



Model Migration (OpenAI to Bedrock)

Model-layer migration from the OpenAI API to Amazon Bedrock, with output parity validated before cutover and the surrounding application left intact.



Move your inference from the OpenAI API to Amazon Bedrock, with target model selection and output parity validated before you cut over.

Teams running on the OpenAI API often reach a point where cost, data residency, or governance pushes them toward AWS, but they do not want to re-architect the whole platform to get there. This engagement migrates the model layer to Amazon Bedrock, re-pointing your inference calls and validating output parity, while leaving the surrounding application intact.

- Target model selection across Bedrock, including GPT on Bedrock, Anthropic Claude, and Amazon Nova, matched to your quality and cost goals
- Prompt and API translation from the OpenAI SDK to the Bedrock Converse or OpenAI-compatible APIs
- Evaluation harness with side-by-side scoring to confirm output parity before cutover
- Guardrails, IAM, and invocation logging configured for governed production use
- Cost & latency comparison with a phased cutover plan

Timeline: 3 to 6 weeks.

Pricing: Scoped per engagement. Contact us.

Get Started Today

Contact us at sales@newmathdata.com to schedule an introduction.

Customer Commitment

- Engineering leads for prompt review, validation, and cutover sign-off
- Finance or cloud operations stakeholders for cost target alignment
- Access to current OpenAI usage, prompts, and representative evaluation data

Who Should Participate

- Engineering and architecture leads
- Finance or cloud operations stakeholders
- Product or business leadership

Benefits

- Lower inference cost and AWS-native governance without a platform rebuild
- Output parity validated against your prompts before cutover
- Flexibility across Bedrock model families, reducing single-provider lock-in

Deliverables

- Target model recommendation with cost and quality rationale
- Migrated inference layer on the Bedrock Converse or OpenAI-compatible APIs
- Evaluation harness and output parity report
- Guardrails, IAM, and logging configuration
- Cost, latency, and cutover plan