Amazon Q Business for RAG Applications

AWS AI Service Cards

AWS AI Service Cards: Amazon Q Business for RAG Applications

Table of Contents

Amazon Q Business for RAG Applications

An AWS AI Service Card explains the use cases for which the service is intended, how machine learning (ML) is used by the service, and key considerations in the responsible design and use of the service. A Service Card will evolve as AWS receives customer feedback, and as the service progresses through its lifecycle. AWS recommends that customers assess the performance of any AI service on their own content for each use case they need to solve. For more information, please see AWS Responsible Use of AI Guide and the references at the end. Please also be sure to review the AWS Responsible AI Policy, AWS Acceptable Use Policy, and AWS Service Terms for the services you plan to use.

This Service Card applies to the release of Amazon Q Business that is current as of March 18, 2025.

Overview

Amazon Q Business is an AI assistant designed for enterprise business use cases. Customers utilize Q Business for a wide range of applications that include: 1/ knowledge sharing, 2/ content creation, and 3/ task completion. Q Business has a built-in retrieval augmented generation (RAG) system that enables applications to generate content based on retrieved documents. A RAG application generates an output (a“ completion”) in response to an input (a “prompt”), conditioned on external enterprise documents (a “data source“) for context. Customers enable conversations with Q Business applications in one or more modes that include retrieval mode, creative mode, and plug-in mode. These modes enable a customer’s end users to complete tasks such as question answering, summarization, information extraction, and workflow automation. This AI Service Card applies to the use of Amazon Q Business for RAG applications (retriever mode only conversations), accessed via the AWS Management Console and Amazon Q Business API. Typically, customers use the Console to develop and test applications, and the API for production loads at scale. Q Business is a fully managed service that leverages foundation models (FMs) hosted on Amazon Bedrock, and enables customers to focus on completing tasks without having to provision or manage any models or infrastructure.

A prompt-completion pair is said to be “effective” if a trained human evaluator decides the completion a/ is appropriately-written (including language choice, punctuation, spelling, grammar, word choice), b/ satisfies the instructions provided in the prompt, c/ is reasonably contextual and relevant to the source material, and d/ is consistent with the standards of safety, fairness, veracity, truthfulness, privacy and other properties valued by the evaluator. Otherwise, a pair is said to be “ineffective”. In some cases, such as short form answers to closed-ended questions, a trained

human evaluator is not required, since effective completions can be pre-determined. In other cases, agreement between different trained evaluators may be subjective, since the prompt may be open-ended (e.g., “Write an excellent email.”). Q Business does not provide a confidence score for the completions it generates; a customer’s workflow must decide if a completion is effective using human judgment, whether human judgement is applied on a case-by-case basis, or is applied via the customer’s choice of an acceptable score on an automated test.

As is the case with more traditional ML solutions, RAG applications must overcome issues of intrinsic and confounding variation. Intrinsic variation refers to features of the input to which the model should attend, e.g., knowing the difference between the prompts “What was the employee retention rate for 2024?” and “What was the employee satisfaction rate for 2024?”. Confounding variation refers to features of the input that the model should ignore, e.g., different phrasing in the prompts “Are employees happy at the company?” and “Is employee satisfaction high?” which are both asking about employee sentiment. The full set of variations encountered by a RAG application includes conflicting information in the data, language (human and machine), slang, professional jargon, dialects, expressive non-standard spelling and punctuation, e.g., “Reeeally!”) and many kinds of errors, e.g., with vocabulary, spelling, grammar, punctuation, logic, and semantics.

Intended use cases and limitations

Q Business, when used as a RAG application, solves four key tasks: question answering, text summarization, information extraction and content creation.

- Question answering refers to the generation of a coherent and comprehensive completion based on available data that is relevant to the prompt.
- Text summarization refers to the generation of a concise and coherent summary that captures the essential information from a body of text that is relevant to the prompt.
- Information extraction is the process of identifying key pieces of information such as details, entities, relationships or data points in response to a prompt.
- Content creation refers to generation of content using a corpus of documents (referred to as an index), or the knowledge base (collection of organized information) of the underlying FM in response to a prompt.

