AI 시대 필수 용어집

Essential AI Glossary

한글 용어 101개 | 영문 용어 299개

기초 개념부터 최신 트렌드까지 완벽 정리

2026년 1월

한글 용어집 (가나다순)

총 101개 용어 수록

용어	정의
가중치 (Weight)	신경망에서 입력 신호의 중요도를 결정하는 학습 가능한 매개변수. 역전파를 통해 최적화됨
강화학습 (Reinforcement Learning)	에이전트가 환경과 상호작용하며 보상을 최대화하는 행동을 학습하는 ML 패러다임
경사 하강법 (Gradient Descent)	손실 함수를 최소화하기 위해 매개변수를 반복적으로 업데이트하는 최적화 알고리즘
과적합 (Overfitting)	모델이 훈련 데이터에 과도하게 맞춰져 새로운 데이터에 대한 일반화 성능이 저하되는 현상
교차 검증 (Cross-Validation)	데이터를 여러 부분으로 나눠 모델 성능을 평가하는 기법. K-Fold가 대표적
그래디언트 소실/폭발 (Vanishing/Exploding Gradient)	심층 신경망에서 역전파 시 그래디언트가 0에 수렴하거나 무한대로 발산하는 문제
기계 학습 (Machine Learning)	명시적 프로그래밍 없이 데이터로부터 패턴을 학습하여 예측/분류를 수행하는 AI 분야
기울기 클리핑 (Gradient Clipping)	그래디언트 폭발 방지를 위해 그래디언트 값을 특정 임계값으로 제한하는 기법
나이브 베이즈 (Naive Bayes)	베이즈 정리에 기반한 확률적 분류기. 특성 간 조건부 독립 가정
노이즈 제거 오토인코더 (Denoising Autoencoder)	입력에 노이즈를 추가한 후 원본을 복원하도록 학습하는 오토인코더 변형
다중 레이블 분류 (Multi-label Classification)	하나의 입력이 여러 클래스에 동시에 속할 수 있는 분류 문제
다층 퍼셉트론 (MLP, Multi-Layer Perceptron)	입력층, 은닉층, 출력층으로 구성된 순방향 신경망의 기본 구조
단어 임베딩 (Word Embedding)	단어를 밀집 벡터로 표현하는 기법. Word2Vec, GloVe, FastText 등
드롭아웃 (Dropout)	학습 시 무작위로 뉴런을 비활성화하여 과적합을 방지하는 정규화 기법
딥러닝 (Deep Learning)	다층 신경망을 사용한 기계학습. 비정형 데이터(이미지, 텍스트, 음성)에 강점
랜덤 포레스트 (Random Forest)	다수의 결정 트리를 앙상블하여 예측하는 알고리즘. 배깅 기법 활용
레이블 (Label)	지도학습에서 입력 데이터에 대한 정답 또는 목표값
로짓 (Logit)	확률의 로그 오즈. 분류 모델의 소프트맥스 이전 출력값
리커런트 신경망 (RNN)	순차적 데이터 처리를 위한 신경망. 이전 상태를 현재 계산에 활용
마스크 언어 모델 (MLM)	BERT의 사전학습 방식. 입력 토큰 일부를 마스킹하고 예측하도록 학습

