

Glossary

This glossary provides definitions for key terms used throughout the NeuroAI Handbook. Terms are organized alphabetically and include cross-references to relevant chapters where the concepts are discussed in detail.

A

Action Potential

The electrical signal that neurons use to transmit information along their axons. Characterized by a rapid rise and fall in voltage caused by ion channel activity. Also called a "spike."

See: [Chapter 2 (Neuroscience Foundations)]

Activation Function

A mathematical function that determines the output of a neuron in artificial neural networks. Common examples include ReLU, sigmoid, and tanh.

See: [Chapter 2 (Neuroscience Foundations)], [Chapter 10 (Deep Learning)]

Attention Mechanism

A component in neural networks that allows the model to focus on specific parts of the input when generating output. Crucial for transformer models.

See: [Chapter 11 (Sequence Models)], [Chapter 20 (Case Studies in NeuroAI)]

Axon

The long, slender projection of a neuron that conducts electrical impulses (action potentials) away from the cell body to target cells.

See: [Chapter 2 (Neuroscience Foundations)]

B

Backpropagation

An algorithm for training artificial neural networks by calculating gradients of the loss function with

respect to the network weights, propagating backward from the output.

See: *[Chapter 10 (Deep Learning)]*

Batch Normalization

A technique used to improve the stability and performance of neural networks by normalizing the activations of each layer.

See: *[Chapter 10 (Deep Learning)]*

Basal Ganglia

A group of subcortical nuclei involved in motor control, procedural learning, and action selection. Often analogized to reinforcement learning systems.

See: *[Chapter 2 (Neuroscience Foundations)]*

C

Channel Capacity

In information theory, the maximum rate at which information can be reliably transmitted over a communication channel.

See: *[Chapter 7 (Information Theory)]*

Convolution

An operation that applies a filter to an input, producing a feature map that indicates where features are located in the input. Fundamental to convolutional neural networks.

See: *[Chapter 10 (Deep Learning)]*

Corpus

A large collection of text used for training language models.

See: *[Chapter 12 (Large Language Models)]*

D

Dendrite

Branched extensions of neural cell bodies that receive signals from other neurons at synapses.

See: *[Chapter 2 (Neuroscience Foundations)]*

Deep Learning

A subset of machine learning based on artificial neural networks with multiple layers that progressively extract higher-level features from raw input.

See: *[Chapter 10 (Deep Learning)]*

Dropout

A regularization technique in neural networks where randomly selected neurons are ignored during training to prevent overfitting.

See: *[Chapter 10 (Deep Learning)]*

E

Efficient Coding Hypothesis

The hypothesis that sensory systems have evolved to efficiently represent natural stimuli by reducing redundancy and maximizing information transmission given metabolic constraints.

See: *[Chapter 7 (Information Theory)]*

Emergent Abilities

Capabilities that appear in large language models only after reaching a certain scale, not present in smaller models.

See: *[Chapter 12 (Large Language Models)]*

Entropy

A measure of uncertainty or randomness in a probability distribution, central to information theory.

See: *[Chapter 7 (Information Theory)]*

F

Fine-tuning

The process of taking a pre-trained model and adapting it to a specific task by training it further on a smaller, task-specific dataset.

See: *[Chapter 12 (Large Language Models)]*

Few-shot Learning

A learning paradigm where a model can learn new tasks or concepts from only a few examples.

See: *[Chapter 12 (Large Language Models)]*

G

Gradient Descent

An optimization algorithm that iteratively adjusts parameters to minimize a loss function by moving in the direction of steepest descent of the gradient.

See: [Chapter 10 (Deep Learning)]

Gated Recurrent Unit (GRU)

A type of recurrent neural network architecture similar to LSTMs but with a simpler structure, designed to capture dependencies in sequential data.

See: [Chapter 11 (Sequence Models)]

H

Hebbian Learning

A theory describing how synaptic connections strengthen when neurons fire together ("cells that fire together wire together").

See: [Chapter 2 (Neuroscience Foundations)]

Hippocampus

A brain structure in the medial temporal lobe critical for forming new episodic memories and spatial navigation.

See: [Chapter 2 (Neuroscience Foundations)]

I

Information Bottleneck

A framework that quantifies the tradeoff between compression (minimal representation) and prediction (preserving relevant information) in neural networks.

See: [Chapter 7 (Information Theory)]

K

KL Divergence (Kullback-Leibler Divergence)

A measure of how one probability distribution differs from a reference probability distribution.

See: *[Chapter 7 (Information Theory)]*

L

Latent Factor Analysis via Dynamical Systems (LFADS)

A deep learning method that uses recurrent neural networks to model neural population dynamics and extract meaningful low-dimensional representations from high-dimensional neural data.

See: *[Chapter 20 (Case Studies in NeuroAI)]*

LSTM (Long Short-Term Memory)

A type of recurrent neural network architecture designed to address the vanishing gradient problem and better capture long-term dependencies in sequential data.

