Tab 1

Google Cloud

HOW TO USE

(1) Create a copy of this document
(2) Complete a 'Find and Replace' for the following:

Placeholder	Description
[Customer Legal Name]	Legal customer business name
[Customer Name]	Dba (Doing business as) name for customer
[SOW Date]	Effective Date
[Start Date]	Project start date
[End Date]	Project end date
[Author]	Name and email of Bitstrapped member who wrote SOW

(3) Delete the contents of this page (Page 1)

Bitstrapped Inc.

[Customer Name]

Gen AI Model and Model Version Evaluation Framework

[SOW Date]

[Author]

Bitstrapped

[Customer Logo]

Statement of Work

This Statement of Work ("SOW") is agreed upon and entered into by and between Bitstrapped Inc. on behalf of itself and its affiliates (hereinafter referred to as ("Bitstrapped") and [Customer Legal Name] (hereinafter referred to as "[Customer Name]" or ("Customer"). This Statement of Work shall be effective on the date of the last signature below, governed by the terms of this SOW, and supersedes any other agreement between the parties related to this subject matter.

Contents

1. Executive Summary

This Statement of Work (SOW) covers an implementation pilot of a Generative AI (Gen AI) solution for LLM performance evaluation and observability. The solution enables organizations building Generative AI or LLM-powered products to monitor the performance of their chosen LLM model, evaluate performance across models and model versions, and compare prompt performance optimizations across models.

The objective is to help Customer adopt the Google Cloud Platform (GCP) to take advantage of Google's open approach to Gen AI models and interoperability through Vertex AI Model Garden. This is achieved by streamlining the process of adopting and testing new models and model versions and avoiding issues with model quality or user experience when changing underlying LLM technologies.

The framework implemented as part of this SOW will enable Customer to overcome common challenges with evaluating and selecting optimal large language models (LLMs), such as measuring consistency of results and sensitivity to context. The deployment will include standard testing protocols, automated monitoring, and feedback loops to detect performance drops and support data-driven model decisions.

Built on GCP, the solution will combine native GCP services and open-source tools into a reusable, customizable framework designed to meet observability standards.

1.1. Project Overview

Bitstrapped will partner with Customer to deploy the Gen AI evaluation solution, ensuring it is functional in all its components to help address LLM evaluation and performance challenges. Bitstrapped will lead the solution deployment on GCP for a production pilot.

This deployment will include a layered architecture supporting:

1. Automated drift detection, with alerting, to flag performance drops.
2. Evaluation pipeline to compare the performance of different models. This includes the ability to utilize the ground truth or human feedback (if either of these are available) or LLM-based evaluation, depending on the availability of data or client-side capture.
3. An ingestion point, storage, and integration of user feedback in the evaluation
4. Visualization for evaluation insights

The framework will support batch evaluation, giving teams ongoing visibility into model quality and enabling safe, data-driven improvements.

1.2. Objectives

- Enhance existing infrastructure to customize a GCP-based Gen AI framework for assessing and monitoring LLM outputs, achieving greater transparency and insight into model behaviour.
- Establish drift detection mechanisms with configurable alert thresholds.
- Create automated evaluation pipelines using both metric-based and LLM-as-judge methodologies.
- Integrate human feedback loops from both domain experts and end-users.
- Provide visualization capabilities for performance metrics and evaluation results.
- Deploy the capabilities using Infrastructure as Code (IaC) with Terraform.

SOW Start Date	[Start Date]
SOW End Date (estimated)	[End Date]

2. Constraints

Constraints establish predefined boundaries to manage this engagement's scope, resources, and expectations. The following table outlines the constraints for this engagement:

Constraint	Description	Limit	Inclusion
LLM Use Cases	The # of existing applications or use cases of LLMs supported	1	Deployment will be piloted with one pre-existing LLM solution
Supported Prompts	The # of existing prompts supported	3	Deployment will be piloted for three pre-existing prompts in the supported use case.
Input Data Format	File formats to run evaluation	n/a	CSV or JSONL file(s)
Human Feedback	Format of human feedback supported	Binary or Numeric	An API will provide an ingestion point for human feedback, which will be stored and integrated with evaluation. It will support binary or numeric feedback scores. Customer is responsible for integrating the ingestion point to the source of feedback, such as the client side, and/or any migration of baseline data.
LLM Models Evaluated	# of LLM models to evaluate	3	The deployment will be piloted using three suitable LLM models to evaluate performance against
Metrics	# of metrics tracked for evaluation	3	The deployment will produce up to three metrics to be consumed in the visualization
Dashboards	# of visualization dashboards	1	The deployment will be piloted with 1 dashboard visualization to consume evaluation insights
Observability Platform	# of monitoring systems for tracking model behaviour	1	One of: Arize, Latitude, or Confident-AI, based on requirements analysis Monitoring setup for 1 to 2 metrics of interest (could be custom or standard metrics, or monitoring the quantitative values representing the human feedback)

