Elite AI Research Paper Reading List: A Focused Guide for Advanced Practitioners

Complete Guide with Prerequisites & Study Resources

Core AI & Theoretical Foundations

1. Scaling Laws for Neural Language Models | Jared Kaplan, Sam McCandlish, Tom Brown, et al. | 2020[1]

- **Core idea:** Establishes power-law relationships between model performance and compute budget, dataset size, and model parameters, spanning over seven orders of magnitude.
- **Prerequisites:** Statistical learning theory, power laws, optimization theory, neural network fundamentals
- **Study First:**
  - Machine Learning Specialization (Andrew Ng) - Coursera[2]
  - Deep Learning Fundamentals - Lightning AI[3]
  - Statistics: mean, variance, regression analysis
- **Links:** arXiv | Code: tensor2tensor

2. Learning-to-Optimize with PAC-Bayesian Guarantees | Michael Sucker, Jalal Fadili, Peter Ochs | 2024[4]

- **Core idea:** First framework to learn optimization algorithms with provable generalization guarantees using PAC-Bayesian theory.
- **Prerequisites:** PAC learning theory, Bayesian inference, convex optimization, gradient descent
- **Study First:**
  - PAC-Bayesian Learning Tutorial - Benjamin Guedj[5]
  - User-friendly Introduction to PAC-Bayes Bounds[6]
  - Convex Optimization - Boyd & Vandenberghe (free online)
- **Links:** arXiv

3. Quantitative Convergence of Trained Neural Networks to Gaussian Processes | Eloy Mosig García, Andrea Agazzi, Dario Trevisan | 2024[7]

- **Core idea:** Provides explicit finite-width error bounds on neural network outputs during training using Neural Tangent Kernel theory.
- **Prerequisites:** Gaussian processes, Neural Tangent Kernels, probability theory, measure theory, Wasserstein distance
- **Study First:**
  - Gaussian Processes: Gaussian Processes for Machine Learning (Rasmussen & Williams)
  - Neural Tangent Kernels: Understanding Neural Tangent Kernels - Distill
  - Measure theory basics from analysis textbooks
- **Links:** arXiv

4. On the Opportunities and Risks of Foundation Models | Rishi Bommasani, Drew Hudson, et al. | 2021[8]

- **Core idea:** Comprehensive analysis of foundation models as a new AI paradigm across modalities.
- **Prerequisites:** Machine learning basics, deep learning fundamentals, natural language processing, computer vision
- **Study First:**
  - Deep Learning Fundamentals Handbook - freeCodeCamp[9]
  - Basic NLP and CV concepts
- **Links:** arXiv

5. Generalization Bounds: Perspectives from Information Theory and PAC-Bayes | Fredrik Hellström et al. | 2024[10]

- **Core idea:** Unified treatment connecting PAC-Bayesian and information-theoretic approaches to generalization bounds.
- **Prerequisites:** Information theory, PAC learning, statistical learning theory, KL divergence, mutual information
- **Study First:**
  - Information theory: Elements of Information Theory (Cover & Thomas)
  - PAC-Bayesian Learning: A Tutorial - Pascal Germain PDF[11]
- **Links:** arXiv

Machine Learning & Statistical Methods

1. Scaling Laws Across Model Architectures: Dense vs MoE Models | Siqi Wang, Zhengyu Chen, Bei Li, et al. | 2024[12]

- **Core idea:** Universal power-law scaling principles across Dense and Mixture-of-Experts models.

- **Prerequisites:** Neural network architectures, Mixture-of-Experts, scaling laws, hyperparameter optimization
- **Study First:**
  - MoE fundamentals: Sparse expert routing, gating mechanisms
  - Scaling Laws for Neural Language Models (prerequisite paper)
- **Links:** ACL Anthology

2. Scaling Laws for Multilingual Language Models | Yifei He, Alon Benhaim, Barun Patra, et al. | 2025[13]

- **Core idea:** Predicting optimal sampling ratios for languages in multilingual training based on language family analysis.
- **Prerequisites:** Multilingual NLP, language families, data mixing strategies, cross-lingual transfer
- **Study First:**
  - Multilingual NLP concepts
  - Language typology basics
  - Cross-lingual representation learning
- **Links:** OpenReview

