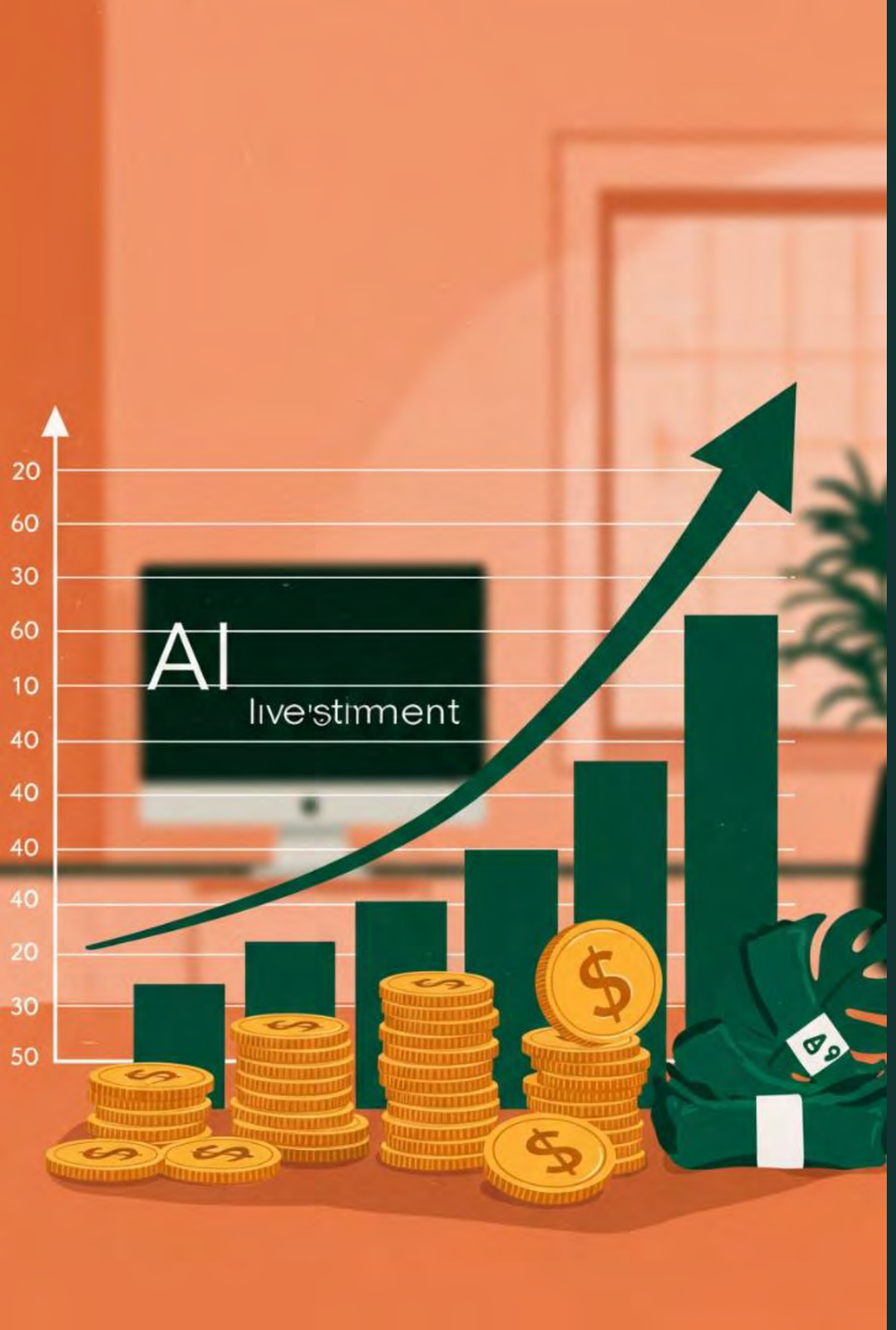
Data-Driven NLP: Quantifying the Revolution from Statistics to Generative AI

Natural Language Processing has transformed how machines understand human language

IP strategy and technology leader with expertise in patent intelligence, data science, and litigation support

- **Director at UnitedLex**:
  - Lead the IP Data Science Team
  - Innovate at the intersection of intellectual property and technology
- **Creator of Vantage for IP**:
  - Patent intelligence platform using natural language processing
  - Analyzes 100,000+ patent assets
  - Identifies competitive threats, assesses patent strength, and uncovers licensing opportunities
- **Technical skills**:
  - Develop deep learning models for patent metrics
  - Design user-friendly platforms for patent analytics



