AWS GenAI RAG Workshop: Outcomes & Strategic Value

Engineering Management Briefing - June 5, 2025

Engineering Team

June 5, 2025

Workshop Overview

AWS GenAI RAG Workshop Summary

- Date: May 29, 2025 (1 day investment)

Start with the concrete facts about the workshop to establish credibility and context.

What We Built

Workshop Overview

AWS GenAI RAG Workshop Summary

- Date: May 29, 2025 (1 day investment)
- Participants: [X] developers from our team

Start with the concrete facts about the workshop to establish credibility and context.

What We Built

Workshop Overview

AWS GenAI RAG Workshop Summary

- Date: May 29, 2025 (1 day investment)
- Participants: [X] developers from our team
- Format: Hands-on technical workshop

Start with the concrete facts about the workshop to establish credibility and context.

What We Built

Workshop Overview

AWS GenAI RAG Workshop Summary

- Date: May 29, 2025 (1 day investment)
- Participants: [X] developers from our team
- Format: Hands-on technical workshop
- Focus: Production-ready RAG implementations

Start with the concrete facts about the workshop to establish credibility and context.

What We Built

Workshop Overview

AWS GenAI RAG Workshop Summary

- Date: May 29, 2025 (1 day investment)
- Participants: [X] developers from our team
- Format: Hands-on technical workshop
- Focus: Production-ready RAG implementations

Start with the concrete facts about the workshop to establish credibility and context.

What We Built

- Complete RAG pipeline from scratch

Workshop Overview

AWS GenAI RAG Workshop Summary

- Date: May 29, 2025 (1 day investment)
- Participants: [X] developers from our team
- Format: Hands-on technical workshop
- Focus: Production-ready RAG implementations

Start with the concrete facts about the workshop to establish credibility and context.

What We Built

- Complete RAG pipeline from scratch
- Cost optimization frameworks

Workshop Overview

AWS GenAI RAG Workshop Summary

- Date: May 29, 2025 (1 day investment)
- Participants: [X] developers from our team
- Format: Hands-on technical workshop
- Focus: Production-ready RAG implementations

Start with the concrete facts about the workshop to establish credibility and context.

What We Built

- Complete RAG pipeline from scratch
- Cost optimization frameworks
- Domain-specific fine-tuning capabilities

Workshop Overview

AWS GenAI RAG Workshop Summary

- Date: May 29, 2025 (1 day investment)
- Participants: [X] developers from our team
- Format: Hands-on technical workshop
- Focus: Production-ready RAG implementations

Start with the concrete facts about the workshop to establish credibility and context.

What We Built

- Complete RAG pipeline from scratch
- Cost optimization frameworks
- Domain-specific fine-tuning capabilities
- Production monitoring systems

Business Value

Immediate ROI Opportunities

- 30-50% reduction through caching

Real-World Applications

Business Value

Immediate ROI Opportunities

- 30-50% reduction through caching
- 40% savings via model tiering

Real-World Applications

Business Value

Immediate ROI Opportunities

- 30-50% reduction through caching
- 40% savings via model tiering
- $X monthly optimization potential

Real-World Applications

Business Value

Immediate ROI Opportunities

- 30-50% reduction through caching
- 40% savings via model tiering
- $X monthly optimization potential
- 3x faster information retrieval

Real-World Applications

Business Value

Immediate ROI Opportunities

- 30-50% reduction through caching
- 40% savings via model tiering
- $X monthly optimization potential
- 3x faster information retrieval
- 80% accuracy improvement

Real-World Applications

Business Value

Immediate ROI Opportunities

- 30-50% reduction through caching
- 40% savings via model tiering
- $X monthly optimization potential
- 3x faster information retrieval
- 80% accuracy improvement
- 50% reduction in manual research

Real-World Applications

Business Value

Immediate ROI Opportunities

- 30-50% reduction through caching
- 40% savings via model tiering
- $X monthly optimization potential
- 3x faster information retrieval
- 80% accuracy improvement
- 50% reduction in manual research

Real-World Applications

1. Customer Support Automation

Business Value

Immediate ROI Opportunities

- 30-50% reduction through caching
- 40% savings via model tiering
- $X monthly optimization potential
- 3x faster information retrieval
- 80% accuracy improvement
- 50% reduction in manual research