3. More Flexible PAC-Bayesian Meta-Learning | Hossein Zakerinia et al. | 2024[14]

- **Core idea:** PAC-Bayesian framework that learns learning algorithms rather than just prior distributions.
- **Prerequisites:** Meta-learning, PAC-Bayesian theory, few-shot learning, algorithm learning
- **Study First:**
  - Meta-learning fundamentals: MAML, gradient-based meta-learning
  - PAC-Bayesian Analysis and Applications Tutorial[15]
- **Links:** PMLR

4. Learning via Surrogate PAC-Bayes | Arthur Picard-Weibel et al. | 2024[16]

- **Core idea:** Uses surrogate training objectives with orthogonal projections to maintain PAC-Bayesian guarantees.
- **Prerequisites:** PAC-Bayesian bounds, surrogate losses, orthogonal projections, optimization theory
- **Study First:**
  - Linear algebra: orthogonal projections, matrix decompositions
  - A Primer on PAC-Bayesian Learning - arXiv[17]
- **Links:** NeurIPS

5. Convergence Analysis for Deep Sparse Coding | Jianfei Li, Han Feng, Ding-Xuan Zhou | 2024[18]

- **Core idea:** Theoretical convergence rates for CNNs in sparse feature learning.
- **Prerequisites:** Sparse coding, optimization theory, convergence analysis, CNN architectures
- **Study First:**
  - Sparse representation theory
  - Convex optimization convergence rates
  - CNN mathematical foundations
- **Links:** arXiv

Deep Learning & Optimization

1. Proportional-Integral-Derivative Accelerated Optimizer (PIDAO) | Shen Chen et al. | 2024[19]

- **Core idea:** Integrates PID control theory into optimization for enhanced landscape exploration.
- **Prerequisites:** Control theory, PID controllers, optimization algorithms, gradient descent variants
- **Study First:**
  - Control systems engineering basics
  - An Overview of Gradient Descent Optimization Algorithms[20]
  - PID control fundamentals
- **Links:** Nature Communications

2. Differential Transformer | Tianzhu Ye, Li Dong, Yuqing Xia, et al. | 2024[21]

- **Core idea:** Differential attention mechanism calculating differences between two attention maps.
- **Prerequisites:** Transformer architecture, attention mechanisms, self-attention, multi-head attention
- **Study First:**
  - Attention Is All You Need (essential foundation)
  - Transformer Attention Mechanism Tutorial - GeeksforGeeks[22]
  - The Transformer Attention Mechanism - Machine Learning Mastery[23]
- **Links:** arXiv | Code

3. Large Language Model Inference Acceleration: A Hardware Perspective | Jinhao Li, Jiaming Xu, et al. | 2024[24]

- **Core idea:** Comprehensive survey of hardware-specific optimization methods for LLM inference.

- **Prerequisites:** Computer architecture, GPU programming, parallel computing, hardware accelerators
- **Study First:**
  - Computer systems architecture
  - AI Inference Performance with Hardware Accelerators[25]
  - CUDA programming basics
- **Links:** arXiv

4. Scaling Laws for Efficient Mixture-of-Experts Language Models | Changxin Tian, Kunlong Chen, et al. | 2025[26]

- **Core idea:** Unified scaling law for MoE models using Efficiency Leverage metric.
- **Prerequisites:** Mixture-of-Experts, expert routing, scaling laws, computational efficiency
- **Study First:**
  - MoE architecture fundamentals
  - Expert gating mechanisms
  - Scaling Laws for Neural Language Models
- **Links:** arXiv

5. Theory, Analysis, and Best Practices for Sigmoid Self-Attention | Apple ML Research | 2025[27]

- **Core idea:** Sigmoid attention as universal function approximator alternative to softmax.
- **Prerequisites:** Attention mechanisms, universal approximation theorem, activation functions, hardware optimization
- **Study First:**
  - Universal approximation theory
  - Attention mechanism mathematics
  - Hardware-aware deep learning
- **Links:** Apple ML Research

Computer Vision

1. VGGT: Visual Geometry Grounded Transformer | Jianyuan Wang, Minghao Chen, et al. | 2025[28]

- **Core idea:** Treats 3D reconstruction as prediction problem using alternating attention mechanisms.
- **Prerequisites:** 3D computer vision, camera geometry, SLAM, structure from motion, multi-view geometry
- **Study First:**