3. Functional Requirements

This section identifies the corresponding solution components:

ID	Requirement	Solution Component
FR-001	**Automated Drift Detection** Monitor drift in the model outputs	Vertex AI Evaluation Service and/or one of the following - Arize, Latitude, Confident-AI
FR-002	**LLM Performance Ingestion** Ingestion points to integrate existing LLM prompt, response, and model metadata for persistence	TBD
FR-003	**Evaluation Pipeline** Each LLM model should consume the prompts included to produce the supported metrics for evaluation. Evaluation will consider: -Drift detection (Batch) -Evaluation (Batch) -Human Feedback (Batch)	Vertex AI Pipelines + GKE + BigQuery
FR-004	**LLM-as-Judge Evaluation** Secondary LLMs should evaluate model outputs against defined criteria.	GKE + Vertex AI Endpoint
FR-005	**Human Feedback Ingestion** Ingestion points to ingest, process, and incorporate human feedback.	Cloud Functions + GCS / Firestore / BigQuery
~~FR-006~~	~~**Prompt Versioning and Catalogue**~~ ~~Maintain a versioned repository of prompts.~~	~~GCS~~
FR-006	**Evaluation History** Record model performance by associating model type, version, prompts, responses, human feedback, and evaluation metrics.	BigQuery
FR-007	**Performance Visualization** Generate artifacts for reviewing evaluation metrics	~~BigQuery + Looker or Looker Studio~~ GCS

Non-Functional Requirements

The following section outlines the quality attributes and constraints that the solution will meet:

ID	Requirement	Solution Component
NFR-001	**Scalability** – Ability to scale to additional LLM use cases, as well as integrate additional models and sources of human feedback This should allow Customer to extend this solution to support other LLM use cases.	GKE
NFR-002	**Reliability** – Ensure consistent evaluation results with error handling and recovery mechanisms for evaluation jobs.	GKE, Cloud Functions
NFR-003	**Data Privacy and Security** – Enforce security measures, including data encryption, access control.	Cloud IAM, Cloud Storage Security, Cloud Logging, VPC
NFR-004	**Extensibility** – Solution must support Model-agnostic evaluation with minimal code changes.	GKE, Cloud Functions

4. Scope of Services

4.1. Success Criteria

The success criteria define the measurable outcomes and benchmarks that indicate the achievement of project goals and satisfactory delivery of the agreed-upon solution, which include the following:

- Complete a pilot of the LLM evaluation deployment with demonstrated functionality for supported use cases, LLM models, and prompts.
- Automated drift detection produces alerts.
- Evaluation pipelines can be triggered both manually and through event-based automation.
- Access to visualization for comparing and benchmarking the computed performance metrics and evaluation results.
- The deployed solution satisfies the non-functional requirements.
- Documentation for future customization and knowledge transfer will enable the customer team to operate and extend the framework.

4.2. Activities

The following section outlines the detailed activities planned to achieve the project objectives and deliverables effectively:

Phase	Activities
Requirements Analysis	• Understand the specific challenges with the Customer LLM deployment that the solution aims to address and the expected outcomes. • Identify existing data sources, technology stack, and workflows related to the project scope. • Review data sets in scope. • Conduct a detailed assessment of evaluation needs and quality standards. • Define specific metrics for automated and human-based assessment. • Define data schemas for prompt storage, evaluation results, and performance tracking. • Design alert thresholds and notification workflow.
Core Infrastructure Implementation	• Provision required GCP resources and configure access controls. • Set up Cloud Storage buckets for prompt cataloging. • Configure BigQuery datasets for evaluation metadata and results. • Configure base integrations between components.
Evaluation Pipeline Development	• Deploy a Vertex AI Pipeline for model evaluation. • Deploy drift detection and alerting mechanisms. • Implement GKE service for LLM-as-judge evaluation. • ~~Provision prompt management and versioning system.~~ • Build API for feedback ingestion (Batch processing job). • Implement visualization of the computed performance metrics and assessment results