메타 학습 (Meta-Learning)	학습하는 방법을 학습하는 접근법. Few-shot learning의 기반 기술
멀티모달 (Multimodal)	텍스트, 이미지, 오디오 등 여러 형태의 데이터를 동시에 처리하는 AI 시스템
멀티헤드 어텐션 (Multi-Head Attention)	Transformer의 핵심 구성요소. 여러 어텐션을 병렬로 수행하여 다양한 표현 학습
모델 경량화 (Model Compression)	프루닝, 양자화, 지식 증류 등을 통해 모델 크기/연산량을 줄이는 기법
모멘텀 (Momentum)	이전 그래디언트의 방향을 현재 업데이트에 반영하는 최적화 기법
미세조정 (Fine-tuning)	사전학습된 모델을 특정 태스크에 맞게 추가 학습하는 전이학습 방식
반지도 학습 (Semi-supervised Learning)	소량의 레이블 데이터와 대량의 비레이블 데이터를 함께 사용하는 학습 방식
배치 정규화 (Batch Normalization)	미니배치 단위로 활성화값을 정규화하여 학습을 안정화하는 기법
백본 (Backbone)	특징 추출을 담당하는 신경망의 기본 구조. ResNet, ViT 등
벡터 데이터베이스 (Vector Database)	고차원 임베딩 벡터의 유사도 검색에 최적화된 데이터베이스. RAG 시스템의 핵심
변분 오토인코더 (VAE)	잠재 공간을 확률 분포로 모델링하는 생성 모델. 데이터 생성 및 표현 학습에 활용
분류 (Classification)	입력 데이터를 사전 정의된 범주 중 하나로 할당하는 지도학습 태스크
분산 학습 (Distributed Training)	여러 GPU/노드에서 병렬로 모델을 학습하는 방식. 데이터/모델 병렬화
비용 함수 (Cost Function)	모델 예측과 실제값 간의 차이를 측정하는 함수. 손실 함수와 동의어
비지도 학습 (Unsupervised Learning)	레이블 없이 데이터의 내재적 구조나 패턴을 발견하는 학습 방식
사전학습 (Pre-training)	대규모 데이터로 일반적인 표현을 학습하는 초기 학습 단계. 전이학습의 기반
서포트 벡터 머신 (SVM)	최대 마진 초평면을 찾는 분류 알고리즘. 커널 트릭으로 비선형 분류 가능
생성적 적대 신경망 (GAN)	생성자와 판별자가 경쟁하며 학습하는 생성 모델. 고품질 이미지 생성에 활용
셀프 어텐션 (Self-Attention)	시퀀스 내 각 위치가 다른 모든 위치와의 관계를 계산하는 메커니즘
소프트맥스 (Softmax)	로짓을 확률 분포로 변환하는 활성화 함수. 다중 클래스 분류의 출력층에 사용
손실 함수 (Loss Function)	모델의 예측 오차를 정량화하는 함수. MSE, Cross-Entropy 등
순환 신경망 (Recurrent Neural Network)	시계열/순차 데이터 처리를 위한 신경망 구조. LSTM, GRU가 대표적
스킵 연결 (Skip Connection)	입력을 출력에 직접 더하는 연결. ResNet에서 도입, 깊은 네트워크 학습 가능하게 함
슬라이딩 윈도우 (Sliding Window)	고정 크기 윈도우를 이동하며 데이터를 처리하는 방식. CNN, 시계열 분석에 활용
시그모이드 (Sigmoid)	출력을 0~1 범위로 압축하는 활성화 함수. 이진 분류, 게이트 메커니즘에 사용
시퀀스-투-시퀀스 (Seq2Seq)	입력 시퀀스를 출력 시퀀스로 변환하는 모델 구조. 번역, 요약에 활용
신경망 (Neural Network)	뉴런을 모방한 연결 구조로 정보를 처리하는 계산 모델

심층 신경망 (DNN)	은닉층이 2개 이상인 신경망. 딥러닝의 기본 구조
아키텍처 검색 (Neural Architecture Search)	최적의 신경망 구조를 자동으로 탐색하는 AutoML 기법
앙상블 (Ensemble)	여러 모델의 예측을 결합하여 성능을 향상시키는 기법. 배깅, 부스팅, 스태킹
어텐션 메커니즘 (Attention Mechanism)	입력의 중요한 부분에 가중치를 부여하여 집중하는 메커니즘
에이전트 (Agent)	환경을 인식하고 목표 달성을 위해 자율적으로 행동하는 AI 시스템
에포크 (Epoch)	전체 훈련 데이터셋을 한 번 완전히 학습하는 단위
역전파 (Backpropagation)	출력 오차를 역방향으로 전파하여 가중치를 업데이트하는 학습 알고리즘
연합 학습 (Federated Learning)	중앙 서버에 데이터를 전송하지 않고 분산된 장치에서 모델을 학습하는 방식
오토인코더 (Autoencoder)	입력을 압축(인코딩)했다가 복원(디코딩)하도록 학습하는 비지도 신경망
온톨로지 (Ontology)	특정 도메인의 개념과 관계를 정형화한 지식 표현 체계
옵티마이저 (Optimizer)	손실 함수를 최소화하도록 가중치를 업데이트하는 알고리즘. Adam, SGD 등
원-핫 인코딩 (One-Hot Encoding)	범주형 변수를 이진 벡터로 변환하는 인코딩 방식
위치 인코딩 (Positional Encoding)	Transformer에서 토큰의 위치 정보를 주입하는 기법. 사인/코사인 함수 사용
은닉층 (Hidden Layer)	입력층과 출력층 사이에 위치한 신경망 계층
이미지 분할 (Image Segmentation)	이미지의 각 픽셀을 클래스로 분류하는 컴퓨터 비전 태스크
인공 신경망 (Artificial Neural Network)	생물학적 뉴런의 연결 구조를 모방한 기계학습 모델
임베딩 (Embedding)	고차원 이산 데이터를 저차원 연속 벡터로 표현하는 기법
자기 지도 학습 (Self-supervised Learning)	레이블 없이 데이터 자체에서 학습 신호를 생성하는 방식. 대조 학습, 마스킹 등
자연어 처리 (NLP)	인간의 언어를 컴퓨터가 이해하고 생성할 수 있게 하는 AI 분야
자연어 이해 (NLU)	텍스트의 의미와 의도를 파악하는 NLP 하위 분야
자연어 생성 (NLG)	구조화된 데이터나 의미로부터 자연스러운 텍스트를 생성하는 기술
잔차 네트워크 (ResNet)	스킵 연결을 도입하여 매우 깊은 네트워크 학습을 가능하게 한 CNN 아키텍처
잠재 공간 (Latent Space)	데이터의 압축된 표현이 존재하는 저차원 공간. 생성 모델의 핵심 개념
적대적 공격 (Adversarial Attack)	입력에 미세한 변형을 가해 모델을 오작동시키는 공격 기법
전이 학습 (Transfer Learning)	한 태스크에서 학습한 지식을 다른 태스크에 적용하는 기법
정규화 (Regularization)	과적합 방지를 위해 모델 복잡도에 페널티를 부여하는 기법. L1, L2 정규화
정밀도/재현율 (Precision/Recall)	분류 모델의 성능 지표. 정밀도는 예측의 정확성, 재현율은 실제 양성 탐지율