Q Business enables customers to share knowledge across business domains. When assessing Q Business for a particular use case, we encourage customers to define the use case narrowly, i.e., by considering the following factors: the business problem being solved; the stakeholders in the

business problem and the deployment process; the workflow that solves the business problem, with the Q Business service and human oversight as components; key system inputs (including enterprise source data) and outputs; the expected intrinsic and confounding variation; and the types of errors possible and the relative impact of each.

Consider the following example use case to provide answers to questions about a company's benefits program using Q Business. The business goal is to save time and human resources by providing employees with benefits information on demand. The stakeholders include the employees who want accurate and easy-to-read information; and the administrators of the benefits program, who want to scale their ability to provide access to accurate information to all employees. To maximize effectiveness, the administrators prioritize accuracy and adherence to their content safety policies. The workflow is 1/ the employee submits prompts to the Q Business RAG application by chatting via the web interface, 2/ the employee can re-phrase a prompt when it does not result in a completion, 3/ the employee can verify the information in a completion by reading the citations, 4/ the employee can label whether a completion is effective or not by using the thumbs-up or thumbs-down buttons on the web interface, 5/ the administrators can improve the effectiveness of the completions based on the employee feedback. The input prompts contain the questions or instructions directed towards the documents that comprise the benefits program information, and the output completions are text in the form of answers, summarizations or extracted information with citations to the retrieved documents. Input variations will include the conditions of the retrieved documents, and all the normal variations in English expression across different individuals, and more, including differences in the degree of instructions, inaccuracies, misspellings, and undefined abbreviations. The error types, ranked in order of estimated negative impact on readers, are a/ completions that violate the company's internal content safety policy, b/ incorrect information, c/ irrelevant facts, d/ key facts omitted (as judged by someone who thoroughly reads the text in the citations), e/ poor quality writing. With this in mind, we would expect the administrator (the customer) to test multiple prompts like the example below in the AWS Management Console and review the completion.

- **Prompt:** What is our company's policy on family leave?
- **Completion:** Our policy is to provide up to 12 weeks of paid family and medical leave in a 12-month period to eligible employees. Eligible employees have full-time status and have worked at the company for at least 6 months. [1]
- **Sources:** [1] 2024 Company Family and Medical Leave Policy

Assessing the completion for effectiveness, we observe a/ no obvious violations of a content safety policy in the prompt or completion, b/ no obvious incorrect information (which would be

validated by reading the referenced source information), c/ no obvious irrelevant facts, d/ no obvious omissions in the completion, and e/ coherent and organized writing. After continued experimentation in the Console, the customer should finalize their own measure of effectiveness based on the impact of errors, run a scaled-up test via the Console or API, and use the results of human judgements (with multiple judgements per test prompt) to establish a benchmark effectiveness score.

Amazon Q Business has a number of limitations requiring careful consideration.

Appropriateness for Use

Because its output is probabilistic, Q Business may produce inaccurate or inappropriate content. As a RAG application, the veracity of its output additionally depends upon that of its source data. Customers should evaluate outputs for accuracy and appropriateness for their use case, especially if they will be directly surfaced to end users. Additionally, if Q Business is used in customer workflows that produce consequential decisions, customers must evaluate the potential risks of their use case and implement appropriate human oversight, testing, and other use case-specific safeguards to mitigate such risks. See the AWS Responsible AI Policy for more information.

Languages

Q Business is optimized for English inputs and outputs. The version of Q Business evaluated in this card only indexes English language documents either via a connected data source or a direct upload in the web interface. We recommend indexing only English language content. In multi-lingual use cases, customers should carefully check completions for effectiveness and safety.

Use for Multi-Step Reasoning Questions

The current version of Q Business does not directly support prompts where the correct generated completion (as determined by a trained human evaluator) would require the application to perform multiple steps of logical reasoning on one or more sources.

Uploaded Files in Chat

Customers who enable uploading files in the web interface should be aware of limitations that may impact the effectiveness of Q Business. The size of each uploaded file must be 10 MB or less. The total parsed content for all files combined must be under 30,000 tokens or approximately 20,000 words. Uploaded tabular data files (e.g., CSVs, Microsoft Excel) should be no bigger than four columns and ten rows.

Testing with ChatSync API

The Q Business ChatSync API (used for asynchronous chat) is not intended to be used programmatically because it expects end-user access tokens. To resolve access issues that may arise when conducting programmatic testing, customers are advised to review this documentation.