NLP Investment Growth

$91.9B

2022 Investment

Private capital in AI

300,000×

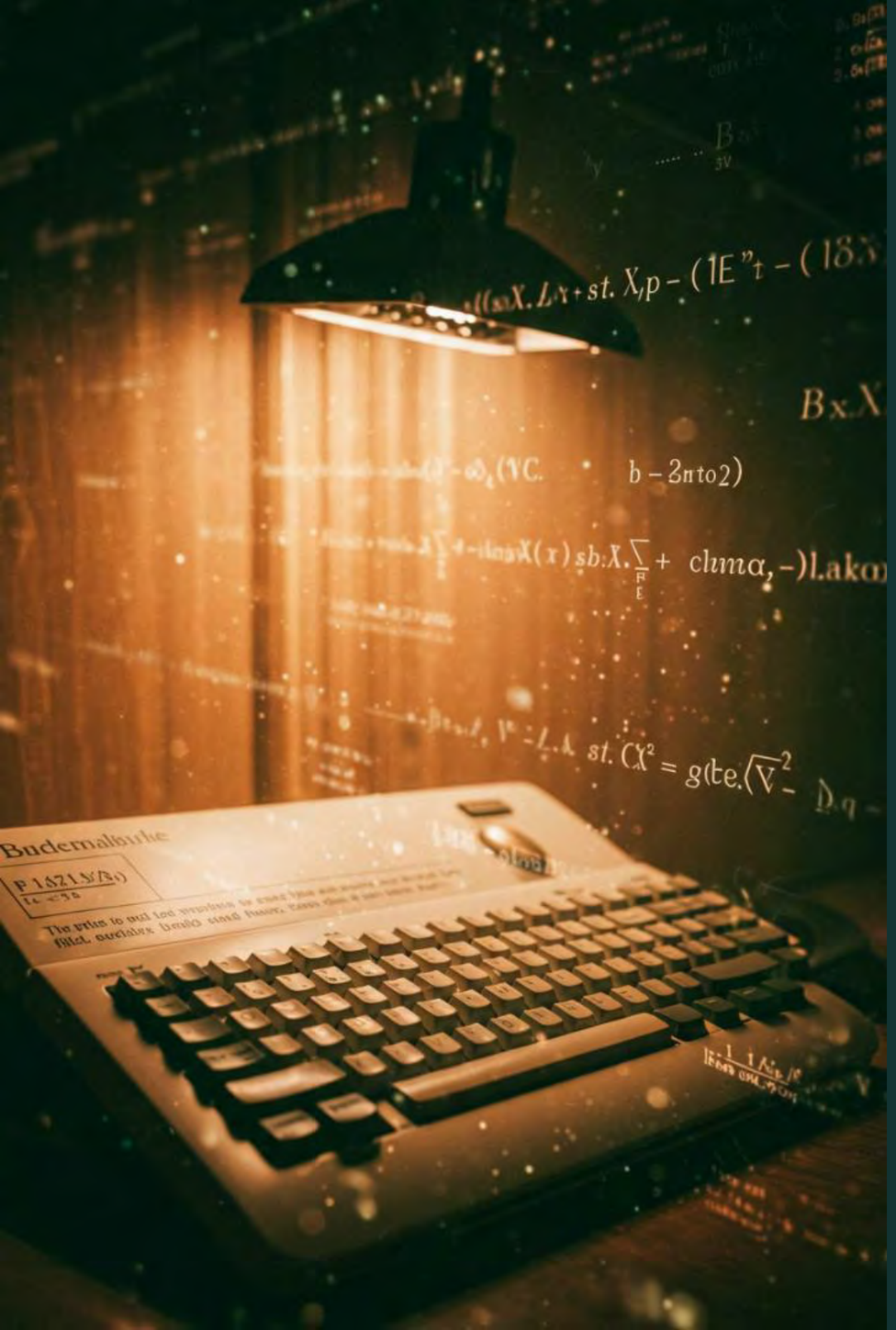
10-Year Growth

Computational demands increase

5TB

Training Data

Modern models vs. 200GB in 2019



Statistical Foundations

Bag of Words

78.2% accuracy in document classification

TF-IDF

31.2% retrieval improvement over simpler methods

N-gram Models

137+ perplexity score on Penn Treebank

The Neural Revolution

Neural Networks

Breakthrough in pattern recognition

Word Embeddings

Captured semantic relationships between words

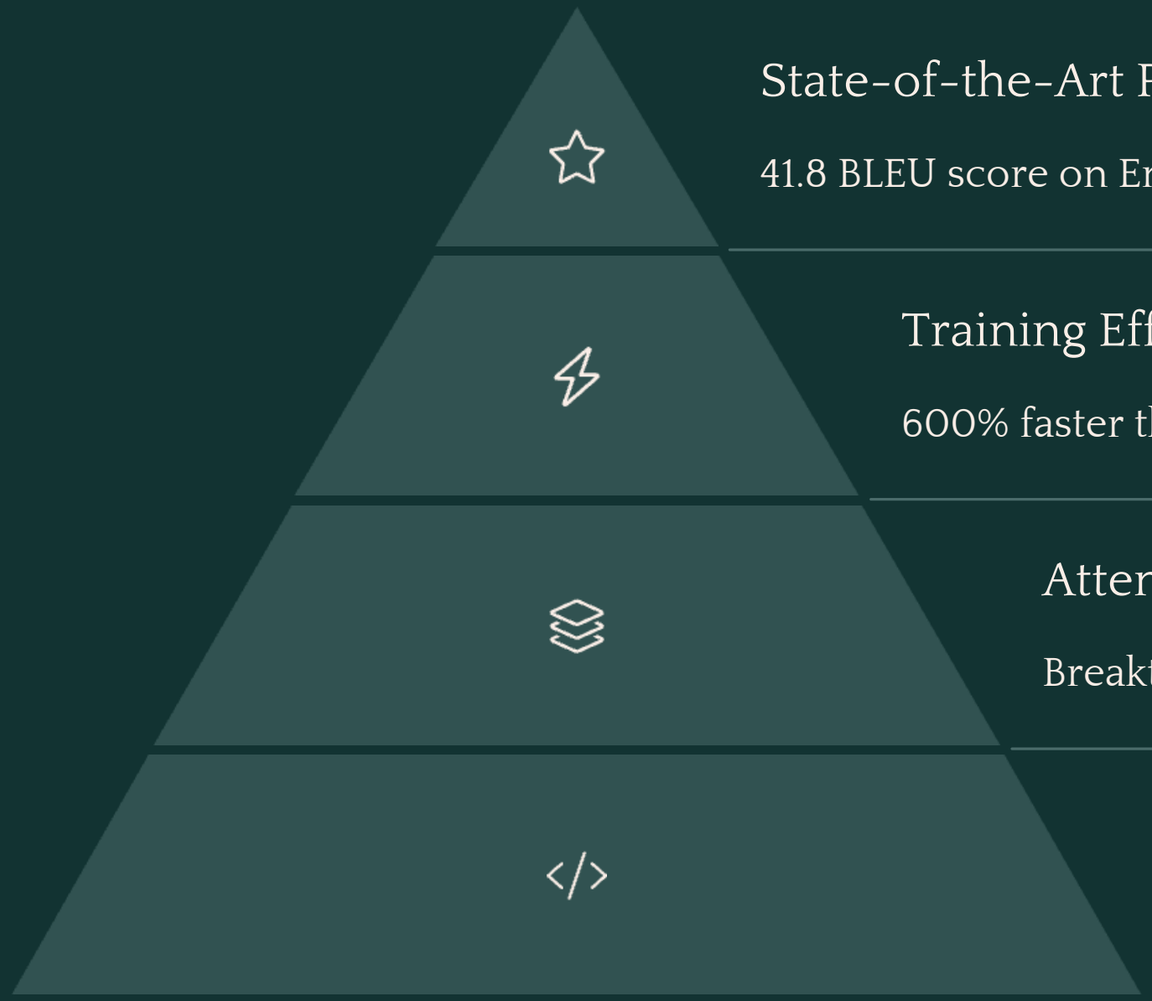
Recurrent Neural Networks

Reduced perplexity to 73.4 on Penn Treebank

Long Short-Term Memory

Solved vanishing gradient problem. Achieved perplexity score of 44.8

The Transformer Revolution



State-of-the-Art Performance

41.8 BLEU score on English-French translation

Training Efficiency

