Complete AI Research Paper Compendium

A comprehensive guide with abstracts and links to every essential AI research paper

---

📚 Phase 1: Foundational Core (Basic ML/AI Principles)

A Few Useful Things to Know About Machine Learning

Authors: Pedro Domingos (2012)
Abstract: Practical wisdom on core ML concepts, common pitfalls, and trade-offs including bias-variance dilemma. Provides an excellent, concise overview of the philosophy and practical side of machine learning.
Link: PDF

Computing Machinery and Intelligence

Authors: Alan Turing (1950)
Abstract: Introduces the Turing Test and poses the fundamental question "Can machines think?" This founding paper of Artificial Intelligence as a field explores machine intelligence through behavioral criteria.
Link: Article

The Perceptron

Authors: Frank Rosenblatt (1958)
Abstract: Introduces the perceptron, the basic computational unit of neural networks. This paper presents a probabilistic model for information storage and organization in the brain, establishing the essential building block for modern deep learning.
Link: Archive

Learning Representations by Back-Propagating Errors

Authors: Rumelhart, Hinton, & Williams (1986)
Abstract: Formalizes the Backpropagation algorithm, the core training method that made deep learning feasible by enabling efficient gradient computation through multiple layers.
Link: Nature

🧠 Phase 2: Deep Learning and Modern AI Pillars

ImageNet Classification with Deep Convolutional Neural Networks (AlexNet)

Authors: Krizhevsky, Sutskever, & Hinton (2012)
Abstract: Introduces AlexNet CNN architecture trained on ImageNet dataset using GPU acceleration. This paper kick-started the modern Deep Learning boom in Computer Vision by demonstrating dramatic improvements over traditional methods.
Link: NIPS 2012

Generative Adversarial Nets (GANs)

Authors: Ian Goodfellow et al. (2014)
Abstract: Introduces a framework for training generative models using two competing neural networks—a Generator that creates samples and a Discriminator that distinguishes real from fake. Fundamental to image generation and unsupervised learning.
Link: arXiv

Deep Residual Learning for Image Recognition (ResNet)

Authors: He et al. (2015)
Abstract: Introduces residual blocks with skip connections enabling training of extremely deep neural networks (150+ layers). Solved the vanishing gradient problem and became a standard architecture in Computer Vision.
Link: arXiv

Playing Atari with Deep Reinforcement Learning (DQN)

Authors: Mnih et al. (2013)
Abstract: Combines deep learning with Q-learning using Experience Replay and Target Networks, enabling agents to learn policies directly from high-dimensional pixel input. A foundational paper for Deep Reinforcement Learning.
Link: arXiv

Attention Is All You Need (Transformer)

Authors: Vaswani et al. (2017)
Abstract: Introduces the Transformer architecture relying solely on self-attention mechanism, eliminating recurrent neural networks. The single most important paper for modern NLP (BERT, GPT) and increasingly influential in Computer Vision.
Link: arXiv

🚀 Phase 3: Specialization - Natural Language Processing

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Authors: Devlin et al. (2018)
Abstract: Introduces pre-training of deep bidirectional Transformer encoder using Masked Language Modeling. Established pre-trained language models as the standard approach for NLP tasks.
Link: arXiv

Language Models are Unsupervised Multitask Learners (GPT-2)

Authors: Radford et al. (2019)
Abstract: Demonstrates the power of large-scale unsupervised pre-training for generative tasks. Shows language models can perform multiple tasks without task-specific training, a key stepping stone to GPT-3 and beyond.
Link: OpenAI

🚀 Phase 3: Specialization - Computer Vision

YOLO: You Only Look Once

Authors: Redmon et al. (2016)
Abstract: A real-time, unified approach for object detection that frames detection as a regression problem. Revolutionary method for faster and more efficient object detection in

a single forward pass.
Link: arXiv

Denoising Diffusion Probabilistic Models (DDPM)