Real-World Applications

1. Customer Support Automation
    - Instant, accurate responses from documentation

Business Value

Immediate ROI Opportunities

- 30-50% reduction through caching
- 40% savings via model tiering
- $X monthly optimization potential
- 3x faster information retrieval
- 80% accuracy improvement
- 50% reduction in manual research

Real-World Applications

1. Customer Support Automation
    - Instant, accurate responses from documentation
    - Reduced ticket resolution time

Business Value

Immediate ROI Opportunities

- 30-50% reduction through caching
- 40% savings via model tiering
- $X monthly optimization potential
- 3x faster information retrieval
- 80% accuracy improvement
- 50% reduction in manual research

Real-World Applications

1. Customer Support Automation
   - Instant, accurate responses from documentation
   - Reduced ticket resolution time
2. Internal Knowledge Management

Business Value

Immediate ROI Opportunities

- 30-50% reduction through caching
- 40% savings via model tiering
- $X monthly optimization potential
- 3x faster information retrieval
- 80% accuracy improvement
- 50% reduction in manual research

Real-World Applications

1. Customer Support Automation
   - Instant, accurate responses from documentation
   - Reduced ticket resolution time
2. Internal Knowledge Management
   - Unified access to company knowledge

Business Value

Immediate ROI Opportunities

- 30-50% reduction through caching
- 40% savings via model tiering
- $X monthly optimization potential
- 3x faster information retrieval
- 80% accuracy improvement
- 50% reduction in manual research

Real-World Applications

1. Customer Support Automation
   - Instant, accurate responses from documentation
   - Reduced ticket resolution time
2. Internal Knowledge Management
   - Unified access to company knowledge
   - Onboarding acceleration

Business Value

Immediate ROI Opportunities

- 30-50% reduction through caching
- 40% savings via model tiering
- $X monthly optimization potential
- 3x faster information retrieval
- 80% accuracy improvement
- 50% reduction in manual research

Real-World Applications

1. Customer Support Automation
   - Instant, accurate responses from documentation
   - Reduced ticket resolution time
2. Internal Knowledge Management
   - Unified access to company knowledge
   - Onboarding acceleration

Business Value

Immediate ROI Opportunities

- 30-50% reduction through caching
- 40% savings via model tiering
- $X monthly optimization potential
- 3x faster information retrieval
- 80% accuracy improvement
- 50% reduction in manual research

Real-World Applications

1. Customer Support Automation
   - Instant, accurate responses from documentation
   - Reduced ticket resolution time
2. Internal Knowledge Management
   - Unified access to company knowledge
   - Onboarding acceleration

Business Value

Immediate ROI Opportunities

- 30-50% reduction through caching
- 40% savings via model tiering
- $X monthly optimization potential
- 3x faster information retrieval
- 80% accuracy improvement
- 50% reduction in manual research

Real-World Applications

1. Customer Support Automation
   - Instant, accurate responses from documentation
   - Reduced ticket resolution time
2. Internal Knowledge Management
   - Unified access to company knowledge
   - Onboarding acceleration

Technical Capabilities Gained

Core Competencies Developed

- Document processing
- Embedding generation
- Vector search
- Response synthesis
- Hybrid search
- Reranking algorithms
- Query optimization
- Fine-tuning strategies

AWS Bedrock Expertise

- Claude 3 model family integration

Technical Capabilities Gained

Core Competencies Developed

- Document processing
- Embedding generation
- Vector search
- Response synthesis
- Hybrid search
- Reranking algorithms
- Query optimization
- Fine-tuning strategies

AWS Bedrock Expertise

- Claude 3 model family integration
- Cost-effective model selection

Technical Capabilities Gained

Core Competencies Developed

- Document processing
- Embedding generation
- Vector search
- Response synthesis
- Hybrid search
- Reranking algorithms
- Query optimization
- Fine-tuning strategies

AWS Bedrock Expertise

- Claude 3 model family integration
- Cost-effective model selection
- Production deployment patterns

Technical Capabilities Gained

Core Competencies Developed

- Document processing
- Embedding generation
- Vector search
- Response synthesis
- Hybrid search
- Reranking algorithms
- Query optimization
- Fine-tuning strategies