See: *[Chapter 11 (Sequence Models)]*

Large Language Model (LLM)

Neural network models with billions to trillions of parameters, trained on vast text corpora, capable of generating human-like text and performing a wide range of language tasks.

See: *[Chapter 12 (Large Language Models)]*

M

Mutual Information

A measure of the mutual dependence between two random variables, quantifying how much information one variable contains about another.

See: *[Chapter 7 (Information Theory)]*

Multi-head Attention

An extension of the attention mechanism that runs multiple attention computations in parallel, allowing the model to focus on different parts of the input for different purposes.

See: *[Chapter 11 (Sequence Models)]*

N

Neuromorphic Computing

Computing systems that mimic the neuro-biological architectures of the brain, often implementing neural networks in hardware for efficiency.

See: [Chapter 2 (Neuroscience Foundations)]

Neurotransmitter

Chemical messengers that transmit signals across synapses from a neuron to a target cell.

See: [Chapter 2 (Neuroscience Foundations)]

O

Overfitting

When a model learns the training data too well, including noise and outliers, resulting in poor generalization to new data.

See: [Chapter 10 (Deep Learning)]

P

Parameter-Efficient Fine-Tuning (PEFT)

Techniques to adapt large language models to new tasks by updating only a small subset of parameters, saving computational resources.

See: [Chapter 12 (Large Language Models)]

Perceptron

The simplest type of artificial neuron, which computes a weighted sum of its inputs, applies a step function, and outputs the result.

See: [Chapter 2 (Neuroscience Foundations)]

PredNet

A deep learning architecture that implements hierarchical predictive coding, modeling how the brain constantly generates predictions about incoming sensory information and learns from prediction errors.

See: [Chapter 20 (Case Studies in NeuroAI)]

Predictive Coding

A neuroscience theory proposing that the brain constantly generates predictions about incoming sensory information and updates its internal models based on prediction errors.

See: *[Chapter 20 (Case Studies in NeuroAI)]*

Prioritized Experience Replay (PER)

A biologically-inspired reinforcement learning technique that preferentially revisits experiences with high learning value, paralleling how the hippocampus selectively consolidates important memories.

See: *[Chapter 20 (Case Studies in NeuroAI)]*

Prompting

The practice of crafting input text to elicit specific behaviors or responses from language models.

See: *[Chapter 12 (Large Language Models)]*

R

Recurrent Neural Network (RNN)

A class of neural networks that have connections between nodes forming a directed graph along a temporal sequence, allowing them to exhibit temporal dynamic behavior.

See: *[Chapter 11 (Sequence Models)]*

Regularization

Techniques used during model training to prevent overfitting by adding a penalty on the complexity of the model.

See: *[Chapter 10 (Deep Learning)]*

Reinforcement Learning from Human Feedback (RLHF)

A technique to align language models with human preferences by training them using human feedback.

See: *[Chapter 12 (Large Language Models)]*

S

Scaling Law

Empirical relationships showing how model performance improves as a function of model size,

dataset size, and computation.

See: *[Chapter 12 (Large Language Models)]*

Self-Attention

A mechanism where a sequence attends to itself, allowing the model to weigh the importance of different positions within the same sequence.

See: *[Chapter 11 (Sequence Models)]*

Spike-Timing-Dependent Plasticity (STDP)

A biological learning mechanism where the strength of connections between neurons is adjusted based on the relative timing of a neuron's action potentials.

See: *[Chapter 2 (Neuroscience Foundations)]*

Synapse

The junction between two neurons where signals are transmitted from one neuron to another.

See: *[Chapter 2 (Neuroscience Foundations)]*

T

Tokenization

The process of converting text into tokens (words, subwords, or characters) that can be processed by language models.

See: *[Chapter 12 (Large Language Models)]*

Transfer Learning

A machine learning technique where a model developed for one task is reused as the starting point for a model on a second task.

See: *[Chapter 10 (Deep Learning)], [Chapter 12 (Large Language Models)]*

Transformer

A neural network architecture that uses self-attention mechanisms to process sequential data, enabling parallel computation and capturing long-range dependencies effectively.

See: *[Chapter 11 (Sequence Models)], [Chapter 12 (Large Language Models)]*

V

Vanishing Gradient Problem

A difficulty encountered in training deep neural networks where gradients become extremely small during backpropagation, making learning ineffective in early layers.

See: [Chapter 10 (Deep Learning)], [Chapter 11 (Sequence Models)]

Vision Transformer (ViT)

A neural network architecture for computer vision that applies the transformer model to image processing, dividing images into patches and processing them using self-attention mechanisms, inspired by how the human visual system allocates attention.

See: [Chapter 20 (Case Studies in NeuroAI)]

Z

Zero-shot Learning

The ability of a model to perform tasks it wasn't explicitly trained on, without requiring any examples.

See: [Chapter 12 (Large Language Models)]