Is All You Need



Week 5: Multi-Head Attention



Multi-Head Attention: 8 Different Perspectives on the Same Sentence

Head 1: Subject-Verb

Head 2: Determiner

Head 3: Relative Clause

Head 4: Adverb Modifier

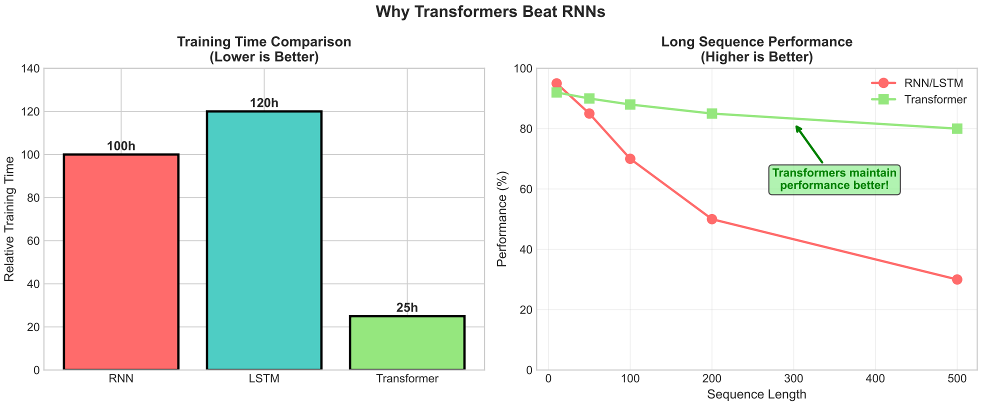
Head 5: Object Relations

Head 6: Sequential

Head 7: Long-range

Head 8: Self-attention

Week 5: Impact



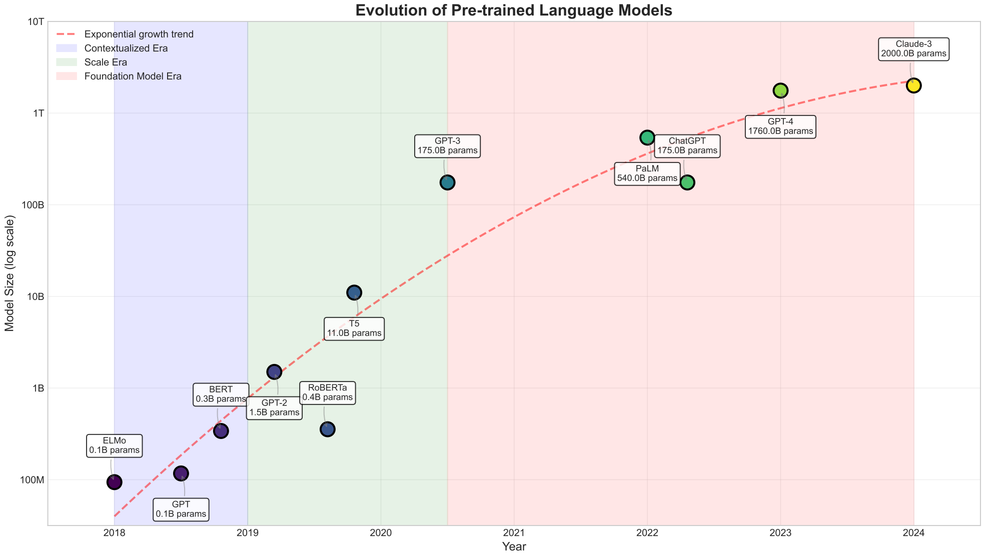
Performance:

- SOTA on all NLP tasks
- 100x faster training
- Scales to billions of parameters
- Transfer learning enabled

Applications:

- Foundation for BERT/GPT
- Computer Vision (ViT)
- Protein Folding
- Multimodal models

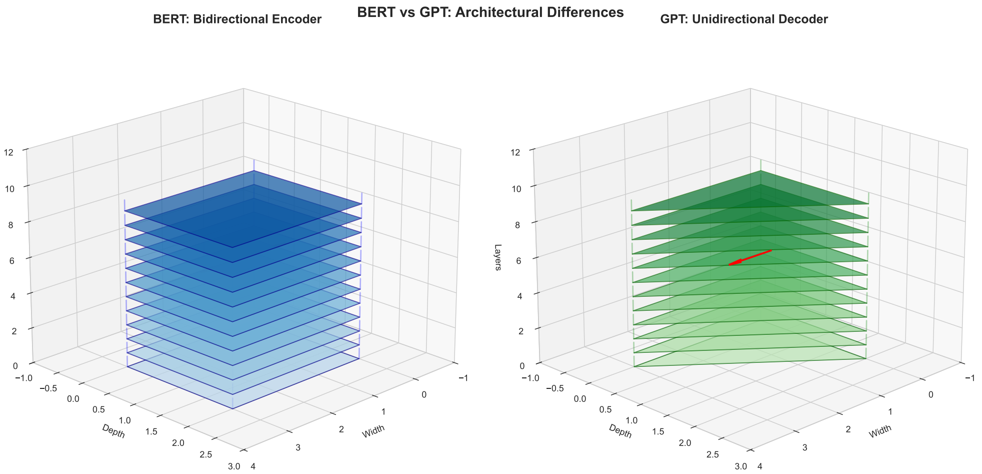
Week 6: BERT and GPT



What We'll Learn:

Key Concepts:

Week 6: BERT vs GPT


BERT vs GPT: Architectural Differences
BERT: Bidirectional Encoder
GPT: Unidirectional Decoder
Depth
Width
Layers
0
2
4
6
8
10
12
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
3.0
−1
0
1
2
3
4

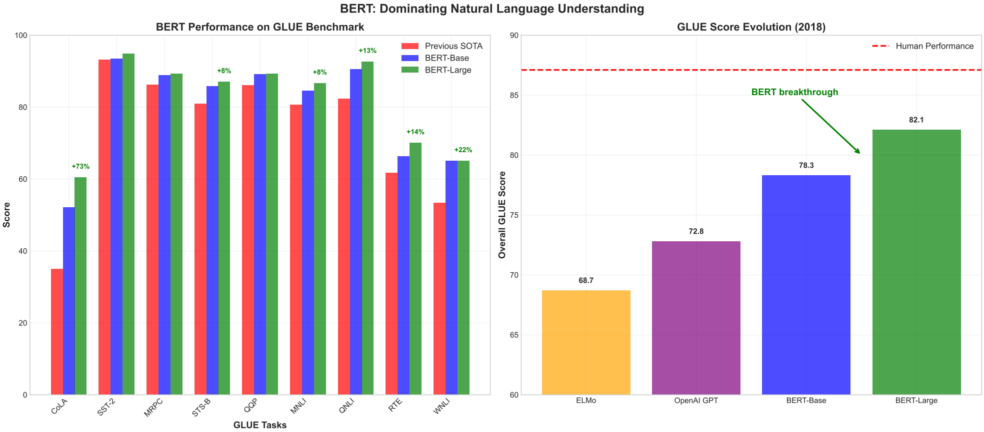
BERT (Bidirectional):

- Masked Language Model
- Sees full context

GPT (Autoregressive):

- Next token prediction
- Left-to-right only

Week 6: Impact



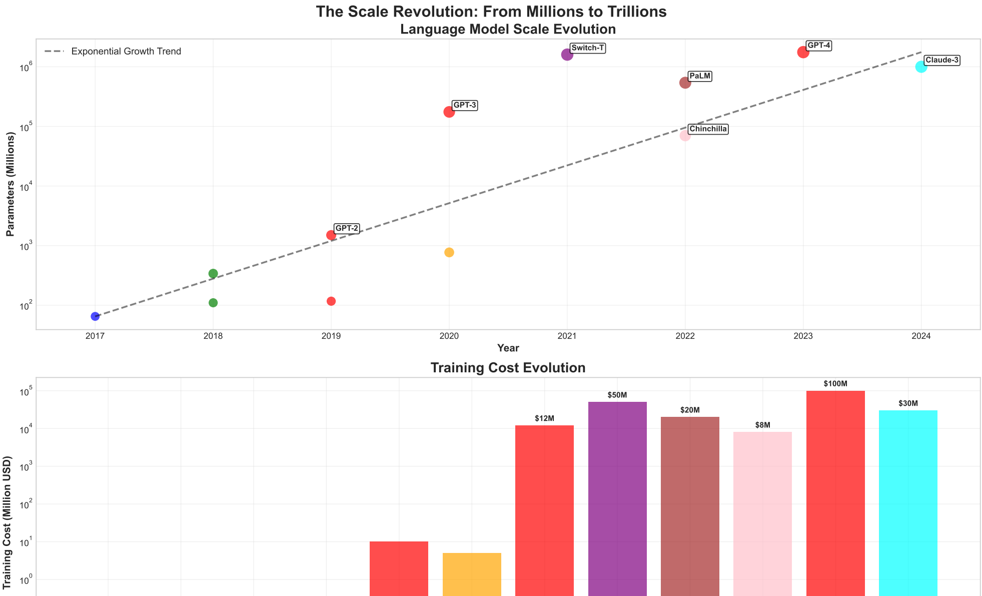
Performance:

- GLUE benchmark SOTA
- Human-level on many tasks
- Few-shot learning
- Zero-shot capabilities

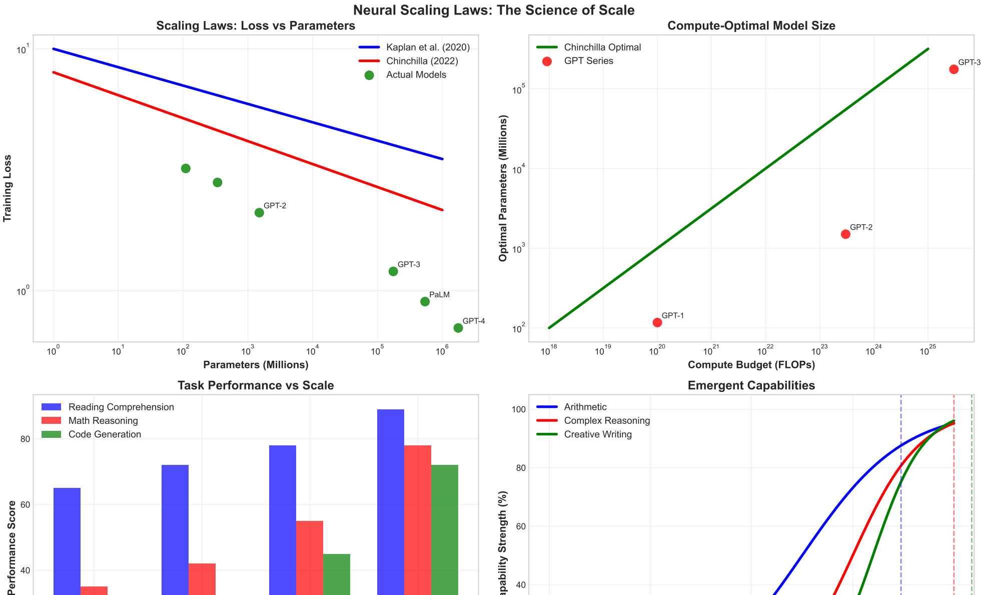
Ecosystem:

- Hugging Face hub
- 100,000+ models
- Easy fine-tuning
- Production ready

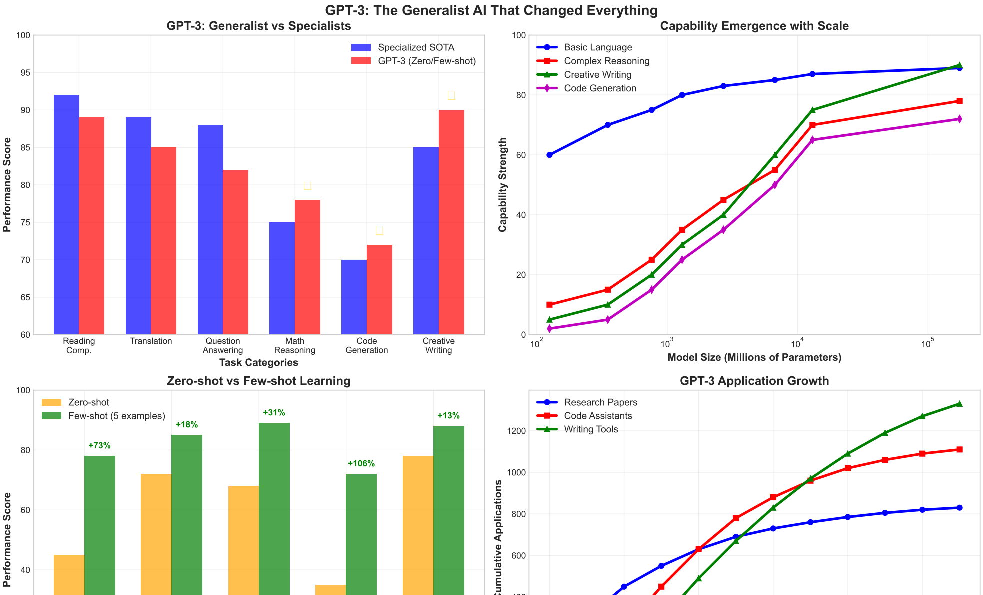
Week 7: Scaling Laws and Giant Models



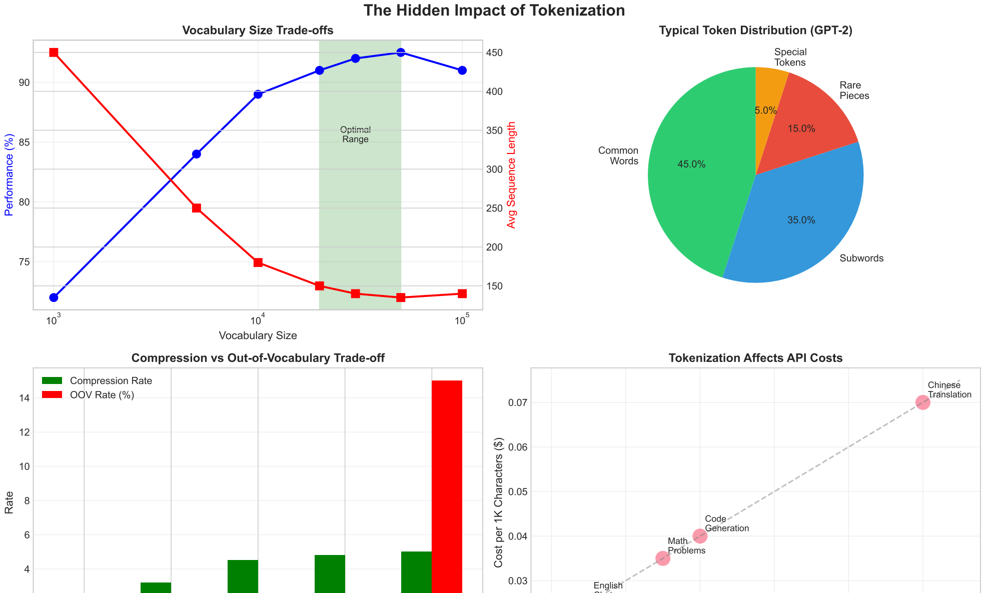
Week 7: Emergent Abilities



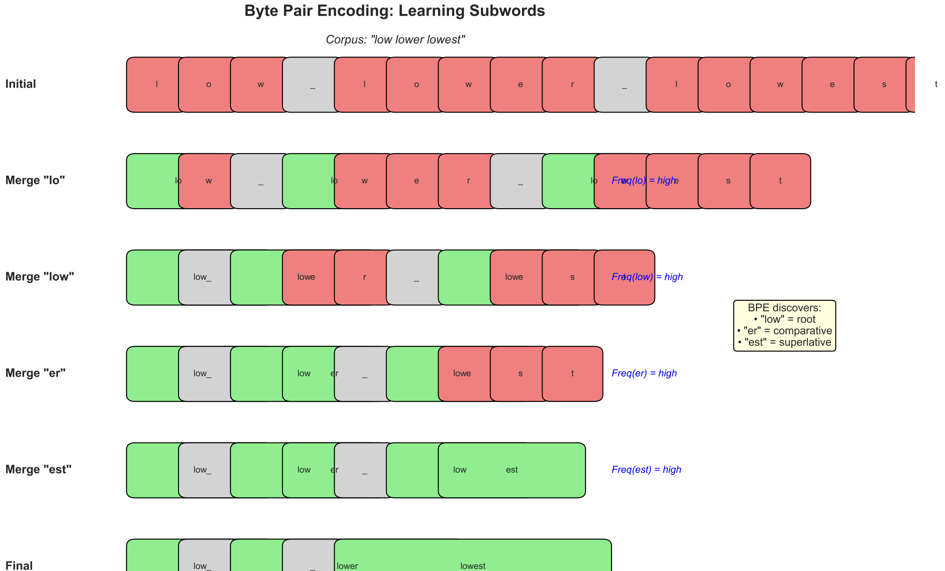
Week 7: Applications



Week 8: Tokenization Strategies