AWS Bedrock Expertise

- Claude 3 model family integration
- Cost-effective model selection
- Production deployment patterns
- Security best practices

Cost Analysis

Workshop Investment vs. Returns

Metric	Value
Workshop Investment	$12,000
Monthly Savings Potential	$5,000
Payback Period	2.4 months
First Year ROI	400%

Cost Optimization Strategies Learned

- Intelligent Caching: Reduce redundant API calls

Cost Analysis

Workshop Investment vs. Returns

Metric	Value
Workshop Investment	$12,000
Monthly Savings Potential	$5,000
Payback Period	2.4 months
First Year ROI	400%

Cost Optimization Strategies Learned

- Intelligent Caching: Reduce redundant API calls
- Model Tiering: Right-size models to queries

Cost Analysis

Workshop Investment vs. Returns

Metric	Value
Workshop Investment	$12,000
Monthly Savings Potential	$5,000
Payback Period	2.4 months
First Year ROI	400%

Cost Optimization Strategies Learned

- Intelligent Caching: Reduce redundant API calls
- Model Tiering: Right-size models to queries
- Batch Processing: Optimize throughput costs

Cost Analysis

Workshop Investment vs. Returns

Metric	Value
Workshop Investment	$12,000
Monthly Savings Potential	$5,000
Payback Period	2.4 months
First Year ROI	400%

Cost Optimization Strategies Learned

- Intelligent Caching: Reduce redundant API calls
- Model Tiering: Right-size models to queries
- Batch Processing: Optimize throughput costs
- Budget Monitoring: Proactive cost management

Implementation Roadmap

Phase 1: Pilot Project (Weeks 1-4)

- Select pilot use case (customer support)

Phase 2: Production Rollout (Weeks 5-8)

Phase 3: Expansion (Weeks 9-12)

Implementation Roadmap

Phase 1: Pilot Project (Weeks 1-4)

- Select pilot use case (customer support)
- Implement basic RAG pipeline

Phase 2: Production Rollout (Weeks 5-8)

Phase 3: Expansion (Weeks 9-12)

Implementation Roadmap

Phase 1: Pilot Project (Weeks 1-4)

- Select pilot use case (customer support)
- Implement basic RAG pipeline
- Measure baseline metrics

Phase 2: Production Rollout (Weeks 5-8)

Phase 3: Expansion (Weeks 9-12)

Implementation Roadmap

Phase 1: Pilot Project (Weeks 1-4)

- Select pilot use case (customer support)
- Implement basic RAG pipeline
- Measure baseline metrics
- Budget: $2,000 (AWS credits available)

Phase 2: Production Rollout (Weeks 5-8)

Phase 3: Expansion (Weeks 9-12)

Implementation Roadmap

Phase 1: Pilot Project (Weeks 1-4)

- Select pilot use case (customer support)
- Implement basic RAG pipeline
- Measure baseline metrics
- Budget: $2,000 (AWS credits available)

Phase 2: Production Rollout (Weeks 5-8)

- Scale to full dataset

Phase 3: Expansion (Weeks 9-12)

Implementation Roadmap

Phase 1: Pilot Project (Weeks 1-4)

- Select pilot use case (customer support)
- Implement basic RAG pipeline
- Measure baseline metrics
- Budget: $2,000 (AWS credits available)

Phase 2: Production Rollout (Weeks 5-8)

- Scale to full dataset
- Implement monitoring

Phase 3: Expansion (Weeks 9-12)

Implementation Roadmap

Phase 1: Pilot Project (Weeks 1-4)

- Select pilot use case (customer support)
- Implement basic RAG pipeline
- Measure baseline metrics
- Budget: $2,000 (AWS credits available)

Phase 2: Production Rollout (Weeks 5-8)

- Scale to full dataset
- Implement monitoring
- Deploy cost optimizations

Phase 3: Expansion (Weeks 9-12)

Implementation Roadmap

Phase 1: Pilot Project (Weeks 1-4)

- Select pilot use case (customer support)
- Implement basic RAG pipeline
- Measure baseline metrics
- Budget: $2,000 (AWS credits available)

Phase 2: Production Rollout (Weeks 5-8)

- Scale to full dataset
- Implement monitoring
- Deploy cost optimizations
- Budget: $5,000/month operational