    - Multiple View Geometry in Computer Vision - Hartley & Zisserman
    - Camera calibration and pose estimation
    - 3D reconstruction fundamentals
- **Links:** CVPR 2025

2. SAM 2: Segment Anything in Images and Videos | Nikhila Ravi, Valentin Gabeur, et al. | 2025[29]

- **Core idea:** Extends segment anything capability to video sequences with temporal consistency.
- **Prerequisites:** Image segmentation, video analysis, temporal consistency, foundation models
- **Study First:**
    - Semantic segmentation concepts
    - Video understanding fundamentals
    - The original SAM paper
- **Links:** ICLR 2025

3. Visual Autoregressive Modeling | Keyu Tian, Yi Jiang, et al. | 2024[29]

- **Core idea:** Autoregressive image generation using next-scale prediction.
- **Prerequisites:** Autoregressive models, image generation, sequence modeling, generative models
- **Study First:**
    - Autoregressive sequence modeling
    - Image representation learning
    - Generative modeling fundamentals
- **Links:** NeurIPS 2024

4. Spatial-Mamba: Effective Visual State Space Models | Chaodong Xiao, Minghan Li, et al. | 2025[30]

- **Core idea:** Structure-aware state fusion using dilated convolutions for 2D vision tasks.
- **Prerequisites:** State space models, Mamba architecture, convolutional neural networks, dilated convolutions
- **Study First:**
    - Mamba: Linear-Time Sequence Modeling
    - Understanding State Space Models - IBM[31]
    - Structured State Space Models Visually Explained[32]
- **Links:** ICLR 2025

5. BioCLIP: A Vision Foundation Model for the Tree of Life | Samuel Stevens, Jiaman Wu, et al. | 2024[33]

- **Core idea:** Biology foundation model trained on TreeOfLife-10M dataset.
- **Prerequisites:** Foundation models, CLIP architecture, contrastive learning, biological taxonomy
- **Study First:**
  - CLIP paper understanding
  - Contrastive learning fundamentals
  - Multi-modal representation learning
- **Links:** CVPR 2024

Natural Language Processing & LLMs

1. Direct Preference Optimization | Rafael Rafailov, Archit Sharma, Eric Mitchell, et al. | 2023[29]

- **Core idea:** Eliminates separate reward model in RLHF by directly optimizing on preference data.
- **Prerequisites:** Reinforcement Learning from Human Feedback (RLHF), reward modeling, policy optimization
- **Study First:**
  - Reinforcement learning fundamentals
  - Policy gradient methods
  - RLHF concepts and reward modeling
- **Links:** NeurIPS 2023

2. Large Language Models: A Survey | Shervin Minaee, Tomas Mikolov, et al. | 2024[34]

- **Core idea:** Comprehensive survey of prominent LLM families and training techniques.
- **Prerequisites:** Natural language processing, neural language models, transformer architecture
- **Study First:**
  - Attention Is All You Need
  - Basic NLP concepts: tokenization, embeddings
  - Language modeling fundamentals
- **Links:** arXiv

3. Are Emergent Abilities of Large Language Models a Mirage? | Rylan Schaeffer, Brando Miranda, Sanmi Koyejo | 2023[29]

- **Core idea:** Challenges notion of emergent abilities as artifacts of evaluation metrics.
- **Prerequisites:** Large language models, emergent properties, evaluation metrics, statistical analysis

- **Study First:**
  - LLM scaling behaviors
  - Evaluation methodology in NLP
  - Statistical hypothesis testing
- **Links:** NeurIPS 2023

4. Not All Tokens Are What You Need for Pretraining | Zhenghao Lin, Zhibin Gou, et al. | 2024[29]

- **Core idea:** Strategic token selection during pretraining improves efficiency.
- **Prerequisites:** Language model pretraining, tokenization, data efficiency, curriculum learning
- **Study First:**
  - Pretraining methodology
  - Tokenization strategies
  - Data-efficient learning concepts
- **Links:** NeurIPS 2024

5. Internal and External Impacts of Natural Language Processing Papers | Yu Zhang | 2025[35]

- **Core idea:** Analyzes citation patterns and external influence of NLP research.
- **Prerequisites:** Research methodology, citation analysis, NLP field evolution, bibliometrics
- **Study First:**
  - NLP history and major milestones
  - Research impact measurement
  - Citation network analysis
- **Links:** arXiv

Generative Models (GANs, Diffusion, etc.)