제로샷 학습 (Zero-shot Learning)	학습 시 보지 못한 클래스에 대해 추론하는 능력
지도 학습 (Supervised Learning)	레이블이 있는 데이터로 입력-출력 관계를 학습하는 방식
지식 그래프 (Knowledge Graph)	엔티티와 관계를 그래프 구조로 표현한 지식 베이스
지식 증류 (Knowledge Distillation)	대형 모델(교사)의 지식을 소형 모델(학생)로 전달하는 모델 압축 기법
차원 축소 (Dimensionality Reduction)	고차원 데이터를 저차원으로 변환하는 기법. PCA, t-SNE, UMAP
청킹 (Chunking)	긴 문서를 작은 단위로 분할하는 RAG 전처리 기법
초매개변수 (Hyperparameter)	학습 전에 설정하는 매개변수. 학습률, 배치 크기, 레이어 수 등
추론 (Inference)	학습된 모델로 새로운 입력에 대해 예측을 수행하는 과정
컨볼루션 신경망 (CNN)	합성곱 연산을 사용하여 공간적 패턴을 학습하는 신경망. 이미지 처리에 강점
컨텍스트 윈도우 (Context Window)	LLM이 한 번에 처리할 수 있는 토큰의 최대 길이
클러스터링 (Clustering)	데이터를 유사성 기반으로 그룹화하는 비지도 학습 태스크. K-means, DBSCAN
토큰 (Token)	텍스트를 처리하는 기본 단위. 단어, 서브워드, 문자 등
토크나이저 (Tokenizer)	텍스트를 토큰으로 분할하는 도구. BPE, WordPiece, SentencePiece
특성 공학 (Feature Engineering)	원시 데이터에서 모델 성능을 향상시키는 특성을 추출/생성하는 과정
파라미터 효율적 미세조정 (PEFT)	전체 모델이 아닌 일부 파라미터만 학습하는 효율적인 미세조정 기법. LoRA, QLoRA
퍼플렉시티 (Perplexity)	언어 모델의 예측 불확실성을 측정하는 지표. 낮을수록 좋음
편향 (Bias)	뉴런의 활성화 임계값을 조절하는 학습 가능한 매개변수
풀링 (Pooling)	CNN에서 특징 맵의 크기를 줄이는 다운샘플링 연산. Max/Average Pooling
프롬프트 엔지니어링 (Prompt Engineering)	LLM에서 원하는 출력을 얻기 위해 입력 프롬프트를 설계하는 기술
피처 (Feature)	모델 입력으로 사용되는 데이터의 개별 속성이나 특성
하이퍼파라미터 튜닝 (Hyperparameter Tuning)	최적의 초매개변수 조합을 찾는 과정. Grid Search, Random Search, Bayesian Optimization
학습률 (Learning Rate)	가중치 업데이트의 크기를 결정하는 하이퍼파라미터
할루시네이션 (Hallucination)	LLM이 사실이 아닌 정보를 생성하는 현상
합성곱 (Convolution)	필터를 입력에 슬라이딩하며 특징을 추출하는 연산
활성화 함수 (Activation Function)	뉴런의 출력에 비선형성을 부여하는 함수. ReLU, Sigmoid, Tanh
회귀 (Regression)	연속적인 값을 예측하는 지도학습 태스크

효율적 어텐션 (Efficient Attention)	어텐션의 $O(n^2)$ 복잡도를 줄이는 기법. Flash Attention, Linear Attention

