Amazon Bedrock Foundation Models 비교 분석

1. 개요

본 보고서는 Amazon Bedrock에서 제공하는 파운데이션 모델들의 특성, 요금, 컨텍스트 길이, 멀티모달 기능 및 지역 가용성에 대한 비교 분석을 제공합니다.[1] 분석 대상 모델은 Nova 시리즈(Micro, Lite, Pro)와 Claude 시리즈(3.5 Sonnet, 3.5 Sonnet v2, 3 Haiku, 3.7 Sonnet)입니다.[1][2]

Amazon Bedrock은 AWS에서 제공하는 서버리스 서비스로, 최고 수준의 파운데이션 모델(FM)을 API를 통해 사용할 수 있게 해주며, 이를 통해 생성형 AI 애플리케이션을 개발할 수 있습니다.[5] 본 보고서는 KB금융그룹이 Azure 기반의 GenAI 플랫폼 확장에 필요한 Amazon Bedrock 모델 검토를 위한 참고 자료로 작성되었습니다.

> **중요 사항:** 본 보고서에 포함된 정보는 작성 시점(2025년 5월)을 기준으로 한 것이며, AWS의 서비스 정책 변경에 따라 달라질 수 있습니다.[5][6]

2. 모델 비교 개요

Amazon Bedrock에서 제공하는 Nova 시리즈와 Claude 시리즈 모델은 성능, 가격, 기능 측면에서 다양한 특성을 가지고 있습니다.[1][2] 각 모델은 특정 사용 사례에 최적화되어 있으며, 요구사항에 따라 적절한 모델을 선택할 수 있습니다.[5]

Nova 시리즈는 Amazon에서 자체 개발한 모델로, 속도와 비용 효율성에 중점을 두고 있으며, Claude 시리즈는 Anthropic에서 개발한 모델로 높은 성능과 안전성을 제공합니다.[1][3]

3. 모델 상세 사양

3.1 모델명 및 주요 특징

각 모델의 주요 특징은 다음과 같습니다: [1][2][3][4]

모델	주요 특징
Nova Micro	- 텍스트 전용 모델 - 가장 빠른 응답 시간 - 200개 이상의 언어 지원 - 텍스트 요약, 번역, 콘텐츠 분류, 채팅, 수학적 추론에 최적화 [1][2]
Nova Lite	- 멀티모달 기능 (텍스트, 이미지, 비디오) - 매우 낮은 비용 - 빠른 처리 속도 - 200개 이상의 언어 지원 - 문서 지원: PDF, CSV, DOC, DOCX, XLS, XLSX, HTML, TXT, MD [1][2]
Nova Pro	- 고급 멀티모달 기능 - 정확성, 속도, 비용의 최적 밸런스 - 200개 이상의 언어 지원 - 문서 지원: PDF, CSV, DOC, DOCX, XLS, XLSX, HTML, TXT, MD - 에이전트 워크플로우 및 함수 호출에 탁월 [1][2]
Claude 3.5 Sonnet v2	- v1보다 2배 증가된 출력 용량(8K 토큰) - 높은 성능 - 균형 잡힌 모델 [3][4]
Claude 3.5 Sonnet	- 높은 성능 - 균형 잡힌 모델 [3][4]
Claude 3 Haiku	- 가장 빠르고 컴팩트한 모델 - 속도에 최적화된 컨텍스트 윈도우 [3][4]
Claude 3.7 Sonnet	- 최신 모델 - 향상된 기능 - 향상된 컨텍스트 처리 [3][4]

3.2 입력 및 출력 요금 (1M 토큰 기준)

각 모델의 입력 및 출력 요금은 다음과 같습니다: [5]

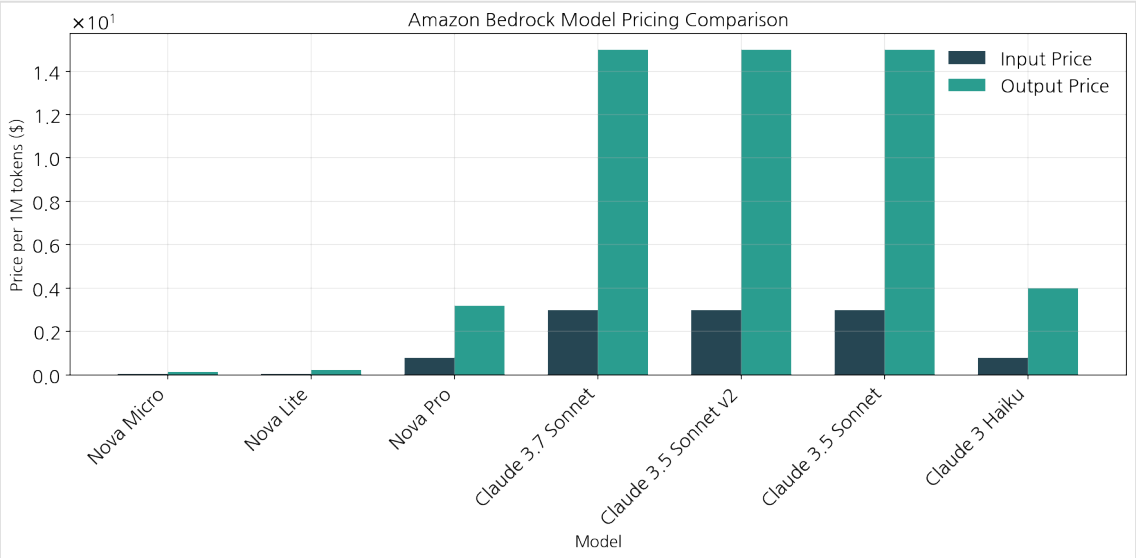

그림 1: Amazon Bedrock 모델 입력 및 출력 요금 비교 (1M 토큰 기준) [5]

위 차트에서 볼 수 있듯이: [5]

- Nova Micro는 가장 저렴한 모델로, 입력 토큰당 $0.04/1M, 출력 토큰당 $0.14/1M의 요금이 부과됩니다.
- Claude 3.7 Sonnet와 Claude 3.5 Sonnet은 가장 비싼 모델로, 입력 토큰당 $3.00/1M, 출력 토큰당 $15.00/1M의 요금이 부과됩니다.
- Nova Pro와 Claude 3 Haiku는 중간 범위의 가격대를 형성하고 있습니다.
- 모든 모델에서 출력 토큰 비용이 입력 토큰 비용보다 3-5배 높습니다.

추가적으로, 일괄 처리 모드(Batch mode)에서는 정상 요금보다 최대 50% 할인된 가격으로 모델을 사용할 수 있습니다: [5]

모델	일반 입력 요금	일반 출력 요금	일괄 처리 입력 요금	일괄 처리 출력 요금
Nova Micro	$0.04/1M	$0.14/1M	$0.02/1M	$0.07/1M
Nova Lite	$0.06/1M	$0.24/1M	$0.03/1M	$0.12/1M
Nova Pro	$0.80/1M	$3.20/1M	$0.40/1M	$1.60/1M
Claude 3.5 Sonnet	$3.00/1M	$15.00/1M	$1.50/1M	$7.50/1M

Claude 3 Haiku	$0.80/1M	$4.00/1M	$0.50/1M	$2.50/1M
Claude 3.7 Sonnet	$3.00/1M	$15.00/1M	해당 없음	해당 없음

3.3 컨텍스트 길이

각 모델의 컨텍스트 윈도우 크기는 다음과 같습니다: [1][2][3][4]

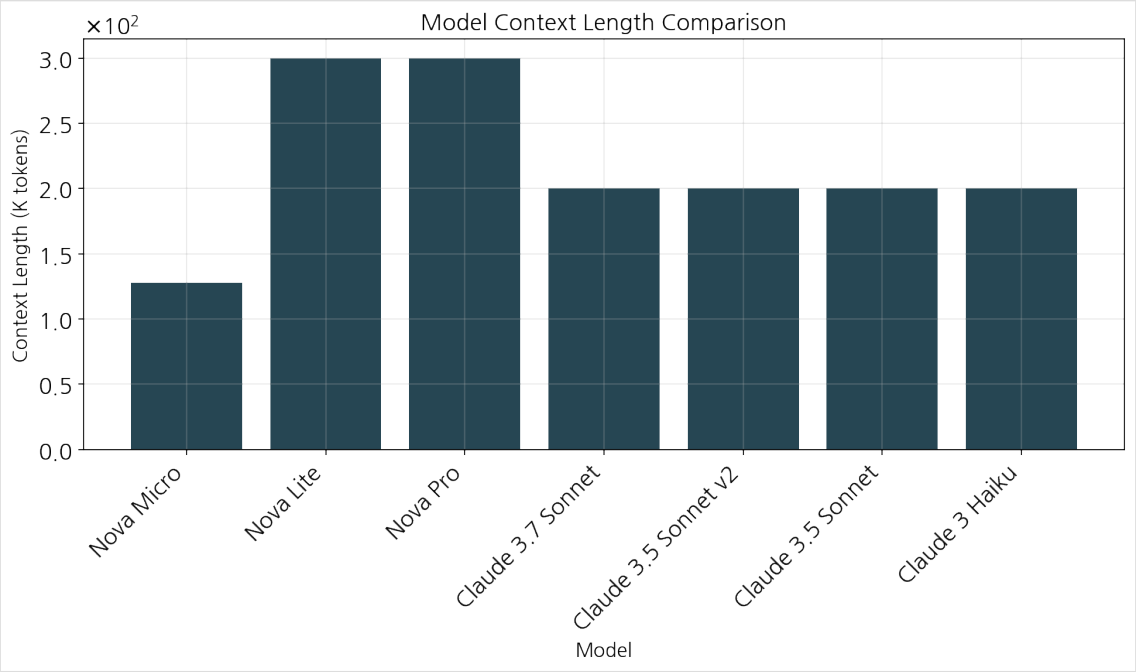

그림 2: Amazon Bedrock 모델별 컨텍스트 길이 비교 (K 토큰) [1][2][3][4]

컨텍스트 길이에 대한 주요 사항: [1][2]

- Nova Lite와 Nova Pro가 가장 긴 컨텍스트 윈도우를 제공하며, 300K 토큰까지 처리 가능합니다.
- Claude 시리즈 모델들은 200K 토큰의 컨텍스트 길이를 제공합니다.
- Nova Micro는 128K 토큰의 컨텍스트 길이를 제공합니다.
- Nova 시리즈 모델들은 모두 최대 5K 토큰의 출력을 지원합니다. [1]
- Claude 3.5 Sonnet v2는 v1에 비해 두 배 증가된 8K 토큰 출력 용량을 제공합니다. [3]

3.4 멀티모달 기능

각 모델의 멀티모달 기능은 다음과 같습니다: [1][2][3][4]

모델	텍스트	이미지	비디오	문서 처리
Nova Micro	✓	✗	✗	✗
Nova Lite	✓	✓	✓	✓
Nova Pro	✓	✓	✓	✓
Claude 시리즈	✓	✓	✗	✓

멀티모달 기능에 대한 추가 정보: [1][2][3]

- Nova Micro는 텍스트 전용 모델로, 이미지, 비디오 또는 문서 처리 기능이 없습니다.
- Nova Lite와 Nova Pro는 텍스트, 이미지, 비디오 입력을 처리할 수 있으며, 문서 처리 기능을 제공합니다.
- Claude 시리즈 모델들은 텍스트와 이미지를 처리할 수 있으며, 문서 분석 기능을 제공하지만 비디오 처리는 지원하지 않습니다.
- Nova Lite와 Nova Pro가 지원하는 문서 형식은 PDF, CSV, DOC, DOCX, XLS, XLSX, HTML, TXT, MD 등이 있습니다. [1]

3.5 지역 가용성

각 모델의 지역 가용성은 다음과 같습니다: [1][2][3][4]

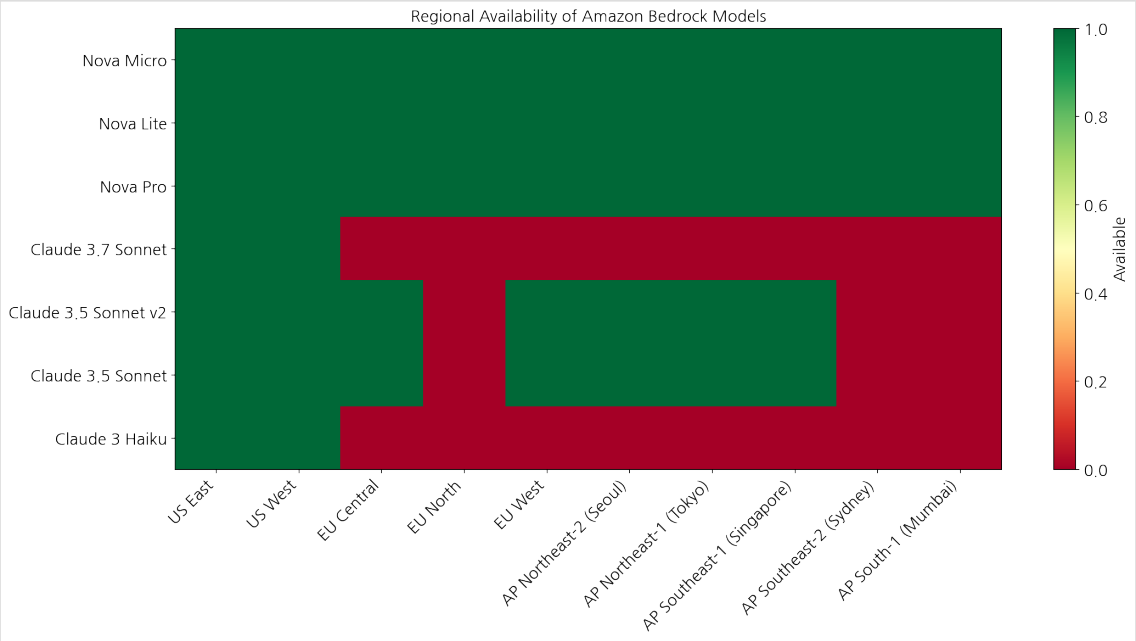

그림 3: Amazon Bedrock 모델의 지역별 가용성 [1][2][3][4]

지역 가용성에 대한 주요 사항: [1][2][3][4]

- Nova 시리즈 모델은 US East (N. Virginia), Asia Pacific (Tokyo)를 포함한 여러 지역에서 사용 가능합니다.
- Nova 시리즈는 다음 지역에서 크로스 리전 추론을 지원합니다:
  - US East (Ohio)
  - US West (Oregon)
  - Europe (Stockholm, Ireland, Frankfurt, Paris)
  - Asia Pacific (Tokyo, Singapore, Sydney, Seoul, Mumbai)
- Claude 3.5 Sonnet와 Claude 3.5 Sonnet v2는 Seoul 리전을 포함하여 다수의 지역에서 사용 가능합니다.
- Claude 3.7 Sonnet와 Claude 3 Haiku는 현재 US East/West 리전에서만 사용 가능합니다.
- Claude 3.7 Sonnet과 Claude 3 Haiku의 Seoul 리전 가용성은 아직 확정되지 않았습니다. [4]

3.6 RI 정책

Amazon Bedrock의 Reserved Instance (RI) 정책은 Provisioned Throughput이라는 형태로 제공됩니다: [5][6]

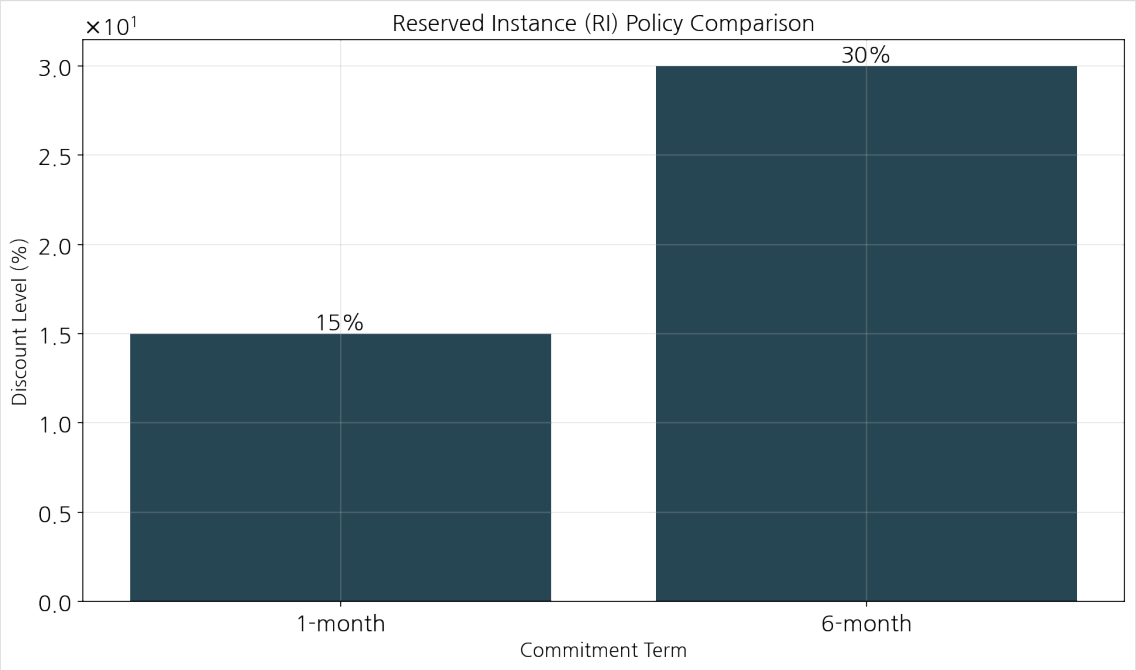

그림 4: Amazon Bedrock RI 정책 비교 [5][6]

RI 정책에 대한 주요 사항: [5][6]

- 약정 옵션:
  - 1개월 약정

    - 6개월 약정 (더 큰 할인 제공)
- 시간당 요금이 부과됩니다.
- 커스텀 모델 및 미세 조정된 모델에는 필수입니다.
- 지역 및 모델 유형에 따라 가격이 다릅니다.

추가적인 비용 최적화 기능:[5]

- 프롬프트 캐싱: 캐시된 토큰에 대해 최대 90% 할인
- 지능형 프롬프트 라우팅: 최대 30% 비용 절감
- 일괄 처리: 주문형 가격보다 최대 50% 할인

4. 결론

Amazon Bedrock은 다양한 요구사항에 맞는 여러 파운데이션 모델을 제공합니다.[1][5] 모델 선택 시 고려해야 할 주요 요소는 다음과 같습니다:

- 비용 효율성이 중요한 경우, Nova Micro와 Nova Lite가 가장 경제적인 옵션입니다.[5]
- 최고의 성능이 필요한 경우, Claude 3.7 Sonnet 또는 Claude 3.5 Sonnet이 적합합니다.[3][4]
- 긴 컨텍스트 처리가 필요한 경우, Nova Lite나 Nova Pro의 300K 토큰 컨텍스트 지원이 유리합니다.[1][2]
- 멀티모달 처리가 필요한 경우, Nova Lite, Nova Pro 또는 Claude 시리즈를 고려할 수 있습니다.[1][2][3]
- Seoul 리전에서 서비스 제공이 중요한 경우, 현재 사용 가능한 Nova 시리즈 및 Claude 3.5 Sonnet/v2를 선택하는 것이 좋습니다.[1][2][4]

KB금융그룹이 Azure 기반의 GenAI 플랫폼을 확장하는 과정에서, Amazon Bedrock의 다양한 모델은 특정 사용 사례에 맞는 옵션을 제공할 수 있습니다.[1][2][3][4][5] 특히 Nova 시리즈는 비용 효율성과 성능 사이의 균형을 찾는 데 도움이 될 것이며, Claude 시리즈는 높은 정확성과 안전성이 요구되는 작업에 적합할 것입니다.[1][2][3][4]

참고문헌

[1]: What is Amazon Nova?

[2]: Introducing Amazon Nova

[3]: Claude 3.5 Sonnet v2: Double Output Tokens on AWS Bedrock

[4]: AWS Bedrock Pricing (Metal Toad)

[5]: AWS Bedrock Pricing (Official)

[6]: AWS Bedrock Service Terms