1. Guiding a Diffusion Model with a Bad Version of Itself | Tero Karras, Miika Aittala, et al. | 2024[29]

- **Core idea:** Novel guidance technique using a "bad" version of the same diffusion model.
- **Prerequisites:** Diffusion models, denoising diffusion probabilistic models, guidance techniques
- **Study First:**
  - Denoising Diffusion Probabilistic Models
  - Classifier-free guidance
  - Diffusion model training

- **Links:** NeurIPS 2024

2. Analytic-DPM: Analytic Estimate of Optimal Reverse Variance | Fan Bao, Chongxuan Li, Jun Zhu, Bo Zhang | 2022[29]

- **Core idea:** Analytical solution for optimal reverse variance in diffusion models.
- **Prerequisites:** Diffusion models, variational inference, ELBO optimization, stochastic differential equations
- **Study First:**
  - Variational autoencoders
  - Understanding Diffusion Models
  - Stochastic processes basics
- **Links:** ICLR 2022

3. Opportunities and Challenges of Diffusion Models | Minsheng Chen et al. | 2024[36]

- **Core idea:** Comprehensive review connecting diffusion models to black-box optimization.
- **Prerequisites:** Diffusion models, generative modeling, optimization theory, controlled generation
- **Study First:**
  - Generative model taxonomy
  - Optimization fundamentals
  - Sampling techniques
- **Links:** Nature Scientific Reports

4. Generative Diffusion Models for Sequential Recommendations | Sharare Zolghadr et al. | 2024[37]

- **Core idea:** DiffuRecSys using diffusion models with offset noise for recommendations.
- **Prerequisites:** Recommendation systems, sequential modeling, diffusion models, collaborative filtering
- **Study First:**
  - Recommendation system fundamentals
  - Sequential pattern mining
  - Matrix factorization techniques
- **Links:** arXiv

5. DGRM: Diffusion-GAN Recommendation Model | Deng Jiangzhou et al. | 2024[38]

- **Core idea:** Integrates diffusion models with GANs for mutual enhancement.
- **Prerequisites:** GANs, diffusion models, recommendation systems, mode collapse, adversarial training
- **Study First:**

    - GAN fundamentals
    - GAN training challenges
    - Hybrid generative models
- **Links:** Pattern Recognition Journal

Transformers / Attention Architectures

1. Attention Is All You Need | Ashish Vaswani, Noam Shazeer, et al. | 2017[39]

- **Core idea:** Introduces Transformer architecture based solely on attention mechanisms.
- **Prerequisites:** Neural networks, RNNs/LSTMs, sequence-to-sequence models, basic linear algebra
- **Study First:**
    - Deep Learning Fundamentals - Lightning AI[3]
    - Sequence modeling basics
    - Tutorial 6: Transformers and Multi-Head Attention - UvA DL[40]
- **Links:** arXiv | Code

2. Mamba: Linear-Time Sequence Modeling with Selective State Spaces | Albert Gu, Tri Dao | 2023[41]

- **Core idea:** Selective state space models achieving Transformer-level performance with linear scaling.
- **Prerequisites:** State space models, recurrent neural networks, linear systems, selective mechanisms
- **Study First:**
    - Introduction to State Space Models - IBM[31]
    - State Space Models Tutorial - HuggingFace[42]
    - Understanding Mamba and SSMs - Towards AI[43]
- **Links:** arXiv | Code

3. Efficiently Modeling Long Sequences with Structured State Spaces | Albert Gu et al. | 2022[29]

- **Core idea:** S4 model for long-range dependencies using structured state space models.
- **Prerequisites:** State space models, long-range dependencies, structured matrices, computational efficiency
- **Study First:**
    - Linear algebra: structured matrices
    - Signal processing basics
    - A Visual Guide to Mamba and State Space Models[44]

- **Links:** ICLR 2022

4. Neural Machine Translation by Jointly Learning to Align and Translate | Dzmitry Bahdanau et al. | 2015[29]

- **Core idea:** Introduces attention mechanism for neural machine translation.
- **Prerequisites:** Neural machine translation, encoder-decoder architectures, RNNs, alignment
- **Study First:**
  - Sequence-to-sequence models
  - RNN fundamentals
  - Machine translation basics
- **Links:** ICLR 2015