English Glossary (A-Z)

Total 299 terms included

Term	Definition
A/B Testing	Comparing two model versions on live traffic to determine which performs better
Ablation Study	Systematic removal of model components to understand their contribution to performance
Activation Function	Non-linear function applied to neuron output. Common types: ReLU, GELU, SiLU, Swish
Activation Maximization	Technique to visualize what patterns a neuron responds to by optimizing input
Adam Optimizer	Adaptive learning rate optimizer combining momentum and RMSprop. Default choice for deep learning
Adversarial Training	Training on adversarial examples to improve model robustness against attacks
Agent	AI system that perceives environment and takes autonomous actions to achieve goals
Agentic AI	AI systems capable of autonomous planning, tool use, and multi-step reasoning to complete tasks
AI Alignment	Research field ensuring AI systems behave according to human values and intentions
AI Safety	Research area focused on preventing harmful AI behaviors and ensuring beneficial outcomes
Anchor Loss	Contrastive learning loss that pulls similar samples together and pushes dissimilar apart
Annotation	Process of labeling data for supervised learning. Can be manual or semi-automated
Anomaly Detection	Identifying rare patterns or outliers that deviate from expected behavior
Attention Mechanism	Mechanism allowing models to focus on relevant parts of input when producing output
Autoencoder (AE)	Neural network that learns compressed representation by encoding then decoding input
AutoML	Automated machine learning - automating model selection, hyperparameter tuning, and architecture design
Autoregressive Model	Model that generates output sequentially, conditioning each token on previous ones. GPT-style
Backpropagation	Algorithm for computing gradients by propagating error backward through network

Bag of Words (BoW)	Text representation counting word occurrences, ignoring order and grammar
Batch Normalization	Normalizing layer inputs within mini-batch to stabilize and accelerate training
Batch Size	Number of training samples processed before updating model parameters
Bayesian Neural Network	Neural network with probability distributions over weights, enabling uncertainty estimation
Beam Search	Search algorithm maintaining top-k candidates at each step for sequence generation
BERT (Bidirectional Encoder Representations)	Bidirectional transformer pre-trained on masked language modeling. Foundation for NLU tasks
Bias-Variance Tradeoff	Balance between model simplicity (high bias) and complexity (high variance)
Bidirectional LSTM	LSTM processing sequences in both forward and backward directions
Binary Cross-Entropy	Loss function for binary classification measuring divergence from true labels
BLEU Score	Metric for evaluating machine translation quality based on n-gram precision
Catastrophic Forgetting	Phenomenon where learning new tasks causes model to forget previously learned tasks
Causal Attention	Attention variant where each position only attends to previous positions. Used in decoder models
Chain-of-Thought (CoT)	Prompting technique encouraging LLMs to show step-by-step reasoning
Checkpoint	Saved model state during training, enabling resumption or evaluation
Class Imbalance	Dataset condition where some classes have significantly more samples than others
Classification	Supervised learning task of assigning inputs to predefined categories
CLIP (Contrastive Language-Image Pre-training)	Model learning joint image-text representations through contrastive learning
Clustering	Unsupervised learning task of grouping similar data points together
Codeformer	Face restoration model using transformer architecture for blind face restoration
Compute Optimal Scaling	Research on optimal allocation of compute between model size and training data (Chinchilla)
Concept Drift	Change in statistical properties of target variable over time, degrading model performance
Conditional Generation	Generating output conditioned on specific input or constraint
Confusion Matrix	Table showing true/false positives/negatives for classification evaluation
Constitutional AI (CAI)	Training approach using principles to guide AI behavior, developed by Anthropic
Contrastive Learning	Self-supervised method learning representations by contrasting positive and negative pairs
Convolutional Neural	Neural network using convolution operations, excelling at spatial pattern