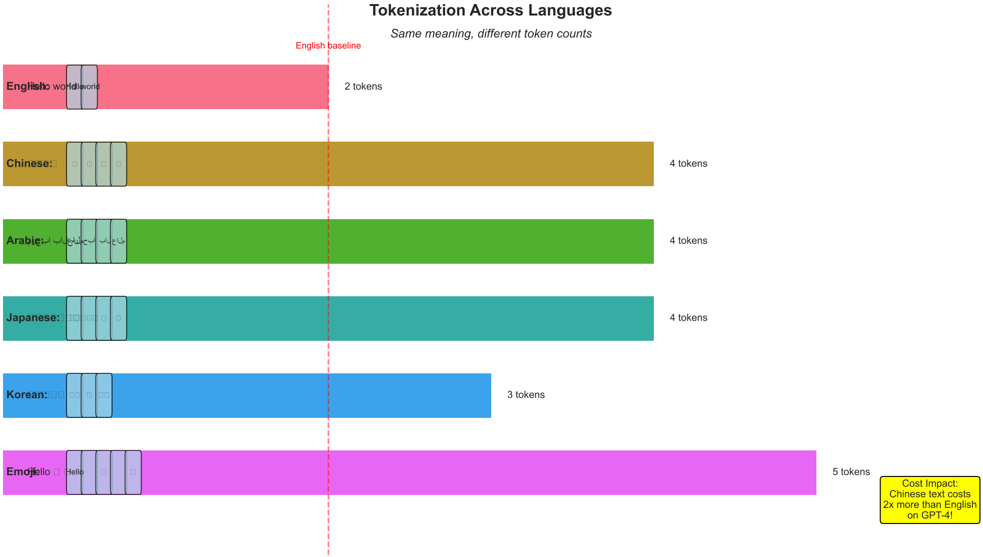


Week 8: Tokenization Methods


Byte Pair Encoding: Learning Subwords

Corpus: "low lower lowest"

Week 8: Applications



Tokenization Across Languages

Same meaning, different token counts

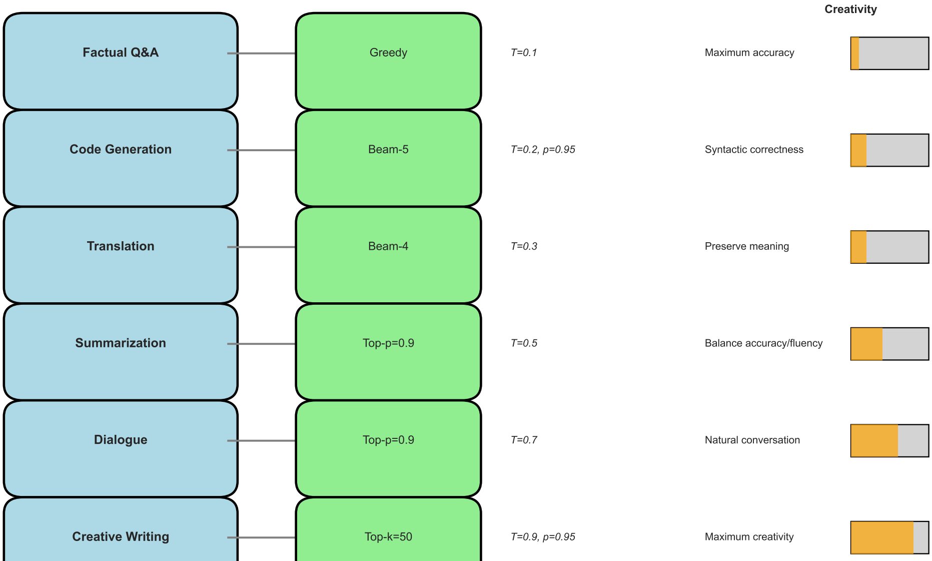
Cost Impact:
Chinese text costs
2x more than English
on GPT-4!