5. Transformer Architecture and Attention Mechanisms in Genome Analysis | Su Rin Choi et al. | 2023[45]

- **Core idea:** Comprehensive analysis of Transformer applications in genomics.
- **Prerequisites:** Transformers, bioinformatics, genomic sequences, computational biology
- **Study First:**
  - Bioinformatics fundamentals
  - DNA/protein sequence analysis
  - Transformer architecture
- **Links:** PMC

Algorithms & Efficient Architectures

1. LtNet: Lightweight Neural Network for Mobile Edge Computing | Liu Liu, Zhifei Xu | 2025[46]

- **Core idea:** Lightweight architecture with H-Swish activation and selective Squeeze-and-Excitation modules.
- **Prerequisites:** Mobile computing, edge AI, neural architecture design, computational efficiency
- **Study First:**
  - Fundamentals of Deep Learning for Computer Vision - NVIDIA[47]
  - MobileNet architectures
  - Edge computing constraints
- **Links:** PMC

2. Faster Cascades via Speculative Decoding | Harikrishna Narasimhan et al. | 2025[29]

- **Core idea:** Speculative decoding approach for cascade models accelerating inference.

- **Prerequisites:** Model cascades, speculative execution, inference optimization, early exit networks
- **Study First:**
  - Model ensemble techniques
  - Early exit strategies
  - Inference acceleration methods
- **Links:** ICLR 2025

3. Data Shapley in One Training Run | Jiachen Wang, Prateek Mittal, et al. | 2025[29]

- **Core idea:** Efficient algorithm to compute Data Shapley values in single training run.
- **Prerequisites:** Shapley values, game theory, data valuation, federated learning
- **Study First:**
  - Cooperative game theory
  - Shapley value computation
  - Data valuation concepts
- **Links:** ICLR 2025

4. Optimizing Lightweight Neural Networks for Efficient Mobile Edge Computing | Liu Liu, Zhifei Xu | 2025[46]

- **Core idea:** Multi-Agent RL framework with LtNet for dynamic resource allocation.
- **Prerequisites:** Multi-agent reinforcement learning, resource allocation, edge computing, neural architecture optimization
- **Study First:**
  - Multi-agent systems
  - Reinforcement learning fundamentals
  - Resource management algorithms
- **Links:** PMC

5. RaNAS: Resource-Aware Neural Architecture Search | Multiple Authors | 2024[48]

- **Core idea:** Resource-aware NAS using graph neural networks for hardware prediction.
- **Prerequisites:** Neural Architecture Search, graph neural networks, hardware-aware ML, resource constraints
- **Study First:**
  - NAS fundamentals
  - Graph neural networks
  - Hardware performance modeling
- **Links:** ACM Digital Library

GenAI & Multimodal Systems

1. What Matters When Building Vision-Language Models? | Hugo Laurençon et al. | 2024[49]

- **Core idea:** Systematic analysis of VLM design choices with Idefics2 model introduction.
- **Prerequisites:** Vision-language models, multimodal learning, cross-modal attention, contrastive learning
- **Study First:**
  - An Introduction to Vision-Language Modeling[50]
  - CLIP architecture understanding
  - Cross-modal representation learning
- **Links:** NeurIPS

2. VL-Mamba: Exploring State Space Models for Multimodal Learning | Yanyuan Qiao et al. | 2024[51]

- **Core idea:** First application of state space models to multimodal tasks.
- **Prerequisites:** State space models, multimodal learning, vision-language tasks, Mamba architecture
- **Study First:**
  - Mamba: Linear-Time Sequence Modeling
  - Multimodal fusion techniques
  - Vision transformer architectures
- **Links:** PMLR

3. A Comprehensive Survey of Retrieval-Augmented Generation | Shailja Gupta et al. | 2024[52]

- **Core idea:** Complete survey of RAG evolution from foundations to current state-of-the-art.
- **Prerequisites:** Information retrieval, language models, vector databases, semantic search
- **Study First:**
  - Information retrieval fundamentals
  - Vector similarity search
  - Dense passage retrieval
- **Links:** arXiv