Authors: Ho et al. (2020)
Abstract: Establishes foundation for modern Diffusion Models by formulating image generation as a denoising process. Technical precursor to models like DALL-E and Stable Diffusion for high-quality image generation.
Link: arXiv

💡 Phase 4: Unique Architectures Beyond Transformers

State Space Models

Mamba: Linear-Time Sequence Modeling with Selective State Spaces

Authors: Gu & Dao (2023)
Abstract: Introduces Selective State Space Model blending recurrence for efficiency with convolutions for parallel training. Achieves linear O(N) scaling in sequence length during inference while maintaining parallel training capability through dynamic information filtering.
Link: arXiv

S4: Efficiently Modeling Long Sequences with Structured State Spaces

Authors: Gu et al. (2021)
Abstract: Revives State Space Models from control theory, demonstrating ability to model extremely long-range dependencies efficiently. Provides mathematical basis for Mamba by connecting continuous-time dynamics to discrete sequence modeling.
Link: arXiv

Memory-Augmented Architectures

Titans + MIRAS: Helping AI have long-term memory

Authors: Google Research (2025)
Abstract: Introduces Titans (Memory-Augmented Controller) and MIRAS (Memory-Informed Recurrent Associative System) framework. Incorporates long-term memory module to compress past data and inject into attention mechanism for recall beyond context window.
Link: Google Research Blog

Recurrent Independent Mechanisms (RIMs)

Authors: Goyal et al. (2019)
Abstract: Proposes recurrent network with multiple nearly-independent recurrent cells communicating sparingly via bottleneck attention. Promotes modularity and specialization with only relevant mechanisms updated per time step.
Link: arXiv

Mixture-of-Experts

Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity

Authors: Fedus et al. (2021)
Abstract: Formalizes modern Mixture-of-Experts where only sparse subset of parameters (experts) activate per token. Enables building trillion-parameter models while keeping computational cost per token nearly constant.
Link: arXiv

🔬 Phase 5: Deep Dive - Large Language Models

Training Stability and Optimization

Adam: A Method for Stochastic Optimization

Authors: Kingma & Ba (2014)
Abstract: Introduces widely-used optimization algorithm combining adaptive learning rates with momentum. Essential for understanding how nearly all modern deep learning

models are trained and optimized.
Link: arXiv

Training Compute-Optimal Large Language Models (Chinchilla)

Authors: Hoffmann et al. (2022)
Abstract: Challenges Kaplan scaling laws, showing optimal model size is smaller with more training data for fixed compute budget. Directly guides current strategy for training competitive LLMs.
Link: arXiv

Efficiently Scaling Transformer Inference

Authors: Pope et al. (2022)
Abstract: Focuses on engineering optimizations (KV Caching, batching) for deploying and running inference on massive LLMs affordably. Critical for understanding cost and latency of production models.
Link: Google Research

Fine-Tuning and Alignment

Finetuned Language Models are Zero-Shot Learners (FLAN)

Authors: Wei et al. (2022)
Abstract: Shows fine-tuning LLMs on diverse task collection significantly improves zero-shot performance on unseen tasks. Formalized shift from task-specific training to general instruction-following.
Link: arXiv

Training language models to follow instructions with human feedback (InstructGPT)

Authors: Ouyang et al. (2022)
Abstract: Introduces core RLHF pipeline for aligning LLMs to human preferences using Reward Model and Proximal Policy Optimization. Foundational paper defining how ChatGPT and similar assistants are aligned.
Link: arXiv

Advanced Architectures

PaLM-E: An Embodied Multimodal Language Model

Authors: Driess et al. (2023)
Abstract: Integrates text, image, and robotic action data into single massive model capable of reasoning about physical tasks. Pushes boundary toward embodied AI where LLMs interact with physical world.
Link: arXiv