Network (CNN)	recognition
Cross-Attention	Attention between two different sequences. Used in encoder-decoder models
Cross-Entropy Loss	Loss function measuring divergence between predicted and true probability distributions
CUDA	NVIDIA's parallel computing platform enabling GPU acceleration for deep learning
Curriculum Learning	Training strategy presenting samples in order of increasing difficulty
Data Augmentation	Artificially expanding training data through transformations (rotation, cropping, noise)
Data Leakage	Information from outside training set inappropriately influencing model, causing overfitting
Data Parallelism	Distributed training strategy splitting data across multiple devices
Decision Boundary	Hypersurface separating different classes in feature space
Decoder	Network component generating output from encoded representation. Used in seq2seq models
Deep Q-Network (DQN)	Reinforcement learning algorithm combining Q-learning with deep neural networks
DeepSpeed	Microsoft's distributed training library with ZeRO optimization for large models
Denoising Diffusion (DDPM)	Generative model learning to reverse gradual noise addition process
Dense Retrieval	Information retrieval using learned dense embeddings rather than sparse keyword matching
Depth-wise Convolution	Efficient convolution applying separate filter per input channel. Used in MobileNet
Differentiable Programming	Programming paradigm where all operations are differentiable for gradient-based optimization
Diffusion Model	Generative model learning to denoise data through iterative refinement. Stable Diffusion, DALL-E
Direct Preference Optimization (DPO)	Simpler alternative to RLHF, directly optimizing policy from preference data
Discriminator	GAN component that distinguishes real from generated samples
Distillation	Transferring knowledge from large teacher model to smaller student model
Distributed Training	Training models across multiple GPUs or nodes for scalability
Domain Adaptation	Adapting model trained on source domain to perform well on different target domain
Dropout	Regularization randomly deactivating neurons during training to prevent overfitting
Early Stopping	Halting training when validation performance stops improving to prevent

	overfitting
Edge AI	Running AI models on edge devices (phones, IoT) rather than cloud servers
Embedding	Dense vector representation of discrete data (words, entities) in continuous space
Emergent Abilities	Capabilities that appear in large models but are absent in smaller ones
Encoder	Network component that compresses input into latent representation
Encoder-Decoder Architecture	Model with encoder processing input and decoder generating output. T5, BART
Ensemble Learning	Combining multiple models to improve prediction accuracy and robustness
Entity Recognition (NER)	NLP task identifying and classifying named entities in text
Epoch	One complete pass through the entire training dataset
Explainability (XAI)	Methods for understanding and interpreting model decisions. SHAP, LIME, attention visualization
Exploration vs Exploitation	Reinforcement learning dilemma between trying new actions and using known good ones
F1 Score	Harmonic mean of precision and recall, balancing both metrics
Feature Extraction	Deriving informative attributes from raw data for model input
Feature Map	Output of convolutional layer representing detected features in input
Feature Store	Centralized repository for storing and serving ML features in production
Federated Learning	Training on decentralized data without transferring it to central server
Few-Shot Learning	Learning to perform tasks from very few examples
Fine-Tuning	Adapting pre-trained model to specific task through additional training
Flash Attention	Memory-efficient attention algorithm reducing GPU memory usage significantly
FLOPS (Floating Point Operations)	Measure of computational cost, used to quantify model training requirements
Foundation Model	Large pre-trained model serving as base for many downstream applications
Fully Connected Layer	Neural network layer where each neuron connects to all neurons in previous layer
Function Calling	LLM capability to generate structured output for invoking external tools/APIs
GAN (Generative Adversarial Network)	Generative model with competing generator and discriminator networks
Gated Recurrent Unit (GRU)	RNN variant with reset and update gates, simpler than LSTM
Gaussian Process	Non-parametric model defining probability distribution over functions
GELU (Gaussian Error Linear Unit)	Activation function combining properties of dropout and ReLU. Used in transformers

Generalization	Model's ability to perform well on unseen data, not just training data
Generative AI	AI systems that create new content (text, images, audio, video, code)
Generative Pre-trained Transformer (GPT)	Autoregressive transformer architecture for text generation. OpenAI's GPT series
Gibbs Sampling	MCMC algorithm for sampling from joint probability distribution
Gradient Accumulation	Accumulating gradients over multiple batches before updating, simulating larger batch size
Gradient Checkpoint	Trading compute for memory by recomputing activations during backward pass
Gradient Descent	Optimization algorithm minimizing loss by iteratively adjusting parameters
Graph Neural Network (GNN)	Neural network operating on graph-structured data. Node classification, link prediction
Greedy Decoding	Text generation selecting highest probability token at each step
Ground Truth	Correct/actual labels in supervised learning dataset
Group Normalization	Normalization dividing channels into groups, alternative to batch norm for small batches
Grounding	Connecting model outputs to real-world knowledge or external data sources
Guardrails	Safety mechanisms limiting AI model behavior to acceptable boundaries
Hallucination	LLM generating false or fabricated information presented as fact
Hidden State	Internal representation passed between time steps in RNNs
Hierarchical Clustering	Clustering method building tree of nested clusters (dendrogram)
High-Rank Adaptation (HiRA)	PEFT method allowing higher-rank updates than standard LoRA
Hugging Face	Platform and library ecosystem for sharing and using transformer models
Hyperparameter	Model configuration set before training (learning rate, layers, batch size)
Image Classification	Computer vision task assigning images to predefined categories
Image Segmentation	Classifying each pixel in image to specific class. Semantic, instance, panoptic
Imitation Learning	Learning by observing and mimicking expert demonstrations
In-Context Learning (ICL)	LLM ability to learn tasks from examples provided in prompt without fine-tuning
Inference	Using trained model to make predictions on new data
Information Retrieval (IR)	Finding relevant documents from large collections based on queries
Input Embedding	Converting discrete inputs (tokens) to continuous vectors
Instance Normalization	Normalization within single sample, used in style transfer
Instruction Tuning	Fine-tuning models on instruction-following datasets for better alignment