Phase 3: Expansion (Weeks 9-12)

Implementation Roadmap

Phase 1: Pilot Project (Weeks 1-4)

- Select pilot use case (customer support)
- Implement basic RAG pipeline
- Measure baseline metrics
- Budget: $2,000 (AWS credits available)

Phase 2: Production Rollout (Weeks 5-8)

- Scale to full dataset
- Implement monitoring
- Deploy cost optimizations
- Budget: $5,000/month operational

Phase 3: Expansion (Weeks 9-12)

- Additional use cases

Implementation Roadmap

Phase 1: Pilot Project (Weeks 1-4)

- Select pilot use case (customer support)
- Implement basic RAG pipeline
- Measure baseline metrics
- Budget: $2,000 (AWS credits available)

Phase 2: Production Rollout (Weeks 5-8)

- Scale to full dataset
- Implement monitoring
- Deploy cost optimizations
- Budget: $5,000/month operational

Phase 3: Expansion (Weeks 9-12)

- Additional use cases

Implementation Roadmap

Phase 1: Pilot Project (Weeks 1-4)

- Select pilot use case (customer support)
- Implement basic RAG pipeline
- Measure baseline metrics
- Budget: $2,000 (AWS credits available)

Phase 2: Production Rollout (Weeks 5-8)

- Scale to full dataset
- Implement monitoring
- Deploy cost optimizations
- Budget: $5,000/month operational

Phase 3: Expansion (Weeks 9-12)

- Additional use cases

Implementation Roadmap

Phase 1: Pilot Project (Weeks 1-4)

- Select pilot use case (customer support)
- Implement basic RAG pipeline
- Measure baseline metrics
- Budget: $2,000 (AWS credits available)

Phase 2: Production Rollout (Weeks 5-8)

- Scale to full dataset
- Implement monitoring
- Deploy cost optimizations
- Budget: $5,000/month operational

Phase 3: Expansion (Weeks 9-12)

- Additional use cases

Team Impact

Skills Development

- Basic AI/ML understanding
- Limited GenAI exposure
- No production RAG experience
- Production-ready skills
- Cost optimization expertise
- Architecture best practices

Knowledge Transfer Plan

1. Internal tech talks (bi-weekly)

Team Impact

Skills Development

- Basic AI/ML understanding
- Limited GenAI exposure
- No production RAG experience

- Production-ready skills
- Cost optimization expertise
- Architecture best practices

Knowledge Transfer Plan

1. Internal tech talks (bi-weekly)
2. Documentation and runbooks

Team Impact

Skills Development

- Basic AI/ML understanding
- Limited GenAI exposure
- No production RAG experience

- Production-ready skills
- Cost optimization expertise
- Architecture best practices

Knowledge Transfer Plan

1. Internal tech talks (bi-weekly)
2. Documentation and runbooks
3. Pair programming sessions

Team Impact

Skills Development

- Basic AI/ML understanding
- Limited GenAI exposure
- No production RAG experience

- Production-ready skills
- Cost optimization expertise
- Architecture best practices

Knowledge Transfer Plan

1. Internal tech talks (bi-weekly)
2. Documentation and runbooks
3. Pair programming sessions
4. Code review standards

Competitive Advantage

Market Positioning

"By 2026, more than 80% of enterprises will have used GenAI APIs and models" – Gartner

- Early mover advantage in our sector

Risk Mitigation

Competitive Advantage

Market Positioning

"By 2026, more than 80% of enterprises will have used GenAI APIs and models" – Gartner

- Early mover advantage in our sector
- Differentiated product capabilities

Risk Mitigation

Competitive Advantage

Market Positioning

"By 2026, more than 80% of enterprises will have used GenAI APIs and models" – Gartner

- Early mover advantage in our sector
- Differentiated product capabilities
- Talent retention through cutting-edge tech

Risk Mitigation

Competitive Advantage

Market Positioning

"By 2026, more than 80% of enterprises will have used GenAI APIs and models" – Gartner

- Early mover advantage in our sector
- Differentiated product capabilities
- Talent retention through cutting-edge tech

Risk Mitigation

- Security best practices learned

Competitive Advantage

Market Positioning

"By 2026, more than 80% of enterprises will have used GenAI APIs and models" – Gartner