Integration and Testing	• Integrate all components into a cohesive workflow. • Integrate existing LLM use case with storage of prompt, response, and metadata • Test full evaluation pipelines with sample prompts and models. • Validate drift detection with simulated anomalies. • Verify monitoring alert propagation and pipeline triggering. • Test human feedback integration into the monitoring and evaluation pipelines.
Production Release	• Environment separation is deployed with Terraform • Supported Repository and Git Branches for Prod deployments
Knowledge Transfer	• Develop comprehensive architecture documentation. • Create user guides for framework operation. • Prepare customization documentation for future deployments. • Conduct weekly knowledge sharing during implementation. • Deliver the final solution with a hands-on demonstration of the workflow scenarios. • Provide handover documentation with operational instructions.
Project Management	• Project Planning and Pre-requisites • Project Kick-Off meeting • Weekly team (internal/external) meetings • Weekly project activity and meeting coordination • Customer team and stakeholder collaboration • Weekly status updates and reporting

4.3. Deliverables

The customer can expect the following key deliverables from this engagement:

- Full implementation of the evaluation framework with all components as a solution pilot, including:
  - Monitoring implementation with alert configuration.
  - Vertex AI Pipelines for automated model assessment.
  - LLM-as-Judge service for automated output scoring.
  - Generation of artifacts for the visualization of metrics.
  - Integrated human feedback
- Fork of Terraform Infrastructure-as-Code deployment
- User guides, operational instructions, and architecture reference.
- Knowledge transfer workshop.
- Results presentation and recommendations

5. Out of Scope

The following items are explicitly excluded from the scope of this project:

Requirement
Model Training and Retraining Training of new foundation models or extensive fine-tuning.
Privacy and Compliance The solution must comply with data privacy regulations (e.g., GDPR, CCPA) and ensure the secure handling of user data.
Cloud Foundations Set up Google Cloud with foundational best practices for administration.
Production Deployment Full-scale deployment and maintenance of the Monitoring solution in a production environment. Deployment will be limited to a pilot in a production environment, for which Customer will be required to manage and support to ensure full operational success.
Support and Maintenance Ongoing support, maintenance, and updates beyond the initial Proof of Concept (PoC) phase.
HITL Interfaces Custom development of sophisticated HITL interfaces.
Integration Integration with existing customer monitoring systems, non-GCP environments or cloud platforms.
Custom Development Custom development of complex evaluation metrics is not specified in the requirements.
BI Foundations Any setup, foundations, or integrations required for Looker or Looker Studio
Third-party tool integration Integration with a non-GCP tool is out of scope

6. Roles and responsibilities

6.1. Partner Roles and Responsibilities

Role	Role Description
Project Manager	• Serve as the primary point of contact for the Customer • Develop and maintain project plans, timelines, and status reports • Coordinate project activities and deliverables across all teams • Facilitate regular project status meetings with the Customer • Manage project risks, issues, and changes in scope • Ensure project deliverables meet the Customer requirements
ML Engineer	• Review provided dataset consisting of prompts and responses. • Design data schemas for evaluation metrics and implement BigQuery integration. • Configure visualization queries • Lead LLM integration and evaluation, and implement model serving components • Design evaluation methodologies • Configure LLM-as-judge functionality
Cloud Engineer	• API implementation • Supporting infrastructure • Build cloud infrastructure • Deploy GCP resources • Configure Vertex AI Pipeline components • Implement pub/sub and event triggers • Deploy GKE evaluation services • Set up the observability tools integration

6.2. Customer Roles and Responsibilities

Customer Role	Responsibilities
Project Sponsor	• Serve as the end decision-maker • Act as a point of escalation for issues • Sign off on deliverables • Provide business criteria for data exploration • Identify priority predictive outcomes
Project Manager	• Align Customer resources to support successful project completion • Help resolve any issues or questions in a timely manner
SME such as Product Manager or Technical Lead	• Collaborate with Bitstrapped to optimize Gen AI for the use case • Collaborate with Bitstrapped on vector database configuration