600% faster than previous architectures

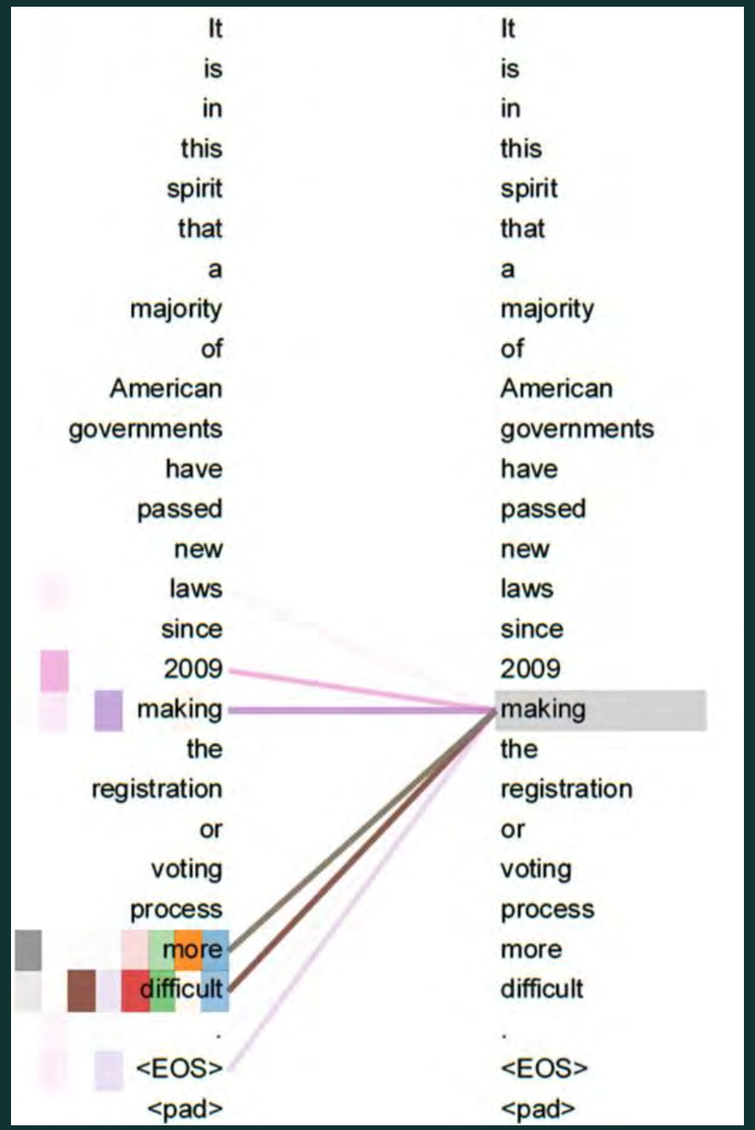
Attention Mechanism

Breakthrough for understanding context

Parallelization

Overcame sequential limitations of RNNs

The Transformer Revolution

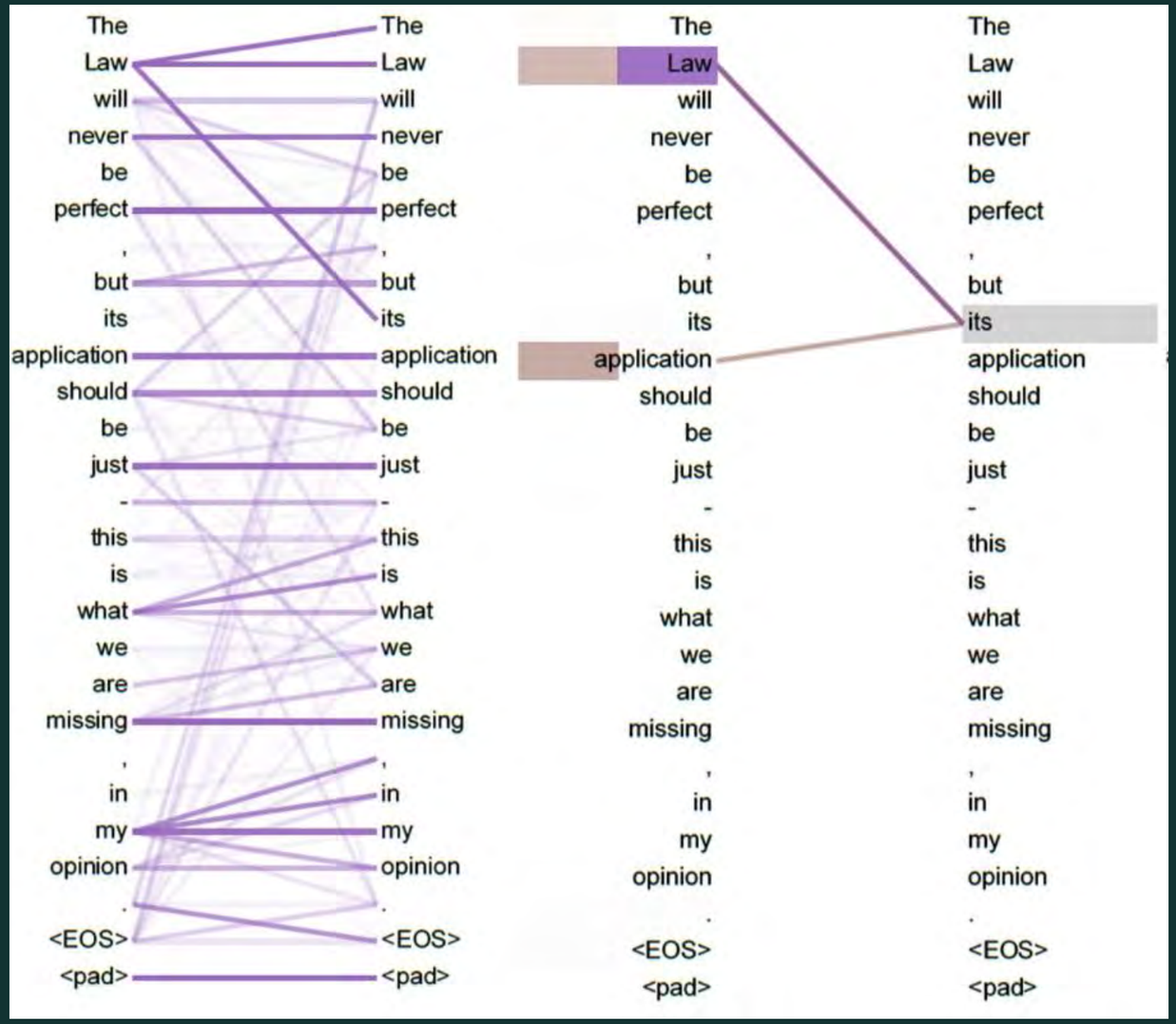


An example of the attention mechanism following long-distance dependencies from the 2017 groundbreaking paper Attention is All You Need.

Attentions here shown only for the word ‘making’. Different colors represent different heads.

Many of the attention heads attend to a distant dependency of the verb ‘making’, completing the phrase ‘making...more difficult’.

The Transformer Revolution



Example of two attention heads involved in anaphora resolution from the 2017 paper

Left: Full attentions for head 5.

Right: Isolated attentions from just the word ‘its’ for attention heads 5 and 6.

BERT's Innovations

94.9% Accuracy

Record-breaking performance on sentiment analysis benchmarks, outperforming previous state-of-the-art models by 7.2%

Transfer Learning

Efficient adaptation to downstream tasks with minimal data, reducing training requirements by up to 60%

Bidirectional Context

Revolutionary approach capturing semantic relationships in both directions, unlike previous unidirectional models

Pre-training

Novel masked language modeling technique enabling deeper contextual understanding across 3.3 billion words

GPT's Scaling Journey

GPT-1 125M Parameters

45.2% on SuperGLUE benchmark, establishing the foundation for larger language models with limited contextual understanding

GPT-2 13B Parameters

Significant intermediate performance gains, demonstrating the benefits of scale with 63.4% accuracy on complex reasoning tasks

GPT-3 175B Parameters

71.8% on SuperGLUE benchmark, achieving near-human performance across multiple reasoning tasks

Multimodal Applications

Healthcare Diagnostics

Improved accuracy when combining image and text data in clinical settings.

Multi-modal assessments outperformed human diagnoses in 80%+ cases.

30–40% improvement in workflow efficiency

Visual Question Answering

Models achieve 84.3% accuracy in understanding images and responding to natural language queries. Applications span from educational platforms to accessibility tools for vision-impaired users.

Content Understanding

Improved video content understanding. VideoChat2 achieved 51.1% MVBench

Accurate, high-quality movie translation

Closed caption generation

Faster retrieval of relevant content

Efficiency Techniques

Model Compression

75% compression with minimal performance loss. Enables deployment on edge devices with limited resources.

Knowledge Distillation

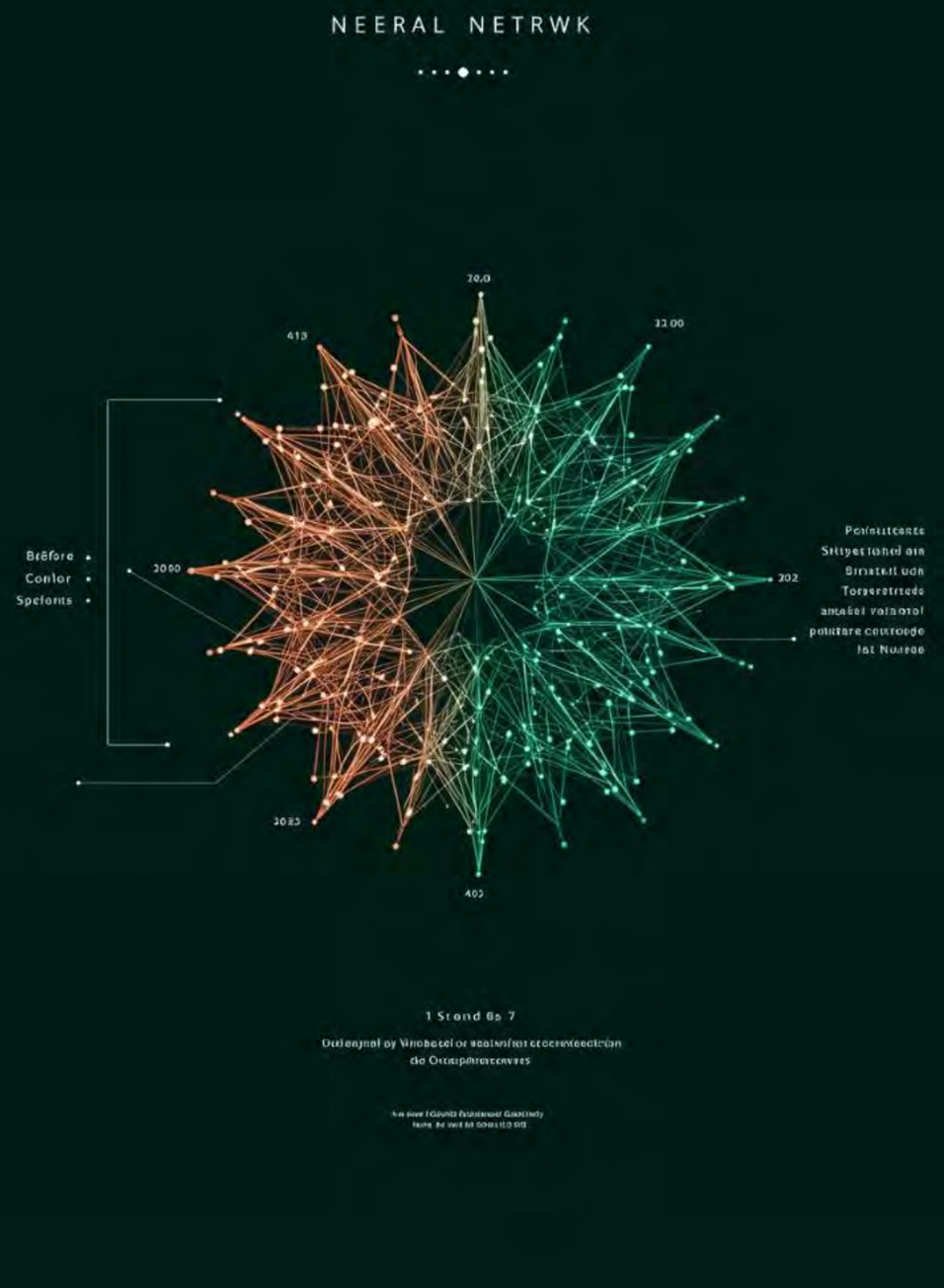
Smaller models learn from larger ones. Preserves capabilities while reducing computational needs.

Quantization

Reducing numerical precision of weights. Maintains accuracy while decreasing memory requirements.

Sparse Attention

Focusing only on relevant connections. Dramatically reduces computation without sacrificing quality.



Future Metrics to Watch

Training Efficiency

Energy consumption per model will be a defining metric in future. Current LLMs requiring up to 300 tons of CO_2 equivalent per training run.

Reasoning Capabilities

Emerging benchmarks will measure multi-step causal reasoning and counterfactual analysis. Current evaluations capture pattern recognition but miss most of human reasoning.

Ethical Alignment

New measurement frameworks will quantify how models reflect diverse human values across cultures. The field is shifting from pure performance metrics to balanced scorecards integrating safety, fairness, and transparency.

Thank you