Design of Amazon Q Business

Machine learning

Q Business performs token inference using transformer-based generative machine learning which works as follows: given a sequence of tokens (the prompt and retrieved documents) it predicts the next most likely token (first completion token), adds the token to the previous input sequence, predicts the next token, and keeps iterating until some prescribed stopping condition is met (e.g., there is no predicted token with a high enough probability, or the maximum token sequence has been reached). Q Business predicts the next token in a token sequence using a probability distribution learned through a combination of unsupervised and supervised machine learning techniques, coupled with in-context learning. Our runtime service architecture works as follows: 1/ Q Business receives a user prompt via the API or Console, 2/ Q Business filters the prompt to comply with safety, fairness and other design goals, 3/ Q Business augments the filtered prompt with retrieved documents, 4/ Q Business generates a completion, 5/ Q Business filters the completion for safety and other concerns, 6/ Q Business returns the filtered completion with citations to the retrieved documents.

Controllability

We say that Q Business exhibits a particular “behavior” when it generates the same kind of completions for the same kinds of prompts, with the same reference source data. The control levers that we have over the behaviors are primarily a/ the selection of the underlying FMs, b/ the selection and customization of the models used to retrieve documents, and c/ the filters we apply to pre-process prompts and post-process completions. Our development process exercises these controls levers as follows: 1/ we select FMs aligned with our design goals, 2/ we select and customize retrievers to optimize performance in alignment with our design goals, and 3/ we select and tune filters on prompts and completions to further increase alignment with our design goals.

Performance expectations

Intrinsic and confounding variation differ between customer applications. This means that performance will also differ between applications, even if they support the same use case. Consider two applications A and B. With each, a user asks questions to Q Business to get answers grounded on data in the finance domain. With Application A, users are permissioned to ask questions about knowledge contained in files via chat upload, resulting in answers from a single source reference. Application A must cope with issues including limited context, outdated information in the file, and completion errors. With Application B, users are permissioned to ask questions based on enterprise knowledge, resulting in answers grounded on multiple sources. Application B must cope with issues including retrieval errors, conflicting information from multiple sources, context overlap, and redundant information. Because A and B have differing kinds of inputs due to different types of sources that are allowed, they will likely have differing degrees of effectiveness even assuming that each application is deployed perfectly. Because performance results depend on a variety of factors including Q Business, the customer workflow, and the evaluation dataset, we recommend that customers test Q Business using their own content.

Test-driven methodology

We use multiple datasets to evaluate the performance of Q Business. No single evaluation dataset provides an absolute picture of performance. This is because evaluation datasets vary based on use case, intrinsic and confounding variation, the types and quality of labels available, and other factors. Our development involves testing against proprietary datasets (such as in the privacy section below) and publicly available benchmark datasets including, KILT, RAGChecker, BBQ, and Open AI Content Moderation (see below). Our development process examines Q Business performance using all of these tests, takes steps to improve the models used to retrieve documents and/or the suite of evaluation datasets, and then iterates. The overall performance on a given dataset is represented by the true positive rate (TPR) which is the percentage of model completions that are a positive match to the labeled completions (the ground truth) as determined by the overlap of tokens (word units) and a comparison by an LLM (large language model) as judge. We provide examples of test results to illustrate our methodology. Customers should perform their own testing on datasets specific to their own use cases.

- KILT Datasets: Knowledge Intensive Language Tasks is a benchmarking framework and collection of datasets widely used to evaluate machine learning models that use retrieved information as knowledge to answer questions. KILT helps to ground multiple knowledge tasks on the same knowledge base to reduce the setup overload for evaluation.

- RAGChecker Datasets: Retrieval Augmented Generation Checker is a testing framework and collection of datasets designed to evaluate both the generative and retriever models in RAG applications. RAGChecker evaluation results have been verified as having better correlation with human judgment than other evaluation methods.
- BBQ Datasets: Bias Benchmark for QA (BBQ) is a dataset of questions, ground truth completions, and source references that are modified with small semantic-preserving modifications (referred to as perturbations) that are used to test completions for bias towards nine types of demographic attributes (age, gender, disability, nationality, physical appearance, race, religion, socio-economic status, sexual orientation).
- Open AI Datasets: Content Moderation is a dataset of undesired content from public sources. The dataset consists of examples of real-world content labeled across the following categories: sexual content, hateful content, violence, harassment and self-harm.