🔬 Phase 6: Cutting-Edge Topics

Large Language Models and Scaling Laws

Scaling Laws for Neural Language Models

Authors: Kaplan et al. (2020)
Abstract: Empirically derives mathematical relationships showing how loss scales predictably with model size, dataset size, and compute. Provides roadmap for building successful LLMs and resource allocation.
Link: arXiv

Language Models are Few-Shot Learners (GPT-3)

Authors: Brown et al. (2020)
Abstract: Showcases emergent ability of massive models (175B parameters) to learn tasks from prompt demonstrations without weight updates. Defined paradigm of using LLMs as general few-shot learners.
Link: arXiv

Chain-of-Thought Prompting Elicits Reasoning in Large Language Models

Authors: Wei et al. (2022)
Abstract: Shows generating intermediate reasoning steps before final answer dramatically improves performance on complex reasoning tasks. Fundamental technique in prompt

engineering for enhancing LLM capability.
Link: arXiv

Multimodal and Generative AI

Learning Transferable Visual Models From Natural Language Supervision (CLIP)

Authors: Radford et al. (2021)
Abstract: Learns visual concepts from 400M (image, text) pairs, creating effective zero-shot classifier enabling image-text alignment. Engine behind modern text-to-image models and major step toward general visual understanding.
Link: arXiv

High-Resolution Image Synthesis with Latent Diffusion Models (Stable Diffusion)

Authors: Rombach et al. (2022)
Abstract: Compresses images into lower-dimensional latent space making diffusion training faster and more efficient. Made high-quality, open-source text-to-image generation accessible.
Link: arXiv

🏛️ Phase 7: Comprehensive Architecture Landscape

Computer Vision Architectures

An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale (ViT)

Authors: Dosovitskiy et al. (2020)
Abstract: Breaks images into patches, treats them as tokens, and applies standard Transformer architecture for global understanding. Demonstrates pure attention mechanisms can match or exceed CNN performance.
Link: arXiv

U-Net: Convolutional Networks for Biomedical Image Segmentation

Authors: Ronneberger et al. (2015)
Abstract: U-shaped encoder-decoder architecture with skip connections between encoder

and decoder to maintain fine-grained spatial details. Foundational for medical image segmentation and dense prediction tasks.
Link: arXiv

Generative Models

Auto-Encoding Variational Bayes (VAE)

Authors: Kingma & Welling (2013)
Abstract: Learns smooth, continuous latent space by enforcing probabilistic structure (KL-divergence loss) on encoded distribution. Enables generation of novel samples and learned representations.
Link: arXiv

Graph Neural Networks

Semi-Supervised Classification with Graph Convolutional Networks (GCN)

Authors: Kipf & Welling (2016)
Abstract: Generalizes CNN convolution to graph structures by aggregating feature information from node's immediate neighbors. Foundation for learning on non-Euclidean data structures.
Link: arXiv

Graph Attention Networks (GAT)

Authors: Veličković et al. (2017)
Abstract: Uses attention mechanism to assign different importance levels to different neighbors, making aggregation dynamic and weighted. Handles heterogeneous graphs and noisy connections robustly.
Link: arXiv

Fundamental Innovations

Kolmogorov-Arnold Networks (KANs)

Authors: Liu et al. (2024)
Abstract: Replaces fixed activation functions in nodes with learnable univariate functions

(B-splines) on edges. Increases interpretability and accuracy compared to traditional MLPs.
Link: arXiv

Neural Ordinary Differential Equations (Neural ODEs)

Authors: Chen et al. (2018)
Abstract: Treats network layers as discretization of continuous-time system governed by ODE. Saves memory and allows solving models with variable precision using adaptive ODE solvers.
Link: arXiv

🔬 Phase 8: Graph Neural Networks Deep Dive

Key GNN Architectures

Inductive Representation Learning on Large Graphs (GraphSAGE)

Authors: Hamilton et al. (2017)
Abstract: Learns generalizable function to sample fixed-size neighborhood and aggregate features using Mean, LSTM, or Pooling. Focuses on inductive learning for unseen nodes in large-scale graphs.
Link: arXiv