Interpretability	Degree to which humans can understand model's decision process
IoU (Intersection over Union)	Metric measuring overlap between predicted and ground truth regions
Jailbreaking	Techniques to bypass AI model safety restrictions through adversarial prompts
Knowledge Distillation	Training smaller model to mimic larger model's outputs
Knowledge Graph (KG)	Graph structure representing entities and their relationships
KV Cache	Caching key-value pairs in transformer attention for efficient autoregressive generation
Label Smoothing	Regularization technique softening hard labels to prevent overconfidence
LangChain	Framework for building applications with LLMs, including chains and agents
Language Model (LM)	Model trained to predict probability of word sequences
Large Language Model (LLM)	Language model with billions of parameters trained on massive text corpora
Latent Space	Lower-dimensional space where data representations exist
Layer Normalization	Normalizing across features within single sample, standard in transformers
Learning Rate Schedule	Strategy for adjusting learning rate during training. Warmup, cosine decay
Leaky ReLU	ReLU variant allowing small gradient for negative inputs to prevent dead neurons
Linear Probe	Evaluation method training only linear classifier on frozen representations
LLaMA	Meta's open-weight large language model family. LLaMA 2, LLaMA 3
LLMOps	Practices for deploying, monitoring, and maintaining LLM applications in production
Logistic Regression	Linear model for binary classification using sigmoid function
Long Short-Term Memory (LSTM)	RNN architecture with gates controlling information flow, handling long sequences
LoRA (Low-Rank Adaptation)	Efficient fine-tuning method adding low-rank matrices to frozen weights
Loss Function	Function measuring model's prediction error, optimized during training
Loss Landscape	Visualization of loss function across parameter space
MAE (Mean Absolute Error)	Average absolute difference between predictions and actual values
Masked Language Modeling (MLM)	Pre-training task predicting masked tokens in input sequence. BERT's approach
Max Pooling	Downsampling by selecting maximum value in each region
Mean Pooling	Aggregating embeddings by taking element-wise mean
Meta-Learning	Learning to learn - acquiring knowledge that transfers across tasks
Mini-Batch	Subset of training data used for single parameter update

Mistral	Open-weight LLM known for efficiency. Mistral 7B, Mixtral MoE
Mixed Precision Training	Using FP16/BF16 with FP32 selectively for faster training
Mixture of Experts (MoE)	Architecture with multiple expert networks, routing inputs to relevant experts
MLflow	Platform for ML lifecycle management including experiment tracking and deployment
MLOps	Practices combining ML and DevOps for reliable model deployment and monitoring
Model Card	Documentation describing model's capabilities, limitations, and intended use
Model Parallelism	Distributing model layers across multiple devices for large models
Model Pruning	Removing unnecessary weights or neurons to reduce model size
Model Quantization	Reducing numerical precision (FP32→INT8) for efficient inference
Momentum	Optimization technique using moving average of past gradients for smoother updates
Monte Carlo Dropout	Using dropout at inference for uncertainty estimation
Multi-Agent System	System with multiple AI agents collaborating or competing on tasks
Multi-Head Attention	Parallel attention mechanisms learning different aspects of relationships
Multimodal AI	AI processing and integrating multiple data types (text, images, audio)
Multi-Task Learning	Training single model on multiple related tasks simultaneously
N-gram	Contiguous sequence of n items (words, characters) from text
Named Entity Recognition (NER)	NLP task identifying and classifying named entities (people, places, organizations)
Natural Language Generation (NLG)	Generating human-readable text from structured data or meaning representations
Natural Language Processing (NLP)	AI field enabling computers to understand and generate human language
Natural Language Understanding (NLU)	Understanding meaning and intent in human language
Negative Sampling	Training technique using sampled negative examples rather than all possible negatives
Neural Architecture Search (NAS)	Automated design of neural network architectures
Neural Machine Translation (NMT)	Translating text between languages using neural networks
Neural Network	Computational model inspired by biological neurons, organized in layers
Next Token Prediction	Core LLM training objective predicting next token given context
Noise Contrastive Estimation	Training by distinguishing real data from noise samples

