Claude Model Selection Guide

Date: January 26, 2025
Purpose: Guide for choosing between Claude 3 Haiku and Claude 3.5 Sonnet
Status: Current Implementation: Claude 3 Haiku

🎯 Current Implementation

Model in Use: `claude-3-haiku-20240307`
Location: `lib/claude.ts`
API Route: `app/api/ai-assistant/route.ts`

📊 Model Comparison

Claude 3 Haiku (Current)

Best For:

- ✅ Lead qualification chatbots
- ✅ Simple Q&A responses
- ✅ Fast response times
- ✅ Cost-effective at scale
- ✅ High-volume interactions

Performance:

- **Speed:** Fastest (low latency)
- **Cost:** Lowest ($0.25 per 1M input tokens, $1.25 per 1M output tokens)
- **Context:** 200K tokens
- **Quality:** Good for structured, predictable responses

Use Cases:

- Customer support chatbots
- Lead qualification
- FAQ responses
- Simple automation tasks

Limitations:

- Less creative than Sonnet
- May struggle with complex reasoning
- Less nuanced responses

Claude 3.5 Sonnet (Upgrade Option)

Best For:

- ✅ Complex reasoning tasks
- ✅ Creative content generation
- ✅ Multi-step problem solving
- ✅ Nuanced conversations
- ✅ Advanced customer support

Performance:

- **Speed:** Moderate (higher latency than Haiku)
- **Cost:** Higher ($3 per 1M input tokens, $15 per 1M output tokens)

- **Context:** 200K tokens
- **Quality:** Superior for complex tasks

Use Cases:

- Advanced customer support
- Content generation
- Complex problem solving
- Multi-turn conversations requiring context

Limitations:

- Higher cost per request
- Slower response times
- May be overkill for simple Q&A

💰 Cost Comparison

Example: 1,000 Chat Messages

Claude 3 Haiku:

- Average input: 500 tokens
- Average output: 200 tokens
- Cost per message: ~$0.0004
- **Total for 1,000 messages: ~$0.40**

Claude 3.5 Sonnet:

- Average input: 500 tokens
- Average output: 200 tokens
- Cost per message: ~$0.0045
- **Total for 1,000 messages: ~$4.50**

Savings with Haiku: ~90% cost reduction

🔄 When to Upgrade to Sonnet

Consider upgrading if:

- ❌ Users report responses are too generic
- ❌ Complex questions aren't being answered well
- ❌ You need more creative/nuanced responses
- ❌ Multi-step reasoning is required
- ❌ Budget allows for 10x cost increase

Stick with Haiku if:

- ✅ Current responses are satisfactory
- ✅ Cost is a concern
- ✅ Fast response times are critical
- ✅ Simple Q&A is sufficient
- ✅ High volume is expected

🚀 How to Switch Models

Step 1: Update Model in `lib/claude.ts`

```
// Current (Haiku)
const MODEL = 'claude-3-haiku-20240307';

// To switch to Sonnet:
const MODEL = 'claude-3-5-sonnet-20241022';
```

Step 2: Verify API Access

Check Anthropic Console:

1. Visit: https://console.anthropic.com/
2. Check your plan tier
3. Verify Sonnet access (may require plan upgrade)

Common Errors:

- `402 Payment Required` → Upgrade plan
- `403 Forbidden` → Model not available on your plan
- `429 Rate Limit` → Too many requests

Step 3: Update Documentation

- Update `app/api/ai-assistant/route.ts` comment
- Update this guide
- Update deployment checklist

Step 4: Test Thoroughly

- Test response quality
- Monitor response times
- Check cost impact
- Verify error handling

📝 API Route Comment

Current Comment (Correct):

```
/**
 * Provides AI-powered lead qualification chatbot using Claude 3 Haiku API
 */
```

If Upgrading to Sonnet:

```
/**
 * Provides AI-powered lead qualification chatbot using Claude 3.5 Sonnet API
 */
```

🔍 Verification Steps

Check Current Model

1. Open `lib/claude.ts`
2. Find `const MODEL = ...`
3. Verify model name matches your plan

Check API Access

1. Visit Anthropic Console
2. Check plan details
3. Verify model availability
4. Check credit balance

Monitor Performance

- Response times
- Error rates
- Cost per request
- User satisfaction

⚠️ Important Notes

1. **Plan Requirements:**
   - Haiku: Available on all plans
   - Sonnet: May require plan upgrade
2. **Cost Impact:**
   - Sonnet is ~10x more expensive
   - Monitor usage carefully
   - Set up billing alerts
3. **Response Times:**
   - Haiku: ~200-500ms
   - Sonnet: ~500-1500ms
   - Consider user experience
4. **Error Handling:**
   - Both models can hit rate limits
   - Both can return 402/403 errors
   - Handle gracefully in code

📚 Resources

- **Anthropic Console:** https://console.anthropic.com/
- **Model Documentation:** https://docs.anthropic.com/claude/docs/models-overview
- **Pricing:** https://www.anthropic.com/pricing
- **API Reference:** https://docs.anthropic.com/claude/reference

---

✅ Decision Matrix

Factor	Haiku	Sonnet
Cost	⭐⭐⭐⭐⭐	⭐⭐
Speed	⭐⭐⭐⭐⭐	⭐⭐⭐
Quality (Simple)	⭐⭐⭐⭐	⭐⭐⭐⭐⭐
Quality (Complex)	⭐⭐⭐	⭐⭐⭐⭐⭐
Best For	High volume, simple Q&A	Complex reasoning, creativity

---

Last Updated: January 26, 2025
Current Model: Claude 3 Haiku
Next Review: When upgrading to Sonnet or if performance issues arise