How Powerful are Graph Neural Networks? (GIN)

Authors: Xu et al. (2019)
Abstract: Proves Graph Isomorphism Network is most expressive among simple Message Passing GNNs (as powerful as 1-WL test) using Sum aggregator and MLP. Optimal for graph classification tasks.
Link: arXiv

Modeling Relational Data with Graph Convolutional Networks (R-GCN)

Authors: Schlichtkrull et al. (2018)
Abstract: Designed for Knowledge Graphs with multiple edge types (relations). Aggregates features differently for each relation type, handling heterogeneous graph

structures.
Link: arXiv

🏎️ Phase 9: Reinforcement Learning Complete Guide

Mathematical Foundations

Q-Learning

Authors: Christopher Watkins (1989)
Abstract: Introduces Q(s,a) function for action-value estimation. Foundation of off-policy learning, allowing agents to learn from actions they didn't actually take through temporal difference updates.
Link: Thesis

Policy Gradient Methods for Reinforcement Learning with Function Approximation

Authors: Sutton et al. (1999)
Abstract: Shifts focus from value functions to directly parameterizing and optimizing policies. Theoretical basis for almost all modern Deep RL algorithms including PPO and A3C.
Link: NIPS 1999

Deep RL Breakthroughs

Human-level control through deep reinforcement learning

Authors: Mnih et al. (2015)
Abstract: Refined DQN achieving human-level performance across diverse Atari games. First demonstration of AI outperforming humans on visually complex tasks using only pixel input.
Link: Nature

Mastering the game of Go with deep neural networks and tree search (AlphaGo)

Authors: Silver et al. (2016)
Abstract: Combines deep RL with Monte Carlo Tree Search defeating world champion.

Landmark in AI history solving problem thought decades away through policy and value network co-training.
Link: Nature

Modern Standards

Proximal Policy Optimization Algorithms (PPO)

Authors: Schulman et al. (2017)
Abstract: Introduces clipped objective function preventing massive destructive policy updates. Industry standard for most RL tasks due to stability, simplicity, and strong empirical performance.
Link: arXiv

Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning (SAC)

Authors: Haarnoja et al. (2018)
Abstract: Introduces entropy maximization encouraging exploration in continuous action spaces. State-of-the-art for continuous control tasks especially robotics applications.
Link: arXiv

Novel RL Paradigms

Decision Transformer: Reinforcement Learning via Sequence Modeling

Authors: Chen et al. (2021)
Abstract: Casts RL as sequence modeling problem using standard Transformer predicting actions based on returns-to-go. Ignores traditional RL loop making offline training straightforward.
Link: arXiv

Mastering Diverse Domains through World Models (DreamerV3)

Authors: Hafner et al. (2023)
Abstract: Agent learns entirely inside learned latent world model achieving extreme sample efficiency. Single algorithm with fixed hyperparameters masters 150+ diverse tasks including Minecraft.
Link: arXiv

Offline Reinforcement Learning: Tutorial, Review, and Perspectives on Open Problems

Authors: Levine et al. (2020)
Abstract: Formalizes Offline RL for learning from fixed datasets without environment interaction. Crucial for real-world applications (healthcare, finance) where exploration is costly or dangerous.
Link: arXiv

LLM Alignment

Direct Preference Optimization: Your Language Model is Secretly a Reward Model (DPO)

Authors: Rafailov et al. (2023)
Abstract: Provides mathematical way to align models without complex RL loop by directly optimizing policy from preference data. Major simplification of RLHF making alignment more accessible.
Link: arXiv

🌎 Phase 10: World Models - Learning to Imagine

Foundational World Models

World Models

Authors: Ha & Schmidhuber (2018)
Abstract: Demonstrates agents can solve complex tasks entirely inside learned simulated environment combining VAE for compression and RNN with Mixture Density Network for stochastic prediction.
Link: arXiv