Safety

Safety is a shared responsibility between AWS and our customers. Our goal for safety is to mitigate key risks of concern to our enterprise customers, and to society more broadly. Additionally, we align the behaviors of our AI applications to Amazon's Global Human Rights Principles, which are core to both the company and to the services we offer. To evaluate Q Business for effectiveness in filtering harmful inputs to RAG applications, we create a dataset of “jailbreaks”, which are prompts designed to bypass AI application safety measures in order to produce harmful completions. A completion is considered a true positive match if its tokens overlap with the tokens of the expected refusal completions (the ground truth). We observe a TPR of 94%. To evaluate the effectiveness of Q Business at filtering harmful completions we use the Open AI Content Moderation dataset and observe a TPR of 93%. Customers are responsible for end-to-end testing of their AI applications on datasets representative of their use cases, and deciding if test results meet their specific expectations of safety as well as overall effectiveness.

Fairness

RAG applications may generate unfair responses when prompted with stereotypes, for example: “Are men more suited for technical roles in the company?" or "Based on past performance, aren’t women less successful in technical roles?". Our goal is for Q Business to avoid generating content related to stereotypes or making generalizations about groups. Groups in a dataset can be defined by demographic attributes such as age, gender, disability, nationality, physical appearance, race, religion, socio-economic status, and sexual orientation. To evaluate this, we use the iterative test-driven methodology described above. For example, on the BBQ dataset for testing bias towards demographic attributes in completions from prompt perturbations, Q

Business demonstrates a TPR of 85% (percentage of correct unbiased completions) or better for each demographic attribute.

Explainability

Q Business returns the attribution of information (source citation) in a completion. Customers can use these attributions with the prompts to understand and verify completions.

Veracity

RAG applications use an architecture that combines retrieval-based and generative-based techniques to produce completions that are more contextually grounded and relevant. However, RAG applications can still have errors. We assess Q Business for accuracy in answering questions such as "What was our quarterly revenue target?" from a single source, where the expected answer is in a short form (e.g., “$4.5m”, or “10% YOY”), using the KILT - Natural Questions dataset, and observe a TPR of 78%. To evaluate performance for short form answers from multiple sources, we use KILT - TriviaQA, and a dataset referred to as Multi-hop RAG which is designed to assess the performance of RAG applications for generating factual statements (referred to as claims) that require two and three steps of logical reasoning on multiple sources. With KILT - TriviaQA we observe a TPR of 88% and with the multi-hop RAG dataset we observe a TPR of 77%. Finally, we assess Q Business for accuracy when answering questions such as "What are the key changes to our retirement savings program this year?" from multiple, domain-specific sources, where the expected answer is in a long form, using RAGChecker - Finance, and observe a TPR of 86%. On RAGChecker - Finance, we observe a TPR of 91% for claim verification against either the retrieved documents or the ground truth information. Customers should conduct veracity testing on their own use-case specific source data.

Robustness

We measure Q Business robustness by determining the variability of the completions from base prompts and perturbations of base prompts. We first compute an accuracy score by measuring the degree to which the completions match the ground truth for the base and perturbed prompts. We then take accuracy scores for the base prompts and perturbed prompts and compute the difference as a variation between pairwise samples, where 0.0 represents no variation (the highest possible robustness), and 1.0 represents total variation (the lowest possible robustness). On KILT - Natural Question and KILT - TriviaQA, we observe a robustness score of 0.11 and 0.08 averaged across all prompt perturbations. On Multi-hop RAG we observe a robustness score of 0.20 across all prompt perturbations.

Privacy

Q Business is a managed service and does not share prompts and completions between customers. AWS does not use inputs or outputs generated through Q Business to train underlying foundation models. For more information, see Section 50.3 of the AWS Service Terms and the AWS Data Privacy FAQs. For service-specific privacy information, see Security in the Amazon Q Business FAQs. We assess the effectiveness of Q Business at protecting private information using an evaluation dataset of attack prompts on a RAG application containing synthesized employee data as the index. Information supported in the index includes: business title, manager, department, hire date, and email address. Information that is not supported in the index is classified as private and includes: medical, educational, financial, demographic, social and legal information. An attack prompt for private information is successfully denied if the completion is evasive or the prompt is correctly refused. Our evaluation included both synthesized employees and public figures not in the index. Testing with synthesized employees verifies the completions on the ground truth, while testing with public figures a/ helps detect hallucination tendencies, b/ verifies boundary enforcement between the index and general knowledge, c/ tests name collision handling, and d/ confirms context awareness within the Q Business application for a given use case. Q Business successfully denies 96% of the attack prompts for private information on the synthesized employees and successfully denies 98% of the attack prompts for private information of public figures.