4. Retrieval Augmented Generation or Long-Context LLMs? | Zhuowan Li, Cheng Li, et al. | 2024[53]

- **Core idea:** Systematic comparison with Self-Route method for dynamic approach selection.
- **Prerequisites:** RAG systems, long-context models, attention mechanisms, computational efficiency
- **Study First:**

- RAG fundamentals
- Long-context attention mechanisms
- Efficiency-accuracy trade-offs

- **Links:** EMNLP

5. An Introduction to Vision-Language Modeling | Florian Bordes et al. | 2024[50]

- **Core idea:** Comprehensive introduction to VLMs covering architectures and applications.
- **Prerequisites:** Computer vision, natural language processing, multimodal learning, foundation models
- **Study First:**
  - CNN fundamentals for vision
  - Transformer fundamentals for language
  - Contrastive learning principles
- **Links:** arXiv

Enhanced 6-Month Reading Roadmap with Prerequisites

Phase 1: Mathematical & Theoretical Foundations (Month 1-2)

Essential Prerequisites to Complete First:

1. **Linear Algebra**: Vector spaces, matrix operations, eigenvalues, SVD
   - **Resource**: 3Blue1Brown Linear Algebra Series
   - **Book**: Linear Algebra Done Right - Sheldon Axler
2. **Probability & Statistics**: Probability distributions, Bayesian inference, hypothesis testing
   - **Resource**: Khan Academy Statistics
   - **Course**: Fundamentals of Statistics - WQU
3. **Calculus**: Derivatives, gradients, chain rule, optimization
   - **Resource**: Khan Academy Calculus
4. **Programming**: Python proficiency, NumPy, basic ML libraries
   - **Resource**: Python for Data Science - DataCamp

Core Theory Papers to Read:

1. **Attention Is All You Need** (Start here - essential foundation)
2. **Scaling Laws for Neural Language Models**
3. **On the Opportunities and Risks of Foundation Models**
4. **Generalization Bounds: PAC-Bayes + Information Theory**

Supplementary Study (parallel to reading):

- **PAC-Bayesian Theory**: Benjamin Guedj's Tutorial[5]
- **Convex Optimization**: Boyd & Vandenberghe (free online)

Phase 2: Deep Learning & Architectures (Month 2-3)

Prerequisites to Strengthen:

- **Deep Learning Fundamentals**: Lightning AI Course[3]
- **Transformer Mathematics**: UvA Tutorial[40]

Architecture Papers:

1. **Mamba: Linear-Time Sequence Modeling**
   - **Prep**: State Space Models Guide[31]
2. **Differential Transformer**
3. **Spatial-Mamba** (for computer vision)
4. **Theory of Sigmoid Self-Attention**

Optimization Papers:

5. **PIDAO: PID Control for Optimization**
   - **Prep**: Control theory basics, PID controllers
6. **Scaling Laws Across Model Architectures**

Phase 3: Domain Specialization (Month 3-4)

Choose Your Primary Track + Prerequisites:

Computer Vision Track:

- **Prerequisites**:
  - Multiple View Geometry - Hartley & Zisserman
  - CNN architectures, image processing fundamentals
- **Papers**: VGGT, SAM 2, Visual Autoregressive Modeling, BioCLIP

NLP & LLMs Track:

- **Prerequisites**:
  - Language modeling fundamentals
  - RLHF concepts, reward modeling
- **Papers**: DPO, LLM Survey, Emergent Abilities, Token Selection

Generative Models Track:

- **Prerequisites**:
  - Understanding Diffusion Models - Lilian Weng
  - VAE fundamentals, GANs basics
- **Papers**: Analytic-DPM, Diffusion Guidance, DGRM, Survey

Phase 4: Efficiency & Hardware (Month 4-5)

Prerequisites to Study:

- **Computer Architecture**: GPU computing, parallel algorithms
- **Hardware-Aware ML**: NVIDIA Deep Learning Fundamentals[47]

Core Papers:

1. **LLM Inference Acceleration: Hardware Perspective**
2. **LtNet: Lightweight Networks for Edge Computing**
3. **Faster Cascades via Speculative Decoding**
4. **Scaling Laws for Efficient MoE Models**

Advanced Theory:

5. **More Flexible PAC-Bayesian Meta-Learning**
6. **Learning via Surrogate PAC-Bayes**