Use Cases:

Best Practices:

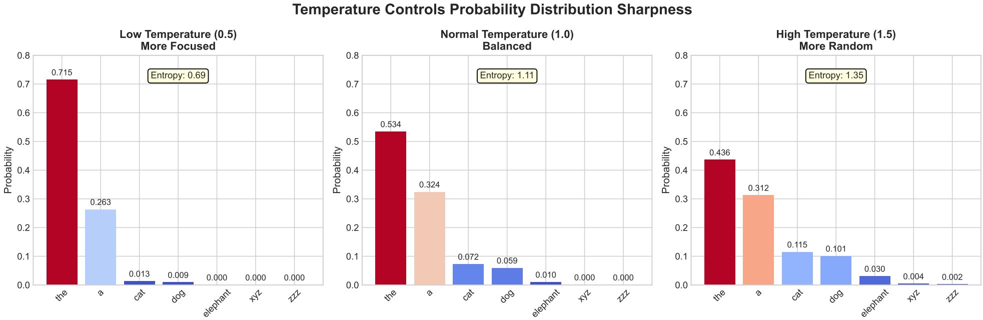
Week 9: Generation Strategies

Decoding Strategy Selection Guide



				Creativity
Factual Q&A	Greedy	*T=0.1*	Maximum accuracy	
Code Generation	Beam-5	*T=0.2, p=0.95*	Syntactic correctness	
Translation	Beam-4	*T=0.3*	Preserve meaning	
Summarization	Top-p=0.9	*T=0.5*	Balance accuracy/fluency	
Dialogue	Top-p=0.9	*T=0.7*	Natural conversation	
Creative Writing	Top-k=50	*T=0.9, p=0.95*	Maximum creativity	

