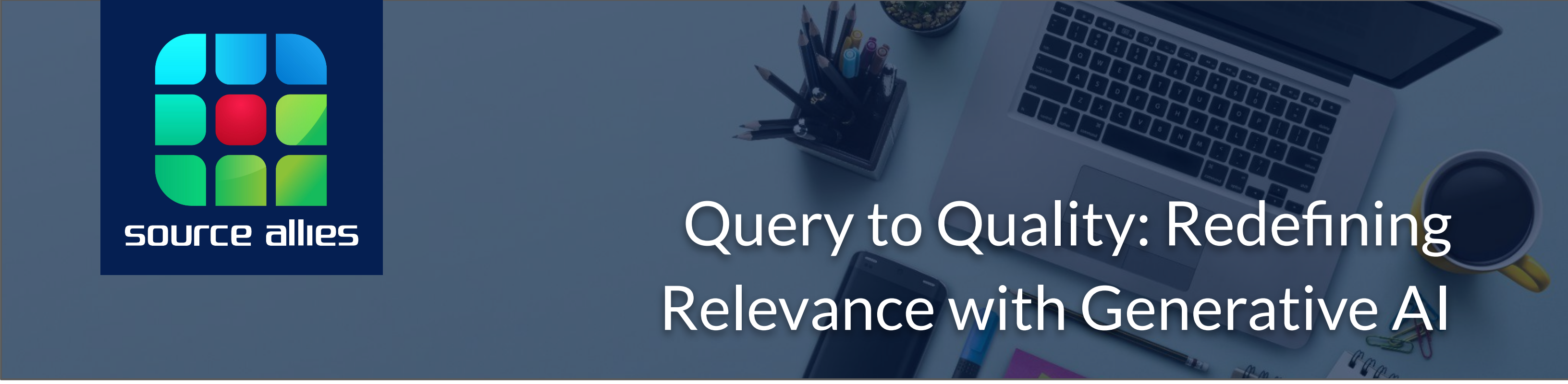

Query to Quality: Redefining Relevance with Generative AI

SUMMARY

Our client, a leading professional services company, wanted to implement an AI solution providing employees the ability to quickly and accurately find company information from an extensive wiki of user generated content. The establishment of trust and real-time observability in the solution was paramount to enable resolution of identified discrepancies, including identifying and remedying or removing outdated documents.

SOLUTION

We engineered a Retrieval Augmented Generation (RAG) solution to improve content discoverability and create a responsive interface for user queries. This provided quick access to pertinent information, insights into FAQ’s, and id of inaccurate responses, achieved through industry best practices like Test-Driven Development and Metrics-Driven Development.

We optimized processes by employing synthetic test data generation, streamlining validation procedures for Subject Matter Experts (SMEs). Presenting pre-prepared test sets for validation eliminated the need for SMEs to generate data from scratch. Leveraging LLM capabilities, we set up real-time evaluations of usage patterns, tracked changes over time, and analyzed metrics derived from our test sets. This approach saved time and expanded the corpus coverage, addressing commonly overlooked questions.

Tech Stack: Tech Stack: Amazon API Gateway, AWS Lambda, Amazon SQS, Amazon S3, Amazon DynamoDB, Amazon Bedrock, Claude v3 Sonnet model for 'chat', Titan v1 model for embeddings, Amazon RDS (Postgres), AwS Cognito.

RESULTS

200%

Increase in size of testset with introduction of Synthetic Testset Generation

6X

Metric-driven Development reduced the time from change to feedback

30%+

Increased accuracy compared to native RAG solution