Dream to Control: Learning Behaviors by Latent Imagination (DreamerV1)

Authors: Hafner et al. (2020)
Abstract: Learns actor-critic policy through backpropagation of imagined trajectories in

latent space using Recurrent State-Space Model. Makes planning far more efficient than pixel-space planning.
Link: arXiv

Mastering Atari with Discrete World Models (DreamerV2)

Authors: Hafner et al. (2021)
Abstract: Extends Dreamer to discrete latent representations matching DQN performance while being more data-efficient. Demonstrates world models can match model-free methods on benchmark tasks.
Link: arXiv

Mastering Diverse Domains through World Models (DreamerV3)

Authors: Hafner et al. (2023)
Abstract: Single algorithm with fixed hyperparameters achieves state-of-the-art across 150+ diverse tasks proving generality and robustness of model-based approach at scale.
Link: arXiv

Mastering Atari, Go, Chess and Shogi by Planning with a Learned Model (MuZero)

Authors: Schrittwieser et al. (2020)
Abstract: Learns implicit world model predicting only necessary planning information (Value, Policy, Reward) without reconstructing observations. Matches AlphaZero without knowing game rules.
Link: Nature

🔍 Phase 11: Mechanistic Interpretability

Foundational Grammar

A Mathematical Framework for Transformer Circuits

Authors: Elhage et al. (2021)
Abstract: Introduces concept of Attention Heads as independent units and Residual Stream as communication bus. Foundational textbook defining vocabulary for interpretability

research.
Link: Anthropic

Toy Models of Superposition

Authors: Elhage et al. (2022)
Abstract: Proves mathematically that models store information in compressed directions leading to polysemanticity (one neuron performing multiple functions). Explains why individual neuron analysis is misleading.
Link: Anthropic

The Geometry of Truth: Emergent Linear Structure in Large Language Model Representations

Authors: Marks & Tegmark (2024)
Abstract: Shows LLMs represent concepts like "True" and "False" as linear directions in representation space. Demonstrates high-level concepts have discoverable geometric structure.
Link: arXiv

Finding Circuits

In-context Learning and Induction Heads

Authors: Olsson et al. (2022)
Abstract: Identifies Induction Heads—specific mechanism allowing LLMs to repeat patterns and learn in-context. First major success finding universal circuit present in almost all models.
Link: Anthropic

Interpretability in the Wild: a Circuit for Indirect Object Identification in GPT-2 small

Authors: Wang et al. (2022)
Abstract: Reverse-engineers circuit GPT-2 uses to identify indirect object in sentences. Masterclass in Causal Scrubbing—proving circuits by breaking pieces and observing failures.
Link: arXiv

Monosemanticity Revolution

Towards Monosemanticity: Decomposing Language Models With Dictionary Learning

Authors: Bricken et al. (2023)
Abstract: Shows dictionary learning (Sparse Autoencoders) can extract clean single-concept features from models. Proves superposition problem from Phase 1 can be fixed.
Link: Anthropic

Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet

Authors: Templeton et al. (2024)
Abstract: Extracts millions of interpretable features (Golden Gate Bridge, scam emails) from production-scale model. Known as "Golden Gate Claude" paper proving interpretability scales to AGI-level models.
Link: Anthropic

Automated Interpretability

Automated Interpretability: A Comprehensive Survey

Authors: Multiple authors (2025)
Abstract: Comprehensive survey of techniques using AI to interpret AI including automated circuit discovery and neuron labeling. Represents shift from manual neuroscience to automated auditing.
Link: arXiv

🛠️ Phase 12: Essential Tools and Frameworks

Core Deep Learning Frameworks

PyTorch - Dynamic computation graph framework favored in research
Documentation: pytorch.org

TensorFlow - Google's framework with strong production deployment tools
Documentation: tensorflow.org