Phase 5: Multimodal & Applications (Month 5-6)

Prerequisites:

- Vision + NLP foundations completed
- Introduction to Vision-Language Modeling[50]

Multimodal Papers:

1. **What Matters When Building Vision-Language Models**
2. **VL-Mamba: State Space Models for Multimodal Learning**
3. **Comprehensive Survey of RAG**
4. **RAG vs Long-Context LLMs**

Implementation Projects:

- Implement 2-3 key architectures from papers
- Focus on reproducible results

Phase 6: Advanced Topics & Integration (Month 6)

Research Synthesis:

1. **Data Shapley in One Training Run**
2. **Internal and External Impacts of NLP Papers**
3. Latest breakthrough papers from your chosen specialization

Capstone Activities:

- **Implementation Portfolio**: 3-5 paper implementations
- **Research Proposal**: Novel research direction based on learned foundations
- **Paper Reviews**: Critical analysis of 5-10 papers

Success Metrics & Checkpoints:

Monthly Checkpoints:

- **Month 1**: Complete all prerequisites, understand PAC-Bayes basics
- **Month 2**: Implement basic Transformer, understand Mamba fundamentals
- **Month 3**: Deep dive implementation in chosen domain
- **Month 4**: Hardware optimization project completed
- **Month 5**: Multimodal system implementation
- **Month 6**: Research portfolio and novel contribution

Prerequisites Validation:

Before advancing to each phase, ensure you can:

- **Phase 1→2**: Derive attention mechanism mathematics, explain PAC-bounds
- **Phase 2→3**: Implement Transformer from scratch, understand SSM equations
- **Phase 3→4**: Complete domain-specific project demonstrating mastery
- **Phase 4→5**: Optimize model for hardware deployment
- **Phase 5→6**: Build and evaluate multimodal system

Study Resources by Difficulty:

Beginner-Friendly Entry Points:

- Machine Learning Crash Course - Google[54]
- Deep Learning Fundamentals Handbook - freeCodeCamp[9]

Intermediate Resources:

- Deep Learning Book - Ian Goodfellow[55]
- Understanding Deep Learning - Simon Prince[56]

Advanced Theory:

- Convex Analysis - Rockafellar
- Convex Optimization Theory - Dimitri Bertsekas (free PDF)[57]

This enhanced roadmap ensures you have proper mathematical foundations before tackling advanced papers, provides clear prerequisites for each paper, and includes specific study resources to fill knowledge gaps. The structured approach guarantees both theoretical understanding and practical implementation skills.

⁂

1. https://arxiv.org/abs/2001.08361
2. https://www.coursera.org/specializations/machine-learning-introduction
3. https://lightning.ai/pages/courses/deep-learning-fundamentals/
4. https://arxiv.org/abs/2404.03290
5. https://bguedj.github.io/icml2019/
6. https://www.nowpublishers.com/article/DownloadSummary/MAL-100
7. https://arxiv.org/html/2509.24544v1
8. https://arxiv.org/abs/2108.07258
9. https://www.freecodecamp.org/news/deep-learning-fundamentals-handbook-start-a-career-in-ai/
10. https://arxiv.org/abs/2309.04381

