New Math DATA

aws PARTNER OF THE YEAR WINNER 2025

Model Migration (Gemini to Bedrock)

Model-layer migration from Google Gemini on Vertex AI to a Bedrock model family, with output parity validated before cutover.

Move your inference from Google Gemini to a model family on Amazon Bedrock, with prompt translation and output parity validated before you cut over.

Teams running on Google's Gemini models through Vertex AI often reach a point where consolidating on AWS makes sense for cost, data residency, or governance, but Gemini itself is not available on Bedrock. This engagement migrates the model layer to a Bedrock model family, translating your prompts and validating output parity, while leaving the surrounding application intact.

- ➔ Target model selection across Bedrock model families such as Anthropic Claude, Amazon Nova, and Meta Llama, matched to your quality and cost goals
- ➔ Prompt and API translation from the Gemini and Vertex AI SDKs to the Bedrock Converse API
- ➔ 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 Gemini usage, prompts, and representative evaluation data

Who Should Participate

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

Benefits

- AWS-native cost and governance without a platform rebuild
- Output parity validated against your prompts before cutover
- Choice 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 API
- Evaluation harness and output parity report
- Guardrails, IAM, and logging configuration
- Cost, latency, and cutover plan