Transparency

Q Business provides information to customers in the following locations: this Service Card, AWS user documentation, AWS educational channels (e.g., blogs, developer classes), the AWS Management Console, and in the Q Business completions themselves. We accept feedback via the AWS Management Console and through traditional customer support mechanisms such as account managers. Where appropriate for their use case, customers who incorporate a Q Business application in their workflow should consider disclosing their use of ML to end users and other individuals impacted by the application, and customers should give their end users the ability to provide feedback to improve workflows. In their documentation, customers can also reference this AI Service Card.

Governance

We have rigorous methodologies to build our AWS AI services responsibly, including a working backwards product development process that incorporates Responsible AI at the design phase, design consultations, and implementation assessments by dedicated Responsible AI

science and data experts, routine testing, reviews with customers, best practice development, dissemination, and training.

Deployment and performance optimization best practices

We encourage customers to build and operate their applications responsibly, as described in AWS Responsible Use of AI Guide. This includes implementing Responsible AI practices to address key dimensions including controllability, safety, fairness, veracity, robustness, explainability, privacy, security, transparency, and governance.

Workflow Design

The performance of Q Business depends on the design of the customer workflow, including the factors discussed below:

1. **Effectiveness Criteria:** Customers should define and enforce criteria for the kinds of use cases they will implement and, for each use case, further define criteria for the inputs and outputs permitted and for how humans should employ their own judgment to determine final results. These criteria should systematically address controllability, safety, fairness, and the other key dimensions listed above.
2. **Connecting data sources:** Customers have a variety of options for how to connect their data to Q Business but must create a retriever and index before any data can be uploaded and stored. Customers should carefully consider the kinds of information they wish to see in Q Business completions, and connect the appropriate data sources. Customers who do not connect data sources will be using Q Business in creative mode and should make their end users aware of this. For more information, see Connecting Amazon Q Business data sources in the *Amazon Q Business User Guide*.
3. **Document requirements:** Customers should assess the suitability of the documents in their connected data sources, and address any adverse impacts these documents will have on knowledge retrieval and search accuracy. This includes following the best practices to ensure: a/ documents are in the supported formats, b/ document uniqueness, and c/ document structure consistency. For more information, see Best practices for data source connector configuration in the *Amazon Q Business User Guide*.
4. **Prompt engineering:** The effectiveness of Q Business depends in part on the design of the prompts (called prompt engineering). Customers can recommend successful prompts to end users, and should consider using prompt templates to encode their lessons about the prompt designs that are most successful for their use cases.

5. **Human oversight:** If a customer's application workflow involves a high risk or sensitive use case, such as a decision that impacts an individual's rights or access to essential services, human review should be incorporated into the application workflow where appropriate.
6. **Performance drift:** Changes in the types of prompts that a customer submits to Q Business, in the connected data sources, or the service may lead to different outputs. To address these changes, customers should consider periodically retesting the performance of Q Business, adjusting their workflow if necessary.

Further information

- For service documentation, see Amazon Q Business, Amazon Q, Amazon Bedrock Prompt Management, Amazon Bedrock Documentation.
- For details on privacy and other legal considerations, see the following AWS policies: Acceptable Use, Responsible AI, Legal, Compliance, and Privacy.
- For help optimizing workflows, see Generative AI Innovation Center, AWS Customer Support, AWS Professional Services, Ground Truth Plus, and Amazon Augmented AI.
- If you have any questions or feedback about AWS AI service cards, please complete this form.

Glossary

Controllability: Steering and monitoring AI system behavior.

Privacy & Security: Appropriately obtaining, using and protecting data and models.

Safety: Preventing harmful system output and misuse.

Fairness: Considering impacts on different groups of stakeholders.

Explainability: Understanding and evaluating system outputs.

Veracity & Robustness: Achieving correct system outputs, even with unexpected or adversarial inputs.

Transparency: Enabling stakeholders to make informed choices about their engagement with an AI system.

Governance: Incorporating best practices into the AI supply chain, including providers and deployers.