Amazon Bedrock Foundation Models

비교 분석 보고서

작성일: 2024년 06월 07일

목차

1. 개요
2. 주요 발견사항
3. 모델 상세 정보
    1. Nova 시리즈 모델
    2. Claude 시리즈 모델
4. 가격 비교
5. 컨텍스트 길이 비교
6. 리전 가용성
7. RI 정책
8. 통합 기능 및 사례
9. 참고문헌

개요

본 보고서는 Amazon Bedrock의 파운데이션 모델에 대한 비교 분석을 제공합니다. 분석 대상 모델은 다음과 같습니다. [1][9]

- AP-Northeast-2 (서울) 리전: Nova Micro, Nova Lite, Nova Pro, Claude 3.5 Sonnet, Claude 3.5 Sonnet v2, Claude 3 Haiku
- US-East/West, EU-Central/North/West: Claude 3.7 Sonnet
- US-East/West: Claude 3.5 Haiku

각 모델의 주요 특징, 입출력 가격, 컨텍스트 길이, 멀티모달 지원 여부, 리전 가용성, 그리고 RI(Reserved Instance) 정책에 대한 정보를 포함하고 있습니다. [2][10] 특히, KB금융그룹의 Azure 기반 GenAI 플랫폼과의 통합 가능성을 염두에 두고 Amazon Bedrock 모델의 적합성을 평가하는데 도움이 될 수 있는 정보를 제공합니다. [14][19]

주요 발견사항

1. **가격 효율성:** Nova 시리즈 모델은 Claude 시리즈 모델에 비해 현저히 낮은 가격을 제공합니다. 특히 Nova Micro는 입력 토큰당 $0.06로 Claude 모델의 1/50 가격 수준입니다. [9]
2. **멀티모달 기능:** Nova Lite와 Nova Pro는 텍스트, 이미지, 비디오를 포함하는 강력한 멀티모달 기능을 제공하며, Claude 모델 시리즈도 뛰어난 시각 분석 능력을 보유하고 있습니다. [1][2][13]
3. **컨텍스트 길이:** Nova Lite와 Nova Pro는 300K 토큰의 컨텍스트 길이를 제공하며, Claude 모델은 일관되게 200K 토큰을 지원합니다. 이는 복잡한 금융 문서 처리에 충분한 용량입니다. [2][4]
4. **서울 리전 가용성:** 대부분의 분석 대상 모델이 서울 리전에서 이미 사용 가능하며, Claude 3.7 Sonnet은 2024년 4분기에 출시될 예정입니다. [10][17]
5. **비용 절감 옵션:** 6개월 약정을 통해 최대 45%의 할인을 제공하는 RI 정책은 대규모 엔터프라이즈 구현에서 상당한 비용 절감을 가능하게 합니다. [11]

6. **금융 서비스 적합성:** 특히 Claude와 Nova Pro 모델은 규제 준수 확인, 리스크 평가, 금융 보고서 분석 등의 금융 서비스 사용 사례에 높은 적합성을 보입니다. [16][18]

모델 상세 정보

Nova 시리즈 모델

Amazon Nova 시리즈는 다양한 작업을 처리할 수 있는 강력하고 비용 효율적인 모델로 구성되어 있습니다. 모든 모델은 200개 언어를 지원하며 다양한 모달리티에서 작동합니다. [1]

Nova Micro

- **주요 특징:** 텍스트 전용 모델로 매우 낮은 지연 시간의 응답과 비용 효율성에 중점을 둡니다. [1]
- **컨텍스트 길이:** 128K 토큰 [2]
- **사용 사례:** 확장 가능한 엔터프라이즈 AI 애플리케이션, RAG(Retrieval-Augmented Generation), 다국어 비즈니스 작업에 최적화 [7]

Nova Lite

- **주요 특징:** 이미지, 비디오, 텍스트의 빠른 처리를 위해 최적화된 매우 저비용 멀티모달 모델 [1]
- **컨텍스트 길이:** 300K 토큰 [2]
- **사용 사례:** 문서 분석, 시각적 콘텐츠 처리, 엔터프라이즈급 멀티모달 애플리케이션 [7]