Week 9: Decoding Methods



Greedy

- Fastest
- Deterministic
- Can be repetitive

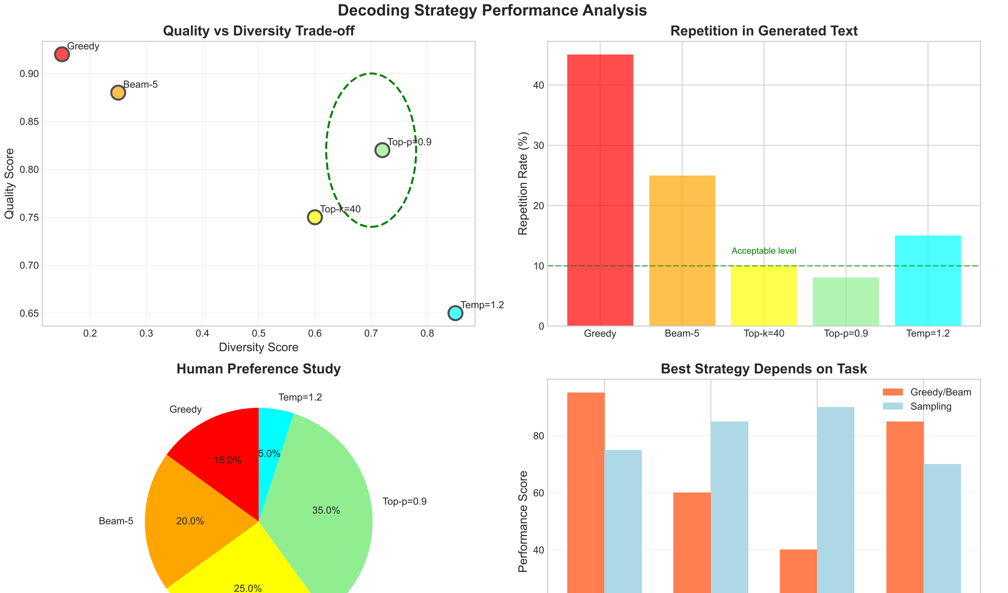
Beam Search

- Better quality
- Multiple hypotheses
- More compute

Sampling

- Creative
- Diverse outputs
- Temperature control

Week 9: Applications



Decoding Strategy Performance Analysis

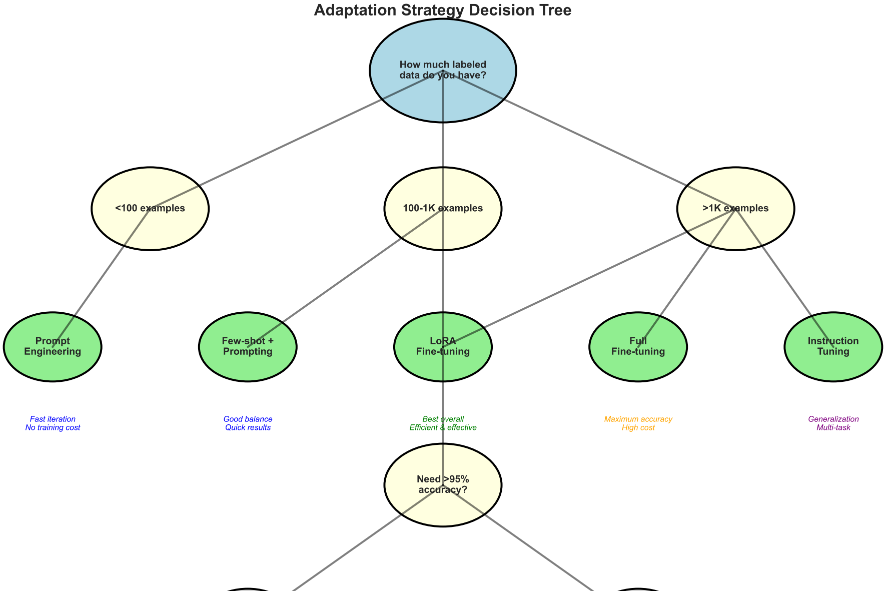
Quality vs Diversity Trade-off

Repetition in Generated Text

Human Preference Study

Best Strategy Depends on Task

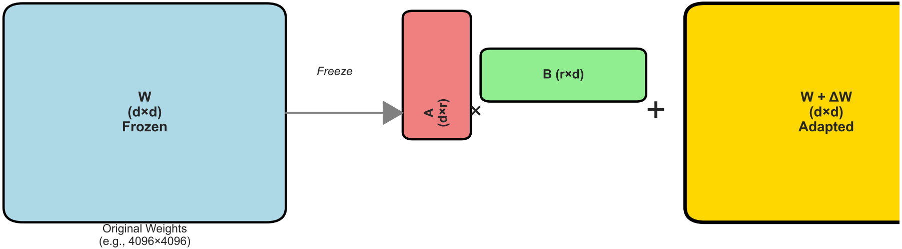
Week 10: Adaptation Techniques



Week 10: Efficient Fine-tuning

LoRA: Low-Rank Adaptation

Instead of updating 16M parameters, update only 32K!



W
(d×d)
Frozen

Original Weights
(e.g., 4096×4096)

Freeze

A
(d×r)

×

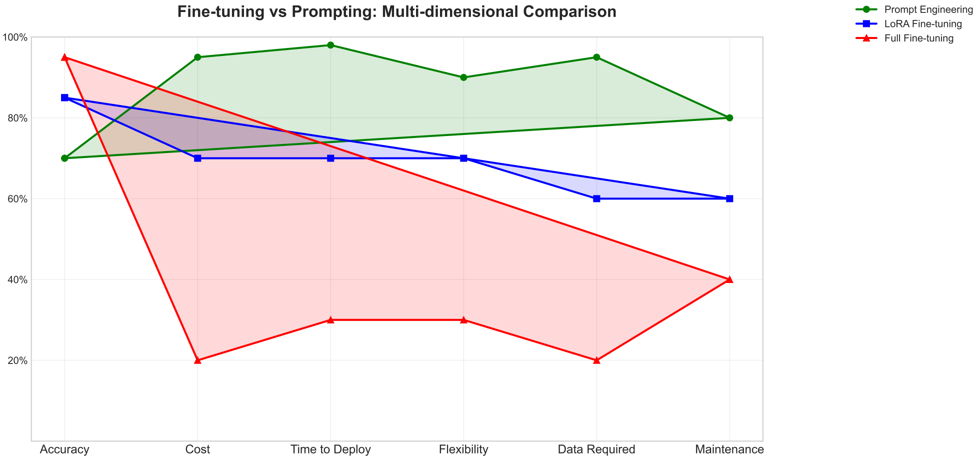
B (r×d)

+

W + ΔW
(d×d)
Adapted

Example: d=4096, r=8
Original: 4096×4096 = 16,777,216 parameters
LoRA: (4096×8) + (8×4096) = 65,536 parameters (0.39%!)

Week 10: Applications



Prompting excels at flexibility and speed

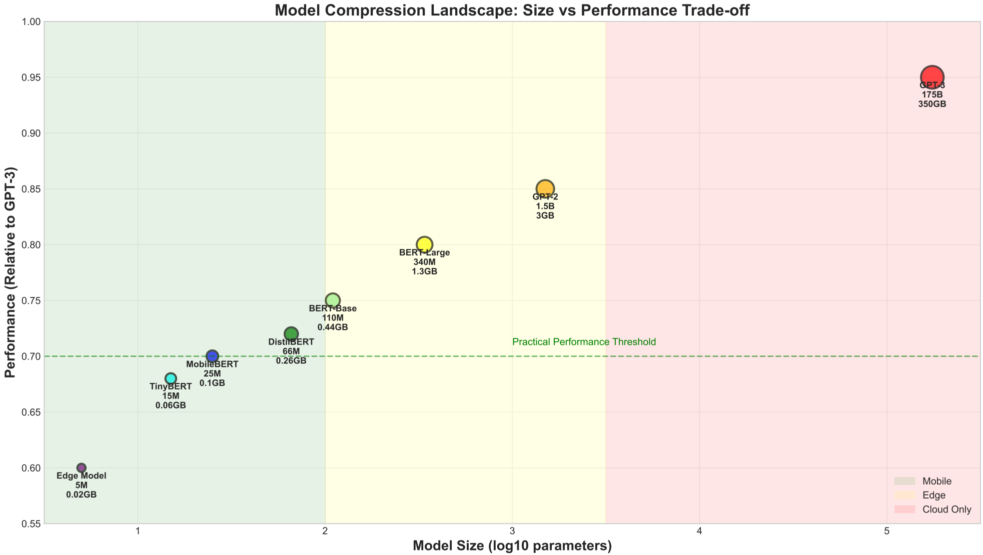
LoRA balances performance and efficiency