NumPy - Foundational numerical computing library
Documentation: numpy.org

Model Development Tools

Hugging Face Transformers - Access to pre-trained models and datasets
Documentation: huggingface.co/docs

Weights & Biases - Experiment tracking and visualization
Documentation: docs.wandb.ai

Jupyter/Colab - Interactive computing environments
Documentation: jupyter.org

High-Performance Computing

DeepSpeed - Microsoft's distributed training library
Documentation: deepspeed.ai

PyTorch FSDP - Fully Sharded Data Parallel for distributed training
Documentation: pytorch.org/docs/stable/fsdp

CUDA/cuDNN - NVIDIA GPU acceleration platform
Documentation: developer.nvidia.com/cuda

Specialized Libraries

PyTorch Geometric (PyG) - Graph Neural Networks library
Documentation: pytorch-geometric.readthedocs.io

TransformerLens - Mechanistic interpretability toolkit
Documentation: github.com/neelnanda-io/TransformerLens

SAE Lens - Sparse Autoencoder training and analysis
Documentation: github.com/jbloomAus/SAELens

📝 Effective Paper Reading Strategy

First Pass (5-10 minutes)

- Read Title, Abstract, and Conclusion
- Examine all Figures and Tables
- Ask: What problem is solved? What is the main result? Is it relevant?

Second Pass (1 hour)

- Read Introduction and Related Work thoroughly
- Read Methodology/Experiments sections conceptually
- Skip dense mathematics initially
- Understand architecture or algorithm conceptually

Third Pass (Deep Dive)

- Re-read Methodology focusing on mathematics and proofs
- Reproduce key equations
- Implement core algorithm to solidify understanding

🎯 Specialization Paths

For LLM Research & Alignment

Focus on: Phases 5-6 (LLMs, Scaling Laws), Phase 9 (RLHF, DPO), Phase 11 (Mechanistic Interpretability)

For Computer Vision

Focus on: Phases 2-3 (CNNs, Vision Transformers), Phase 7 (Generative Models, Diffusion)

For Robotics & Embodied AI

Focus on: Phase 9 (SAC, continuous control), Phase 10 (World Models, DreamerV3)

For Graph Learning & Drug Discovery

Focus on: Phase 8 (GNNs, molecular property prediction)

For AI Safety & Interpretability

Focus on: Phase 11 (Mechanistic Interpretability, Sparse Autoencoders)

For Efficient AI & Edge Computing

Focus on: Phase 4 (State Space Models, MoE), Phase 6 (Model Compression)

📚 Essential Resources

Conferences

- **NeurIPS** (Neural Information Processing Systems)
- **ICML** (International Conference on Machine Learning)
- **ICLR** (International Conference on Learning Representations)
- **CVPR** (Computer Vision and Pattern Recognition)
- **ACL** (Association for Computational Linguistics)

Online Resources

- **arXiv.org** - Latest research preprints
- **Papers with Code** - Papers with implementation code
- **Distill.pub** - Visual explanations of ML concepts
- **Anthropic Research** - Interpretability and safety research

Books

- **Deep Learning** by Goodfellow, Bengio, Courville
- **Reinforcement Learning: An Introduction** by Sutton & Barto
- **Speech and Language Processing** by Jurafsky & Martin

🚀 Your Research Journey

1. **Master Foundations** (Phases 1-2): 2-3 months
2. **Choose Specialization** (Phase 3): 1-2 months
3. **Explore Cutting Edge** (Phases 4-6): 3-6 months
4. **Deep Dive Specialization** (Phases 7-11): 6-12 months
5. **Build & Implement** (Phase 12): Ongoing
6. **Contribute Original Research**: 12+ months

Remember: Becoming an AI researcher is a marathon, not a sprint. Focus on understanding deeply rather than covering everything quickly. Implement papers, reproduce results, and contribute to open source projects to solidify your knowledge.

Good luck on your research journey! 🎓 🚀