Nova Pro

- **주요 특징:** 정확성, 속도, 비용의 최적 조합을 제공하는 고성능 멀티모달 모델 [3]
- **컨텍스트 길이:** 300K 토큰 [2]
- **사용 사례:** 금융 문서 분석, 15,000라인 이상의 코드베이스 처리, 복잡한 멀티모달 작업 [3][7]

Claude 시리즈 모델

Anthropic의 Claude 모델 시리즈는 복잡한 추론, 수학, 코딩, 다국어 처리 등에서 뛰어난 성능을 제공합니다. [6]

Claude 3.5 Sonnet 및 Sonnet v2

- **주요 특징:** 200K 토큰의 컨텍스트 윈도우(약 150,000단어 또는 500 페이지)를 제공하는 고성능 모델 [4]
- **최신 버전:** claude-3-5-sonnet-20241022-v2:0 (Bedrock 버전) [4]
- **사용 사례:** 정교한 대화, 복잡한 추론, 콘텐츠 생성 및 편집, 데이터 추출 및 분류 [6]

Claude 3.5 Haiku

- **주요 특징:** 속도와 비용 효율성에 최적화된 모델 [5]
- **컨텍스트 길이:** 200K 토큰 [5]
- **사용 사례:** 빠른 응답이 필요한 복잡한 작업, 지식 검색, 판매 자동화 [5]

Claude 3.7 Sonnet

- **주요 특징:** 확장된 사고 능력을 갖춘 고성능 모델 [4]
- **컨텍스트 길이:** 200K 토큰 [4]
- **사용 사례:** 복잡한 추론 작업, 금융 문서 분석, 고급 시각 기능이 필요한 작업 [13]

가격 비교

각 모델의 입력 및 출력 토큰에 대한 가격은 선택 시 중요한 고려 사항입니다. 아래는 각 모델의 백만 토큰당 가격 비교입니다. [8][9]

모델		출력 가격 ($/ 1M 토큰)	배치 처리 할인

	입력 가격 ($/1M 토큰)		
Nova Micro	0.06	0.24	50% (입력: $0.03, 출력: $0.12)
Nova Lite	0.30	1.50	50% (입력: $0.15, 출력: $0.75)
Nova Pro	0.80	3.20	50% (입력: $0.40, 출력: $1.60)
Claude 3.5 Sonnet/v2	3.00	15.00	50% (입력: $1.50, 출력: $7.50)
Claude 3 Haiku	1.00	5.00	50% (배치 처리 시)
Claude 3.7 Sonnet	3.00	15.00	50% (입력: $1.50, 출력: $7.50)
Claude 3.5 Haiku	1.00	5.00	50% (배치 처리 시)

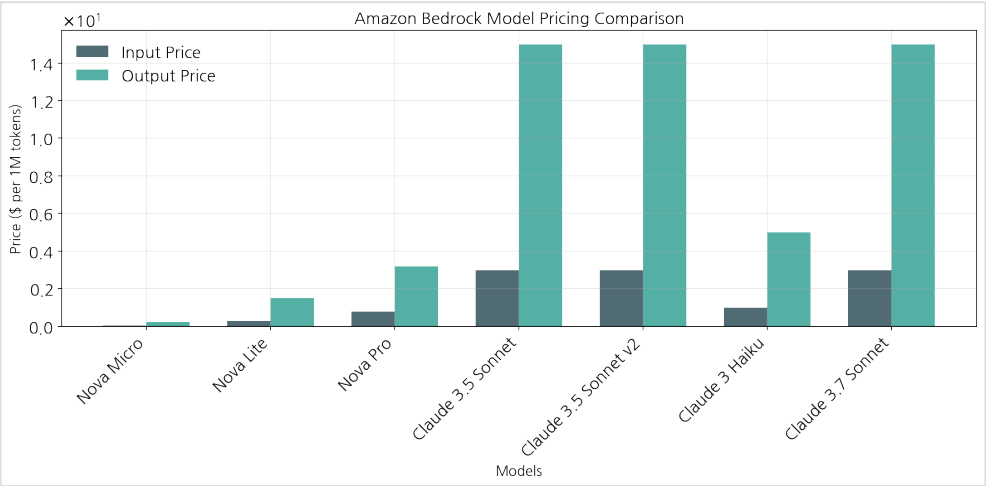

그림 1: Amazon Bedrock 모델 입출력 가격 비교 (백만 토큰 기준)