Full fine-tuning for maximum accuracy

Domains:

Best Practices:

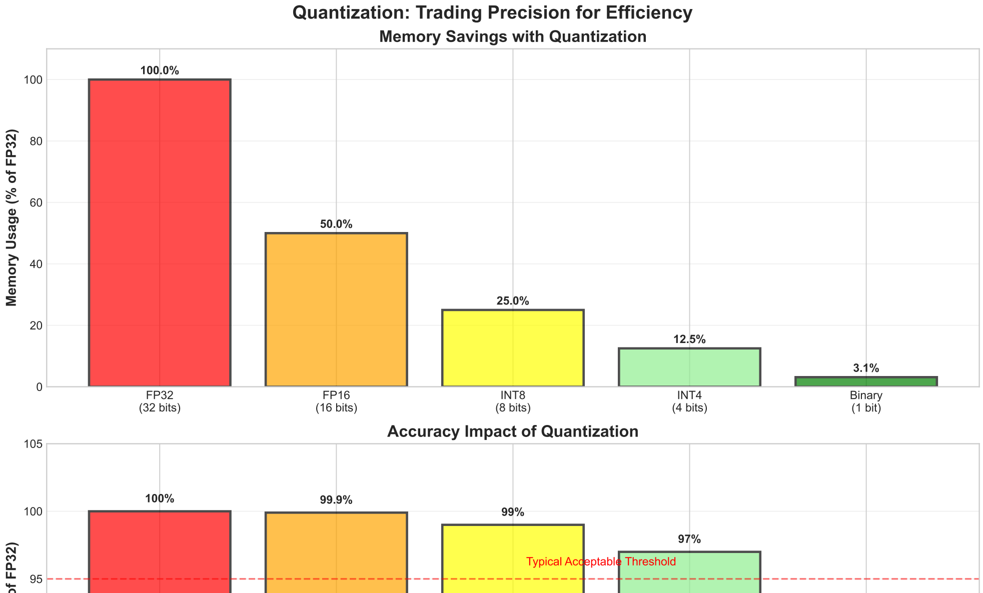
Week 11: Model Optimization



What We'll Learn:

Goals:

Week 11: Optimization Techniques

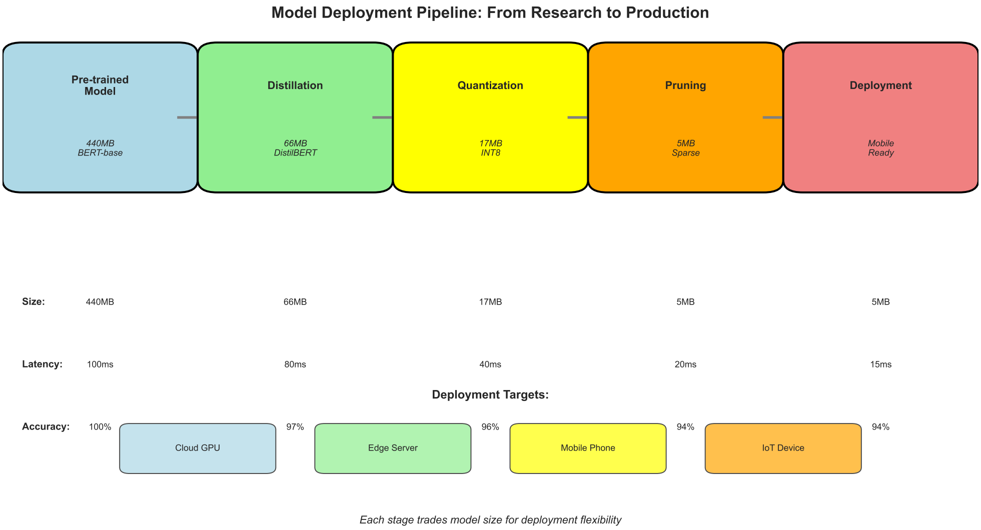


Quantization: Trading Precision for Efficiency

Memory Savings with Quantization

Accuracy Impact of Quantization

Week 11: Deployment



Model Deployment Pipeline: From Research to Production

Size:	440MB	66MB	17MB	5MB	5MB
Latency:	100ms	80ms	40ms	20ms	15ms
Accuracy:	100%	97%	96%	94%	94%

Deployment Targets:

Each stage trades model size for deployment flexibility

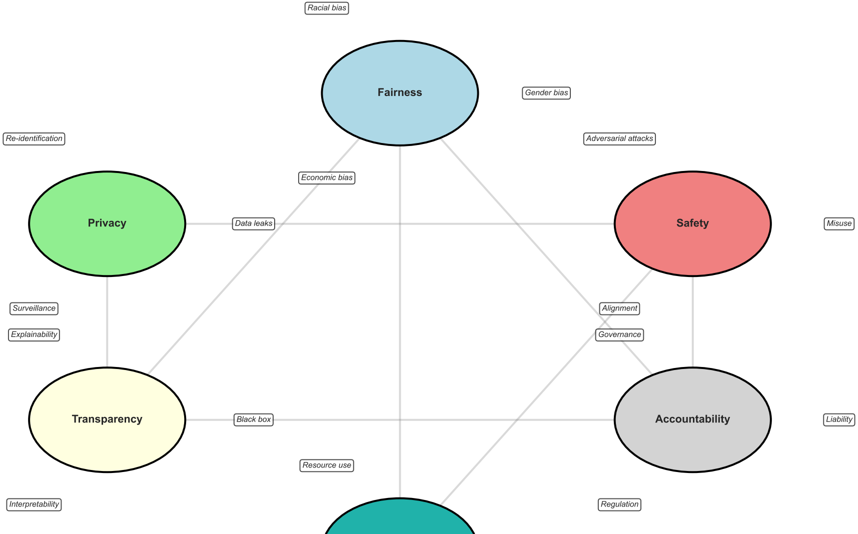
Platforms:

Metrics:

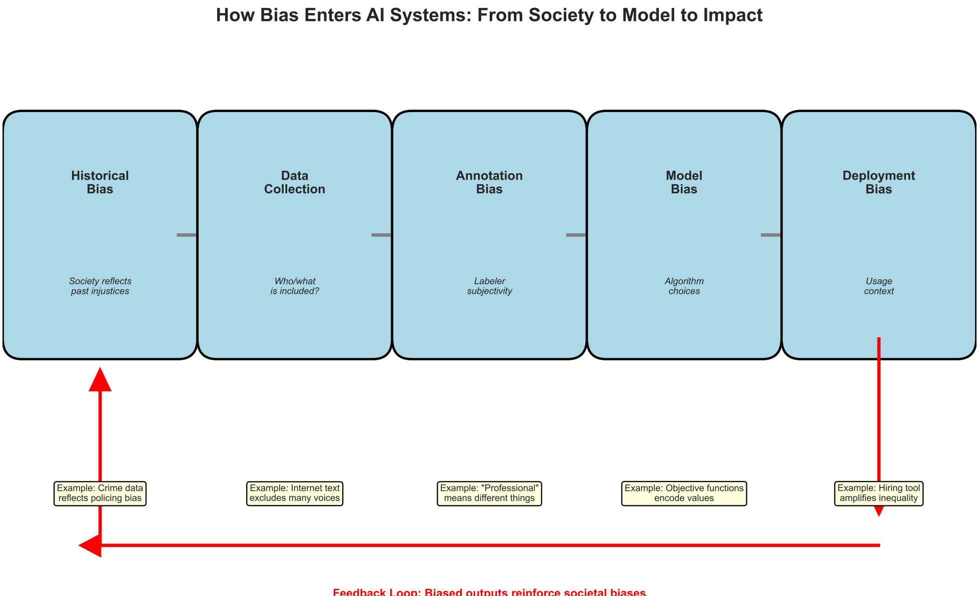
Week 12: Responsible AI

The AI Ethics Landscape: Interconnected Challenges

Each ethical dimension affects and is affected by others



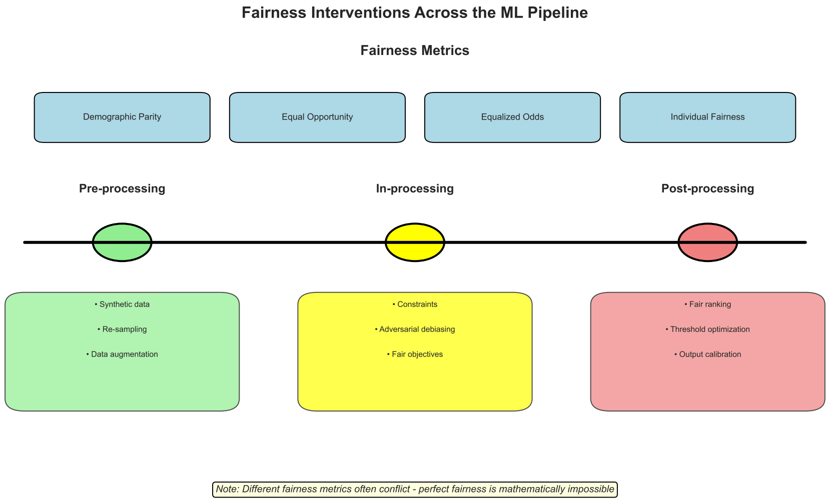
Week 12: Bias and Fairness



How Bias Enters AI Systems: From Society to Model to Impact

Historical Bias
Society reflects past injustices
Data Collection
Who/what is included?
Annotation Bias
Labeler subjectivity
Model Bias
Algorithm choices
Deployment Bias
Usage context
Example: Crime data reflects policing bias
Example: Internet text excludes many voices
Example: "Professional" means different things
Example: Objective functions encode values
Example: Hiring tool amplifies inequality
Feedback Loop: Biased outputs reinforce societal biases

Week 12: Future Directions



Fairness Interventions Across the ML Pipeline

Fairness Metrics

- Demographic Parity
- Equal Opportunity
- Equalized Odds
- Individual Fairness

Pre-processing

- Synthetic data
- Re-sampling
- Data augmentation

In-processing

- Constraints
- Adversarial debiasing
- Fair objectives

Post-processing

- Fair ranking
- Threshold optimization
- Output calibration

Note: Different fairness metrics often conflict - perfect fairness is mathematically impossible

Technical:

Societal: