What strategies can be employed to optimize the performance and scalability of a Generative AI model deployed on AWS, considering the potential computational and memory constraints of serverless architectures like AWS Lambda?