- Early mover advantage in our sector
- Differentiated product capabilities
- Talent retention through cutting-edge tech

Risk Mitigation

- Security best practices learned
- Cost control mechanisms

Competitive Advantage

Market Positioning

"By 2026, more than 80% of enterprises will have used GenAI APIs and models" – Gartner

- Early mover advantage in our sector
- Differentiated product capabilities
- Talent retention through cutting-edge tech

Risk Mitigation

- Security best practices learned
- Cost control mechanisms
- Vendor lock-in avoidance strategies

Competitive Advantage

Market Positioning

"By 2026, more than 80% of enterprises will have used GenAI APIs and models" – Gartner

- Early mover advantage in our sector
- Differentiated product capabilities
- Talent retention through cutting-edge tech

Risk Mitigation

- Security best practices learned
- Cost control mechanisms
- Vendor lock-in avoidance strategies
- Compliance considerations addressed

Recommendations

Immediate Actions

1. Approve pilot project (2 developers, 4 weeks)

Medium-term Strategy

Success Metrics

Recommendations

Immediate Actions

1. Approve pilot project (2 developers, 4 weeks)
2. Allocate AWS budget ($2,000 for pilot)

Medium-term Strategy

Success Metrics

Recommendations

Immediate Actions

1. Approve pilot project (2 developers, 4 weeks)
2. Allocate AWS budget ($2,000 for pilot)
3. Schedule weekly progress reviews

Medium-term Strategy

Success Metrics

Recommendations

Immediate Actions

1. Approve pilot project (2 developers, 4 weeks)
2. Allocate AWS budget ($2,000 for pilot)
3. Schedule weekly progress reviews

Medium-term Strategy

1. Establish AI Center of Excellence

Success Metrics

Recommendations

Immediate Actions

1. Approve pilot project (2 developers, 4 weeks)
2. Allocate AWS budget ($2,000 for pilot)
3. Schedule weekly progress reviews

Medium-term Strategy

1. Establish AI Center of Excellence
2. Budget for ongoing training (quarterly)

Success Metrics

Recommendations

Immediate Actions

1. Approve pilot project (2 developers, 4 weeks)
2. Allocate AWS budget ($2,000 for pilot)
3. Schedule weekly progress reviews

Medium-term Strategy

1. Establish AI Center of Excellence
2. Budget for ongoing training (quarterly)
3. Partner with AWS (credits, support)

Success Metrics

Recommendations

Immediate Actions

1. Approve pilot project (2 developers, 4 weeks)
2. Allocate AWS budget ($2,000 for pilot)
3. Schedule weekly progress reviews

Medium-term Strategy

1. Establish AI Center of Excellence
2. Budget for ongoing training (quarterly)
3. Partner with AWS (credits, support)

Success Metrics

- Cost reduction: Target 30% by Q3

Recommendations

Immediate Actions

1. Approve pilot project (2 developers, 4 weeks)
2. Allocate AWS budget ($2,000 for pilot)
3. Schedule weekly progress reviews

Medium-term Strategy

1. Establish AI Center of Excellence
2. Budget for ongoing training (quarterly)
3. Partner with AWS (credits, support)

Success Metrics

- Cost reduction: Target 30% by Q3
- Efficiency gain: 50% faster information access

Recommendations

Immediate Actions

1. Approve pilot project (2 developers, 4 weeks)
2. Allocate AWS budget ($2,000 for pilot)
3. Schedule weekly progress reviews

Medium-term Strategy

1. Establish AI Center of Excellence
2. Budget for ongoing training (quarterly)
3. Partner with AWS (credits, support)

Success Metrics

- Cost reduction: Target 30% by Q3
- Efficiency gain: 50% faster information access
- Customer satisfaction: +10 NPS points

Recommendations

Immediate Actions

1. Approve pilot project (2 developers, 4 weeks)
2. Allocate AWS budget ($2,000 for pilot)
3. Schedule weekly progress reviews

Medium-term Strategy

1. Establish AI Center of Excellence
2. Budget for ongoing training (quarterly)
3. Partner with AWS (credits, support)

Success Metrics

- Cost reduction: Target 30% by Q3
- Efficiency gain: 50% faster information access
- Customer satisfaction: +10 NPS points