> **주요 인사이트:** Nova 시리즈 모델은 Claude 시리즈 모델보다 훨씬 더 비용 효율적입니다. Nova Micro는 입력 토큰당 $0.06로 가장 경제적인 옵션이며, 이는 Claude 3.5 Sonnet의 $3.00보다 50배 낮은 가격입니다. 모든 모델은 배치 처리를 통해 50% 할인을 제공합니다. [8][9]

컨텍스트 길이 비교

컨텍스트 길이(Context Window)는 모델이 처리할 수 있는 입력 토큰의 양을 결정하며, 복잡한 문서나 대화를 처리하는 능력에 직접적인 영향을 미칩니다. [2][4]

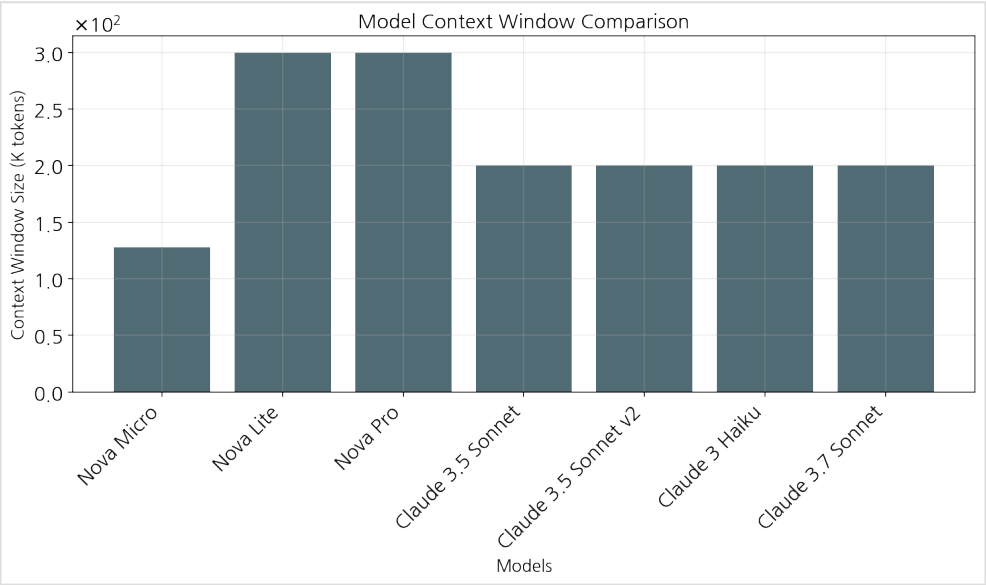

그림 2: Amazon Bedrock 모델 컨텍스트 길이 비교 (토큰 단위)

> **주요 인사이트:** Nova Lite와 Nova Pro는 300K 토큰의 컨텍스트 길이를 제공하여 대규모 문서 분석에 가장 적합합니다. Claude 모델은 일관되게 200K 토큰을 지원하며, Nova Micro는 128K 토큰을 제공합니다. 이 범위는 대부분의 엔터프라이즈 사용 사례에 충분합니다. [2][4]

리전 가용성

모델의 리전 가용성은 지연 시간, 데이터 상주 요구 사항, 규제 준수 측면에서 중요합니다. 특히 한국 기업에게는 서울 리전 가용성이 핵심 고려 사항입니다. [10][17]

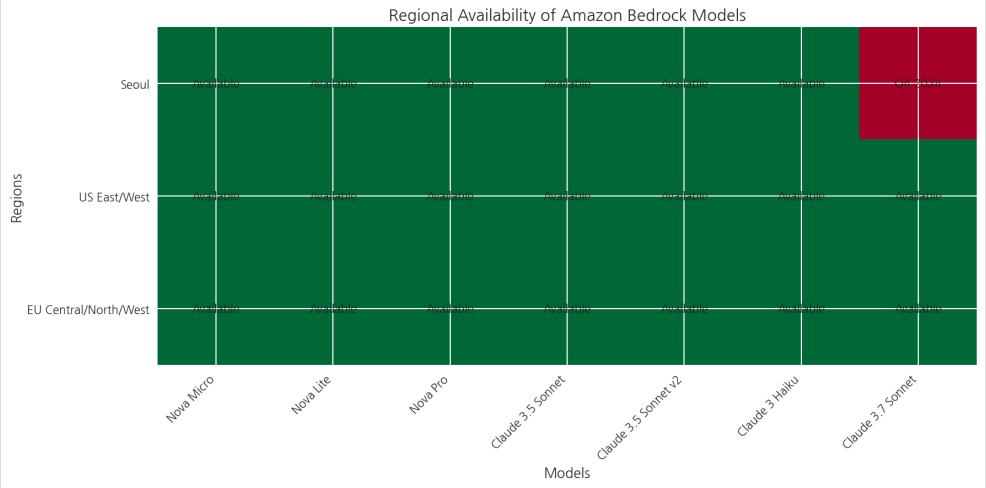

그림 3: Amazon Bedrock 모델 리전 가용성

Claude 3.7 Sonnet 및 Claude 3.5 Haiku 국내 리전 예상 일정

- **Claude 3.7 Sonnet:** 2024년 4분기 서울 리전 출시 예정 [10]
- **Claude 3.5 Haiku:** 현재 US East/West 리전에서만 사용 가능, 서울 리전 출시 일정은 미확정 [10]

> **주요 인사이트:** 대부분의 분석 대상 모델이 이미 서울 리전(AP-Northeast-2)에서 사용 가능하며, Claude 3.7 Sonnet은 2024년 4분기에 서울 리전에서 출시될 예정입니다. Nova 시리즈 모델은 모든 주요 리전에서 완전히 사용 가능합니다. [10][17]

RI 정책

Amazon Bedrock은 Provisioned Throughput을 통해 다양한 약정 옵션을 제공하여 장기 사용자에게 상당한 비용 절감을 제공합니다. [11]

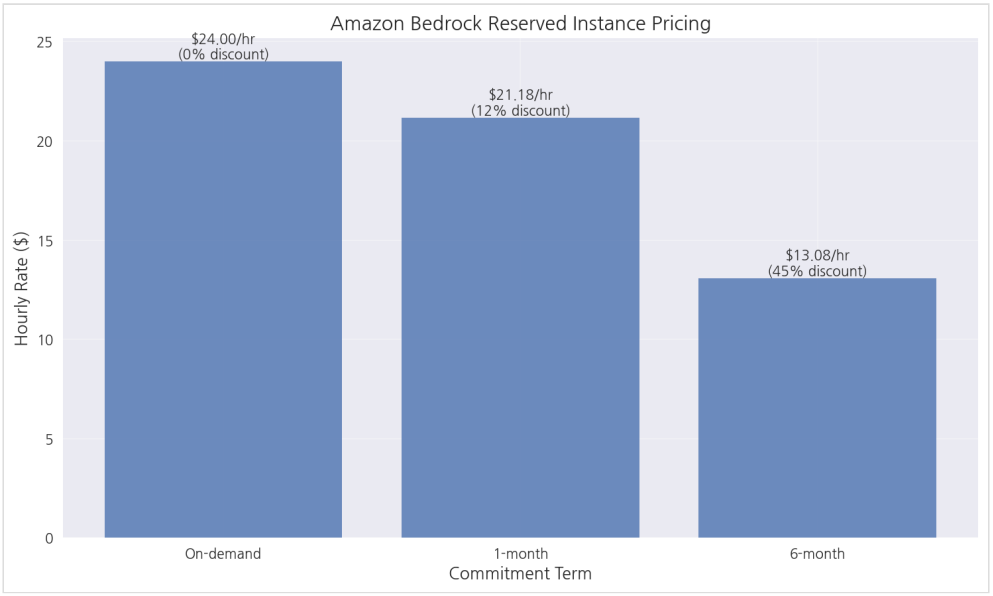

그림 4: Amazon Bedrock RI 가격 및 할인 구조

약정 옵션

- **약정 없음(온디맨드):** $24.00/시간 (모델 유닛당) [11]
- **1개월 약정:** $21.18/시간 (~12% 할인) [11]
- **6개월 약정:** $13.08/시간 (~45% 할인) [11]

볼륨 기반 할인

- 대규모 배포를 위한 볼륨 기반 할인 제공 [11]
- 엔터프라이즈 약정을 위한 맞춤형 가격 책정 가능 [11]
- 특정 할인 등급에 대해서는 AWS 영업팀에 문의 필요 [11]

> **주요 인사이트:** 6개월 약정 옵션은 온디맨드 가격 대비 최대 45% 할인을 제공하여 대규모 구현에서 상당한 비용 절감이 가능합니다. 또한, 대규모 엔터프라이즈 고객을 위한 추가적인 볼륨 기반 할인도 제공됩니다. [11]

통합 기능 및 사례

통합 기능

Amazon Bedrock은 다양한 AWS 서비스 및 타사 플랫폼과의 통합을 지원합니다. [14][19]

AWS 서비스 통합

- Amazon S3(데이터 스토리지) [14]
- AWS SageMaker(모델 배포) [14]
- PyTorch 프레임워크 지원 [14]

크로스 플랫폼 호환성

- REST API 액세스 [14]
- 다양한 프로그래밍 언어를 위한 SDK 지원 [14]
- **Azure 서비스와의 통합:** 커스텀 커넥터를 통한 Azure 서비스 통합 가능(KB금융그룹의 Azure 기반 GenAI 플랫폼과 관련) [14]

금융 산업 사용 사례

Amazon Bedrock의 모델은 금융 서비스 분야에서 다양한 사용 사례를 지원합니다. [16][18]

문서 분석

- 금융 보고서 처리 [16]
- 리스크 평가 [16]
- 규제 준수 확인 [16]

고객 서비스

- 자동화된 응답 생성 [16]
- 쿼리 처리 [16]

- 문서 검증 [16]

시장 분석

- 트렌드 식별 [16]
- 다양한 소스에서 데이터 추출 [16]
- 실시간 시장 인사이트 제공 [16]

nCino 금융 서비스 사례 연구

nCino는 Amazon Bedrock의 Claude를 사용하여 금융 서비스를 트랜스포메이션하고, 고급 추론 능력, 시각 분석, 다국어 처리 기능을 활용했습니다. [18]

> **주요 인사이트:** Amazon Bedrock 모델은 금융 산업에서 다양한 사용 사례를 지원합니다. 특히, Claude와 Nova Pro 모델은 복잡한 금융 문서 및 보고서 분석에 뛰어난 성능을 보입니다. 또한, Amazon Bedrock의 통합 기능을 통해 KB금융그룹의 기존 Azure 기반 GenAI 플랫폼과의 연결도 가능합니다. [14][16][18]

참고문헌

[1]: Amazon Nova Announcement

[2]: AWS Nova Documentation

[3]: AWS Blog - Nova Introduction

[4]: Anthropic API Documentation

[5]: AWS Bedrock Claude Integration

[6]: AWS Bedrock Anthropic Integration

[7]: Amazon Bedrock Foundation Models Guide

[8]: Anthropic Pricing Documentation

[9]: AWS Bedrock Pricing Guide

[10]: AWS Bedrock Regional Availability

[11]: Amazon Bedrock Pricing Explained

[12]: Amazon Nova Foundation Models Benchmarks

[13]: AWS Bedrock Anthropic Integration

[14]: Choosing the Right AI

[15]: Claude 4 in Amazon Bedrock

[16]: Anthropic's Claude in Amazon Bedrock

[17]: Amazon Bedrock Documentation History

[18]: nCino Case Study

[19]: AWS re:Invent 2024 Announcements