11. https://www.pascalgermain.info/talks/pacbayes-tutorial-2023.pdf
12. https://aclanthology.org/2024.emnlp-main.319.pdf
13. https://openreview.net/forum?id=T2h2V7Rx7q
14. https://proceedings.mlr.press/v235/zakerinia24a.html
15. https://sites.google.com/site/ecmlpkddtutorialpacbayes/
16. https://proceedings.neurips.cc/paper_files/paper/2024/file/5fba70900a84a8fb755c48ba99420c95-Paper-Conference.pdf
17. https://arxiv.org/pdf/1901.05353.pdf
18. https://arxiv.org/abs/2408.05540
19. https://www.nature.com/articles/s41467-024-54451-3
20. https://www.ruder.io/optimizing-gradient-descent/
21. https://hippocampus-garden.com/deep_learning_2024/
22. https://www.geeksforgeeks.org/nlp/transformer-attention-mechanism-in-nlp/
23. https://www.machinelearningmastery.com/the-transformer-attention-mechanism/
24. https://arxiv.org/abs/2410.04466
25. https://www.aiacceleratorinstitute.com/improving-ai-inference-performance-with-hardware-accelerators/
26. https://arxiv.org/html/2507.17702v1
27. https://machinelearning.apple.com/research/iclr-2025
28. https://www.basic.ai/blog-post/cvpr-2025-top-papers-award-winners-and-notable-research
29. https://github.com/SarahRastegar/Best-Papers-Top-Venues
30. https://openreview.net/forum?id=iDe1mtxqK5
31. https://www.ibm.com/think/topics/state-space-model
32. https://towardsdatascience.com/structured-state-space-models-visually-explained-86cfe2757386/
33. https://cvpr.thecvf.com/Conferences/2024/News/Awards
34. https://arxiv.org/abs/2402.06196
35. https://arxiv.org/abs/2505.16061
36. https://academic.oup.com/nsr/article/11/12/nwae348/7810289
37. https://arxiv.org/abs/2410.19429
38. https://www.sciencedirect.com/science/article/abs/pii/S0031320324004436
39. https://papers.neurips.cc/paper/7181-attention-is-all-you-need.pdf
40. https://uvadlc-notebooks.readthedocs.io/en/latest/tutorial_notebooks/tutorial6/Transformers_and_MHAttention.html
41. https://arxiv.org/abs/2312.00752
42. https://huggingface.co/blog/lbourdois/get-on-the-ssm-train
43. https://towardsai.net/p/l/understanding-mamba-and-selective-state-space-models-ssms
44. https://newsletter.maartengrootendorst.com/p/a-visual-guide-to-mamba-and-state
45. https://pmc.ncbi.nlm.nih.gov/articles/PMC10376273/
46. https://pmc.ncbi.nlm.nih.gov/articles/PMC12218929/

47. https://learn.nvidia.com/courses/course-detail?course_id=course-v1%3ADLI+C-FX-01+V3
48. https://dl.acm.org/doi/10.1145/3703353
49. https://proceedings.neurips.cc/paper_files/paper/2024/file/a03037317560b8c5f2fb4b6466d4c439-Paper-Conference.pdf
50. https://arxiv.org/abs/2405.17247
51. https://proceedings.mlr.press/v262/qiao24a.html
52. https://arxiv.org/abs/2410.12837
53. https://aclanthology.org/2024.emnlp-industry.66/
54. https://developers.google.com/machine-learning/crash-course/prereqs-and-prework
55. https://www.deeplearningbook.org
56. https://udlbook.github.io/udlbook/
57. http://web.mit.edu/dimitrib/www/Convex_Theory_Entire_Book.pdf
58. https://www.mccormick.northwestern.edu/computer-science/academics/courses/descriptions/496-1.html
59. https://www.guvi.in/blog/prerequisites-for-machine-learning/
60. https://www.wqu.edu/deep-learning-lab
61. https://www.baeldung.com/cs/machine-learning-how-to-start
62. https://pub.towardsai.net/step-by-step-exploration-of-transformer-attention-mechanisms-e9d36548d2d8
63. https://www.kaggle.com/learn/intro-to-deep-learning
64. https://www.youtube.com/watch?v=eMlx5fFNoYc&vl=en
65. https://www.geeksforgeeks.org/machine-learning/machine-learning-prerequisites/
66. http://www.d2l.ai/chapter_attention-mechanisms-and-transformers/index.html
67. https://www.reddit.com/r/learnmachinelearning/comments/s88qei/can_someone_tell_me_all_the_prerequisites/
68. https://learn.sas.com/course/view.php?id=583
69. https://www.mpi-inf.mpg.de/departments/computer-vision-and-machine-learning/teaching/courses-1/ss-2022-convex-analysis
70. https://bguedj.github.io/icml2019/material/main.pdf
71. https://www.colorado.edu/cs/csci-5254-convex-optimization-and-its-applications
72. https://cfp.pydata.org/berlin2025/talk/GRZ3RG/
73. https://carmamaths.org/jon/Preprints/Books/CaNo2/cano2f.pdf
74. https://anr.fr/Project-ANR-18-CE23-0015
75. https://neptune.ai/blog/state-space-models-as-natural-language-models
76. https://ocw.mit.edu/courses/6-253-convex-analysis-and-optimization-spring-2012/
77. https://www.math.columbia.edu/department/pinkham/Optimizationbook.pdf