Budget Proposal

Quarterly Investment Plan

Quarter	Investment	Expected Return	ROI
Q2 2025	$15,000	$15,000	100%
Q3 2025	$20,000	$45,000	225%
Q4 2025	$25,000	$75,000	300%
Q1 2026	$30,000	$120,000	400%

Resource Allocation

- 2 FTE for initial implementation

Budget Proposal

Quarterly Investment Plan

Quarter	Investment	Expected Return	ROI
Q2 2025	$15,000	$15,000	100%
Q3 2025	$20,000	$45,000	225%
Q4 2025	$25,000	$75,000	300%
Q1 2026	$30,000	$120,000	400%

Resource Allocation

- 2 FTE for initial implementation
- 20% time for knowledge transfer

Budget Proposal

Quarterly Investment Plan

Quarter	Investment	Expected Return	ROI
Q2 2025	$15,000	$15,000	100%
Q3 2025	$20,000	$45,000	225%
Q4 2025	$25,000	$75,000	300%
Q1 2026	$30,000	$120,000	400%

Resource Allocation

- 2 FTE for initial implementation
- 20% time for knowledge transfer
- AWS infrastructure budget

Budget Proposal

Quarterly Investment Plan

Quarter	Investment	Expected Return	ROI
Q2 2025	$15,000	$15,000	100%
Q3 2025	$20,000	$45,000	225%
Q4 2025	$25,000	$75,000	300%
Q1 2026	$30,000	$120,000	400%

Resource Allocation

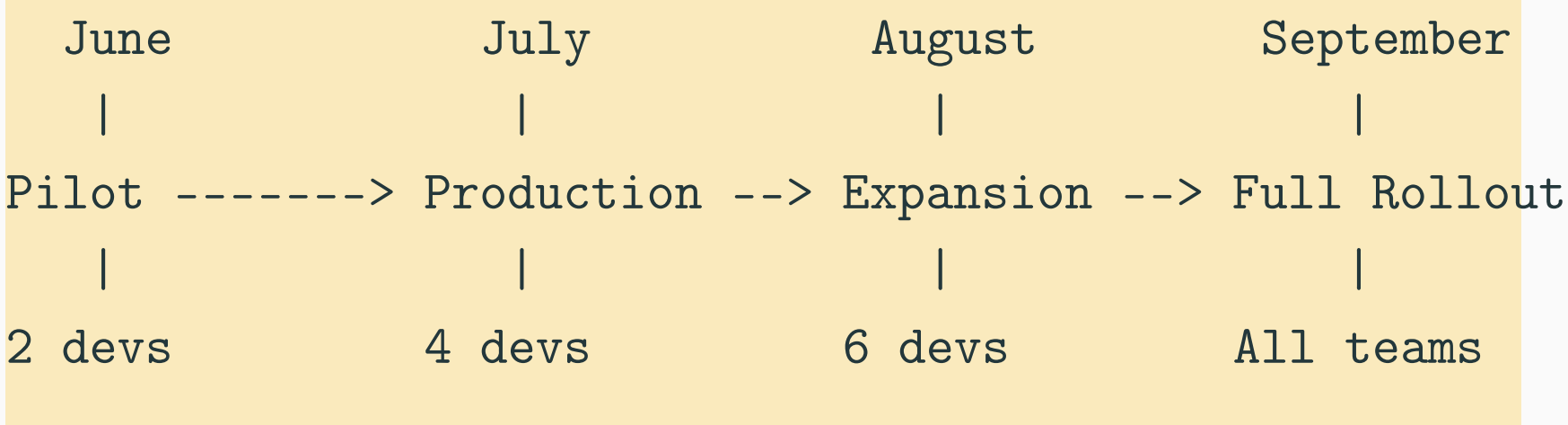
- 2 FTE for initial implementation
- 20% time for knowledge transfer
- AWS infrastructure budget
- Training and certification funds

