Автоформатировать


```json
{
  "project": "Interconnected AI Architecture: Parameter Scaling Constructor",
  "version": "1.0",
  "date": "2026-02-12",
  "status": "Production Ready",
  "documents": [
    {
      "id": 1,
      "filename": "interconnected_ai_architecture_v1.md",
      "type": "Architecture & Theory",
      "sections": 10,
      "pages": "~50",
      "key_topics": [
        "Base Encoder (Frozen)",
        "Sparsely-Gated MoE",
        "LoRA Adapters",
        "Multi-Task Learning",
        "Parameter Scaling Trajectories",
        "Implementation Guidelines"
      ]
    },
    {
      "id": 2,
      "filename": "practical_implementation_guide.md",
      "type": "Code & Implementation",
      "sections": 3,
      "pages": "~30",
      "key_topics": [
        "Complete PyTorch Implementation",
        "MoE Layer from Scratch",
        "Training Loop",
        "Monitoring & Metrics",
        "Production Optimization"
      ]
    },
    {
      "id": 3,
      "filename": "final_metrics_and_recommendations.md",
      "type": "Analysis & Recommendations",
      "sections": 7,
      "pages": "~40",
      "key_topics": [
        "Comparative Analysis",
        "Resource Efficiency",
        "Usage Scenarios",
        "Hyperparameter Tuning",
        "Deployment Checklist"
      ]
    },
    {
      "id": 4,
      "filename": "quick_reference_cheatsheet.md",
      "type": "Quick Reference",
      "sections": "12 sections",
      "pages": "1-2",
      "key_topics": [
        "Architecture Overview",
        "Component Parameters",
        "Key Metrics",
        "Debugging Tips",
        "Quick Start Code"
      ]
    }
  ],
  "visualizations": [
    {
      "filename": "interconnected_ai_architecture.png",
```

```
      "type": "Architecture Diagram"
    },
    {
      "filename": "scaling_comparison.png",
      "description": "Parameter scaling methods comparison",
      "type": "Comparative Chart"
    },
    {
      "filename": "multitask_distribution.png",
      "description": "Multi-task parameter distribution",
      "type": "Distribution Chart"
    }
  ],
  "key_metrics": {
    "total_parameters_b": 11.8,
    "effective_parameters_b": 4.1,
    "parameter_efficiency": "41% reduction vs Dense",
    "training_speedup": "4× faster",
    "inference_speedup": "3.3× faster",
    "memory_savings": "75% (activations), 41% (weights)",
    "accuracy_drop": "-0.7% (acceptable)",
    "branch_expansion_cost": "2-4 hours",
    "shared_parameters_pct": 75,
    "branches_supported": "3-100+ with scaling"
  },
  "architecture_components": {
    "base_encoder": {
      "params_b": 1.2,
      "frozen": true,
      "shared_pct": 100,
      "type": "Dense Transformer"
    },
    "moe_layer_1": {
      "params_b": 4.8,
      "num_experts": 8,
      "k": 2,
      "active_pct": 25,
      "type": "Sparse MoE"
    },
    "moe_layer_2": {
      "params_b": 4.8,
      "num_experts": 8,
      "k": 2,
      "active_pct": 25,
      "type": "Sparse MoE"
    },
    "adapters": {
      "params_m": 120,
      "count": 3,
      "type": "LoRA (rank=8)",
      "total_m": 360
    },
    "task_heads": {
      "params_m": 50,
      "count": 3,
      "type": "Linear Output",
      "total_m": 150
    }
  },
  "recommendations": {
    "optimal_for": [
      "5-50 interdependent tasks/branches",
      "Resource-constrained deployments",
      "Rapid iterative expansion",
      "Multi-task reinforcement learning"
    ],
    "not_recommended_for": [
      "1-3 simple tasks (use dense)",
```

```
      "Unlimited budget (Dense simpler)"
    ],
    "deployment_stages": {
      "stage_1_prototyping": "2 weeks",
      "stage_2_optimization": "2 weeks",
      "stage_3_scaling": "4 weeks",
      "stage_4_production": "ongoing"
    }
  },
  "usage_scenarios": {
    "scenario_a_small": {
      "branches": 10,
      "total_params_b": 12.5,
      "gpu": "1x A100 80GB",
      "training_days": 30
    },
    "scenario_b_medium": {
      "branches": 30,
      "total_params_b": 25.5,
      "gpu": "2x A100 80GB",
      "training_days": 45
    },
    "scenario_c_large": {
      "branches": 100,
      "total_params_b": 147,
      "gpu": "4x A100 80GB",
      "training_days": 90
    }
  },
  "learning_rates": {
    "moe_layers": "1e-4",
    "gating_networks": "5e-4",
    "lora_adapters": "5e-4",
    "task_heads": "1e-3",
    "warmup_steps": 500,
    "total_steps": 50000
  },
  "quality_assurance": {
    "checks": [
      "Base encoder frozen",
      "Aux loss decreasing",
      "Expert utilization balanced",
      "Branch-specific accuracy monitored",
      "Gradient clipping enabled",
      "Checkpoints saved regularly"
    ],
    "monitoring_metrics": [
      "Total loss",
      "Aux loss",
      "Per-branch accuracy",
      "Expert utilization",
      "Training speed",
      "Memory usage"
    ]
  },
  "related_papers": [
    {
      "title": "Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts
Layer",
      "authors": "Shazeer et al.",
      "year": 2017,
      "venue": "ICLR",
      "url": "https://arxiv.org/pdf/1701.06538.pdf"
    },
    {
      "title": "LoRA: Low-Rank Adaptation of Large Language Models",
      "authors": "Hu et al.",
      "year": 2022,
```

```
      "url": "https://arxiv.org/pdf/2106.09685.pdf"
    },
    {
      "title": "A Survey on Mixture of Experts in Large Language Models",
      "authors": "Shen et al.",
      "year": 2024,
      "venue": "arxiv",
      "url": "https://arxiv.org/pdf/2407.06204.pdf"
    },
    {
      "title": "Multi-Task Reinforcement Learning Enables Parameter Scaling",
      "authors": "Arnob et al.",
      "year": 2025,
      "venue": "arxiv",
      "url": "https://arxiv.org/html/2503.05126v3"
    }
  ]
}
```