Normalization	Scaling data to standard range or distribution for stable training
Nucleus Sampling (Top-p)	Text generation sampling from smallest set of tokens exceeding probability p
Object Detection	Computer vision task locating and classifying objects in images
Objective Function	Function model optimizes during training (minimize loss, maximize reward)
Offline Reinforcement Learning	RL learning from fixed dataset without environment interaction
ONNX	Open format for representing deep learning models, enabling framework interoperability
Open-Source Model	Model with publicly available weights and often code. LLaMA, Mistral
Optimizer	Algorithm updating model parameters to minimize loss. Adam, AdamW, SGD
Out-of-Distribution (OOD)	Data samples significantly different from training distribution
Overfitting	Model memorizing training data, performing poorly on new data
Padding	Adding values to sequences to achieve uniform length for batch processing
Parameter-Efficient Fine-Tuning (PEFT)	Fine-tuning updating only small subset of parameters. LoRA, QLoRA, Adapters
Perplexity	Measure of language model uncertainty; lower is better
Pipeline Parallelism	Distributing model layers across devices with micro-batch pipelining
Policy Gradient	RL method directly optimizing policy through gradient ascent on expected reward
Pooling	Downsampling operation reducing spatial dimensions while preserving important features
Position Embedding	Encoding token position in sequence for transformers lacking inherent order
Post-Training Quantization (PTQ)	Quantizing model after training without retraining
Pre-Training	Initial training phase on large general dataset before task-specific fine-tuning
Precision	Fraction of positive predictions that are actually positive (TP/(TP+FP))
Prediction Head	Task-specific layer(s) on top of model backbone for specific outputs
Principal Component Analysis (PCA)	Dimensionality reduction finding orthogonal axes of maximum variance
Prompt	Input text provided to LLM to elicit desired response
Prompt Engineering	Designing effective prompts to achieve desired LLM outputs
Prompt Injection	Attack attempting to override model instructions through malicious input
Proximal Policy Optimization (PPO)	RL algorithm with clipped objective for stable policy updates
QLoRA	Quantized LoRA combining 4-bit quantization with low-rank adaptation
Quantization	Reducing numerical precision of model weights for efficiency

Quantization-Aware Training (QAT)	Training with simulated quantization for better quantized performance
Query-Key-Value (QKV)	Three projections in attention mechanism computing weighted representations
RAG (Retrieval-Augmented Generation)	Combining retrieval of relevant documents with LLM generation for grounded responses
Random Forest	Ensemble of decision trees using bagging for robust predictions
Recall	Fraction of actual positives correctly identified (TP/(TP+FN))
Receptive Field	Region of input influencing particular CNN neuron's output
Recommendation System	AI system suggesting items based on user preferences and behavior
Recurrent Neural Network (RNN)	Neural network with loops for processing sequential data
Red Teaming	Adversarial testing to identify vulnerabilities in AI systems
Regularization	Techniques preventing overfitting by constraining model complexity
Reinforcement Learning (RL)	Learning through environment interaction, maximizing cumulative reward
Reinforcement Learning from Human Feedback (RLHF)	Aligning models using human preference feedback. Key for ChatGPT, Claude
ReLU (Rectified Linear Unit)	Activation function outputting max(0, x), standard in deep learning
Representation Learning	Learning useful features/representations from raw data automatically
Residual Connection	Skip connection adding input to output, enabling deep network training
ResNet	CNN architecture with residual connections enabling very deep networks
Reward Hacking	RL agent exploiting reward function in unintended ways
Reward Model	Model predicting human preferences for RLHF training
RMSprop	Optimizer adapting learning rates using moving average of squared gradients
RNN (Recurrent Neural Network)	Network with recurrent connections for sequential data processing
ROC-AUC	Area under ROC curve, measuring classifier performance across thresholds
RoPE (Rotary Position Embedding)	Efficient position encoding using rotation matrices. Used in LLaMA, Mistral
Safety Filter	Content moderation system filtering harmful AI outputs
Sampling Temperature	Parameter controlling randomness in text generation. Higher = more diverse
Scaling Laws	Predictable relationships between model size, data, compute, and performance
Self-Attention	Attention where query, key, value all come from same sequence
Self-Supervised Learning	Learning representations from unlabeled data using pretext tasks

Semantic Search	Search based on meaning rather than keyword matching, using embeddings
Semantic Similarity	Measuring meaning similarity between texts using vector representations
Semi-Supervised Learning	Learning from both labeled and unlabeled data
Sentence Embedding	Fixed-length vector representation of entire sentence
Sentiment Analysis	NLP task determining emotional tone or opinion in text
Sequence-to-Sequence (Seq2Seq)	Model architecture mapping input sequence to output sequence
SGD (Stochastic Gradient Descent)	Gradient descent using random sample subsets for updates
SHAP (SHapley Additive exPlanations)	Explainability method using game theory to attribute feature importance
Sigmoid Function	Activation squashing output to (0,1) range, used for probabilities
Similarity Search	Finding nearest neighbors in embedding space. Used in RAG, recommendations
Skip Connection	Direct path bypassing layers, enabling gradient flow in deep networks
Softmax	Function converting logits to probability distribution summing to 1
Sparse Attention	Attention mechanism attending to subset of positions for efficiency
Speculative Decoding	Acceleration technique using small model to draft tokens verified by large model
Speech Recognition (ASR)	Converting spoken audio to text. Whisper, DeepSpeech
Speech Synthesis (TTS)	Generating spoken audio from text. ElevenLabs, Bark
Stable Diffusion	Open-source latent diffusion model for image generation
State-Space Model (SSM)	Sequence model using state space formulation. Mamba architecture
Stochastic Gradient Descent	Gradient descent with random sampling for computational efficiency
Stop Token	Special token signaling end of generated sequence
Stride	Step size for convolution or pooling operations
Structured Prediction	Predicting complex outputs with interdependent components
Style Transfer	Applying artistic style of one image to content of another
Subword Tokenization	Tokenizing into subword units. BPE, WordPiece, SentencePiece
Supervised Fine-Tuning (SFT)	Fine-tuning on labeled instruction-response pairs before RLHF
Supervised Learning	Learning from labeled data with known input-output pairs
Support Vector Machine (SVM)	Classifier finding maximum margin hyperplane between classes
Synthetic Data	Artificially generated data for training when real data is scarce

System Prompt	Hidden instructions defining AI assistant's behavior and persona
t-SNE	Dimensionality reduction for visualizing high-dimensional data in 2D/3D
Tanh	Activation function outputting values in (-1, 1) range
Temperature (Softmax)	Parameter controlling output distribution sharpness in generation
Tensor	Multi-dimensional array, fundamental data structure in deep learning
TensorFlow	Google's open-source deep learning framework
Test Set	Held-out data for final model evaluation after training complete
Text Classification	Categorizing text into predefined classes. Sentiment, spam, topic
Text Generation	Producing coherent text given prompt or context
Token	Basic unit of text processing. Word, subword, or character
Tokenizer	Tool converting text to tokens and vice versa. BPE, SentencePiece
Top-k Sampling	Text generation sampling from k highest probability tokens
Training Loop	Iterative process of forward pass, loss computation, backpropagation, update
Training Set	Data used to train model parameters
Transfer Learning	Applying knowledge from one task/domain to improve performance on another
Transformer	Architecture using self-attention for parallel sequence processing. Foundation of modern LLMs
Transformer Block	Basic unit with multi-head attention and feed-forward network
Tree of Thoughts (ToT)	Prompting technique enabling exploration of multiple reasoning paths
Triton	Language and compiler for writing efficient GPU kernels. NVIDIA's inference server
Truncation	Cutting sequences exceeding maximum length for model processing
UMAP	Dimensionality reduction preserving global and local structure better than t-SNE
Underfitting	Model too simple to capture data patterns, performing poorly on training and test
Unsupervised Learning	Learning patterns from unlabeled data without ground truth
Upsampling	Increasing spatial resolution. Transposed convolution, interpolation
Validation Set	Data for hyperparameter tuning and model selection during training
Vanishing Gradient	Gradient becoming near-zero in deep networks, preventing learning
Variational Autoencoder (VAE)	Autoencoder learning probabilistic latent space for generation
Vector Database	Database optimized for storing and searching embedding vectors. Pinecone, Weaviate

Vector Embedding	Dense numerical representation of data in continuous space
Vision Language Model (VLM)	Model processing both images and text. GPT-4V, LLaVA, Claude
Vision Transformer (ViT)	Transformer architecture applied to image patches for computer vision
Vocabulary	Set of all tokens known to tokenizer/model
Warmup	Gradually increasing learning rate at training start for stability
Weight Decay	Regularization adding penalty proportional to weight magnitudes
Weight Initialization	Strategy for setting initial parameter values. Xavier, He initialization
Weight Sharing	Using same parameters across different parts of model
Whisper	OpenAI's speech recognition model supporting multiple languages
Word2Vec	Algorithm learning word embeddings from co-occurrence patterns. Skip-gram, CBOW
XGBoost	Gradient boosting library known for performance on tabular data
Zero-Shot Learning	Performing tasks without any task-specific training examples
Zero-Shot Prompting	Using LLM without examples, relying only on instruction

AI 기술은 빠르게 진화하고 있습니다.
이 용어집이 AI 시대를 이해하는 데 도움이 되길 바랍니다.