Next Steps

Decision Points

1. ☐ Approve pilot project scope

Timeline

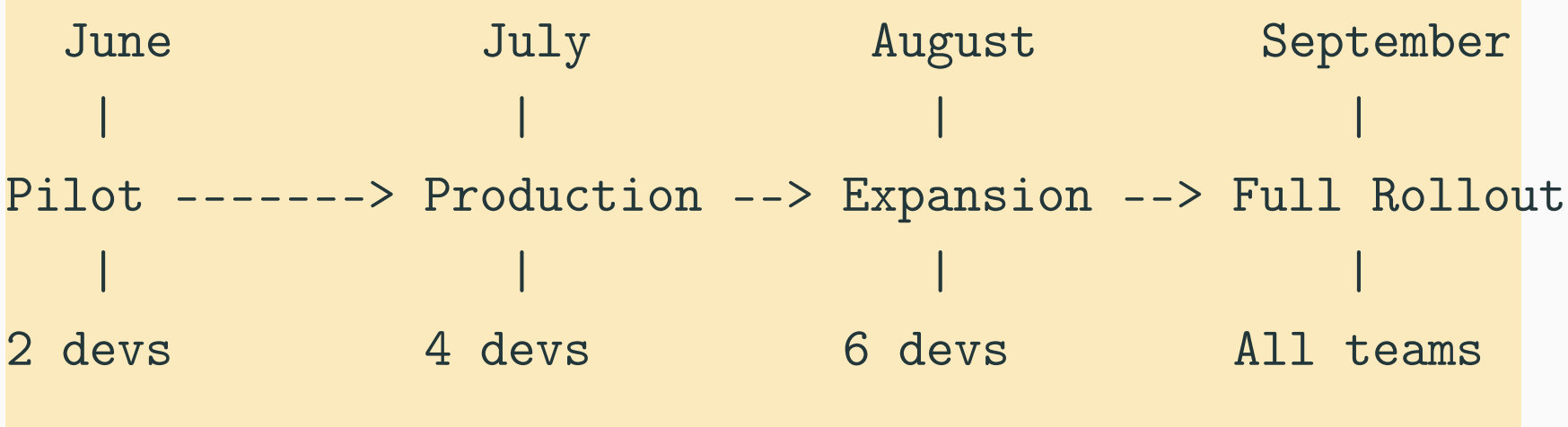


Next Steps

Decision Points

1. ☐ Approve pilot project scope
2. ☐ Allocate Q2 budget ($15,000)

Timeline

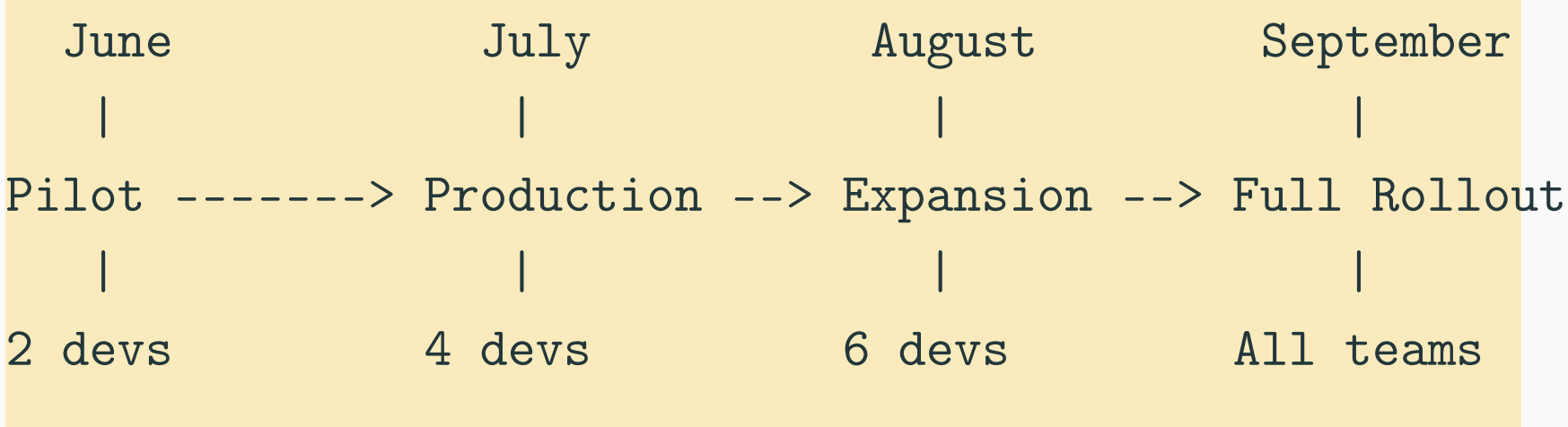


Next Steps

Decision Points

1. ☐ Approve pilot project scope
2. ☐ Allocate Q2 budget ($15,000)
3. ☐ Assign project team (2 FTE)

Timeline

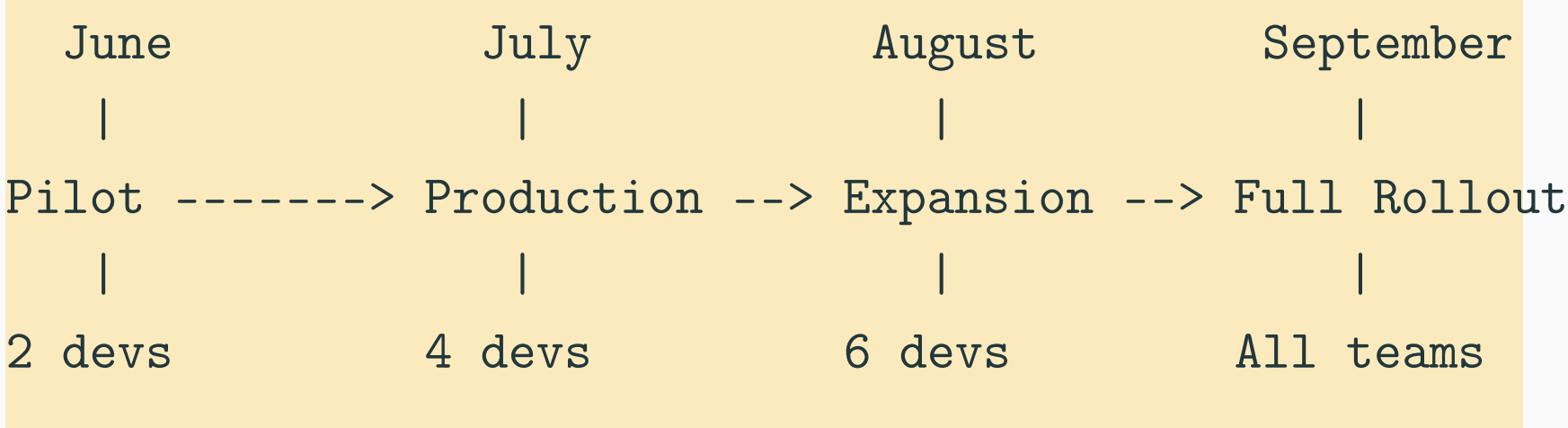


Next Steps

Decision Points

1. ☐ Approve pilot project scope
2. ☐ Allocate Q2 budget ($15,000)
3. ☐ Assign project team (2 FTE)
4. ☐ Schedule bi-weekly reviews

Timeline



Questions & Discussion

Key Takeaways

- Proven ROI: 400% first-year return

Ready to lead in the GenAI era?

Contact & Resources

- Workshop Materials: `github.com/company/rag-workshop`
- Technical Lead: [Name]
- Project Sponsor: [Name]
- AWS Contact: [Name]

Questions & Discussion

Key Takeaways

- Proven ROI: 400% first-year return
- Low Risk: Phased approach with checkpoints

Ready to lead in the GenAI era?

Contact & Resources

- Workshop Materials: `github.com/company/rag-workshop`
- Technical Lead: [Name]
- Project Sponsor: [Name]
- AWS Contact: [Name]

Questions & Discussion

Key Takeaways

- Proven ROI: 400% first-year return
- Low Risk: Phased approach with checkpoints
- High Impact: Customer & developer benefits

Ready to lead in the GenAI era?

Contact & Resources

- Workshop Materials: `github.com/company/rag-workshop`
- Technical Lead: [Name]
- Project Sponsor: [Name]
- AWS Contact: [Name]

Questions & Discussion

Key Takeaways

- Proven ROI: 400% first-year return
- Low Risk: Phased approach with checkpoints
- High Impact: Customer & developer benefits
- Strategic Value: Competitive differentiation

Ready to lead in the GenAI era?

Contact & Resources

- Workshop Materials: `github.com/company/rag-workshop`
- Technical Lead: [Name]
- Project Sponsor: [Name]
- AWS Contact: [Name]