7. Dependencies/Assumptions

7.1. General assumptions

- The project timeline was based on information interpreted from related conversations with Customer to date. To the extent material discrepancies are uncovered, this SOW may need to be altered to accommodate timeline, resources, and/or cost changes.
- Customer will provide mutually agreed-upon personnel (IT, etc.) and resources (documentation, systems, etc.) on time.
- Customer Project sponsors will help resolve any issues or questions related to the work on time.
- The Customer team will assist with acquiring and implementing the Google Cloud environment and other tools required for development and deployment.
- Google Cloud usage fees are not included in the costs specified in the Financials Section. Customer shall be responsible for all Google Cloud usage fees.
- The new system has no compliance / regulatory / data location requirements for customer data.
- Partner does not warrant or otherwise represent the functionality of any vendor product used during this project. The respective vendor must provide such representations.
- Work related to this SOW is expected to be 100% remote. Travel to customer sites and on-site services are not required.
- This SOW is based upon the accuracy of the information the Customer provides.
- Any additions, changes, or modifications to the Services specified herein are outside the scope of this SOW.
- All Partner work will be performed remotely at Partner's facilities or resource locations.
- Customer shall make access for Partner resources available to the systems, applications, and knowledgeable personnel during designated time frames, which will be established during the project kick-off meeting. Failure to provide this timely access could delay the completion of the services.
- Customer already has Gen AI models serving inference and has the ability to add additional models as part of this evaluation framework deployment, and/or is using Bitstrapped to do this work as part of a separate SOW

7.2 Customer Responsibilities

- Align stakeholders to attend the kickoff meeting and regular weekly calls as part of the engagement; provide timely feedback on deliverables.
- Identify and ensure the availability of subject matter experts with the necessary skills and information regarding the technical infrastructure and the technical and functional application technologies.
- Customer shall be responsible for addressing any internal dependencies (including, but not limited to, business approval, network and security approval)
- Customer shall assign a named Project Sponsor who will be working with Partner.
- Identify data and content conducive to the knowledge base and model goals
- Provide knowledge sharing on datasets, including domain expertise and pre-processing requirements.
- Provide business and end-user requirements for Gen AI.
- Participate in training, feedback, and optimization exercises.
- Participate in regular project status meetings and engage in knowledge transfer sessions and

workshops to understand the end-to-end solution.

- Assume responsibility for the ongoing maintenance and management of the solution post-deployment.
- Deploy any Gen AI models to serve inference
- Maintain a prompt catalogue and versions
- Data preparation and cleaning
- Data integrations or third-party tool integrations
- Scaling out evaluation pipelines, observability dashboards, and other functional requirements beyond the limits of the Constraints section
- Integration with a human feedback source

8. Change Requests

Any change to this SOW shall not take effect unless and until a Change Request to this SOW is fully executed by Customer and Partner. The Change Request shall include the following information:

Date of Master Services Agreement (Optional)	
Date of Change Request	
Name and Date of Impacted SOW	
Description of modified Services	
Description of modified Deliverables	
Impact on resources	
Impact on Project Timeline	
Resulting change in Cost	
Effective Date of approval	

9. Project Timeline

The project timeline is based on an estimated start date of approximately two weeks after all parties sign this SOW. The start date is subject to change based on the availability of critical team members from the Customer and Partner.

The anticipated project duration is approximately **8 weeks**. This estimate assumes timely reviews and/or approvals from Customer.

Deliverable	**WK 01**	**WK 02**	**WK 03**	**WK 04**	**WK 05**	**WK 06**	**WK 07**	**WK 08**
Requirements Analysis and Customization								
Core Infrastructure Implementation								
Evaluation Pipeline Development								
Integration and Testing								
Pilot Release								
Documentation & Knowledge Transfer								

10. Financials

Partner will provide the Services on a fixed fee basis per the table below, including the total DAF amount as a line item.

Service	Cost
Requirements Analysis	$3,248
Core Infrastructure Implementation	$9,744
Evaluation Pipeline Development	$19,488
Integration and Testing	$9,744
Release and Deployment	$9,744
Documentation & Knowledge Transfer	$3,248
Project Management	$3,248
Subtotal	**$58,464**
Partner Services Funds (PSF)	**$58,464**
Total Due from Customer	**$0**

Google DAF Investment. Customer acknowledges and agrees that the Google DAF Investment is contingent upon meeting the Success Criteria shown in this document. The total of all Google payments shall not exceed the predetermined amount for the selected Consumption Pack. If Customer terminates this SOW or does not permit Partner to meet the Success Criteria for any reason, then Customer will be obligated to pay Partner under this SOW.

Estimated Project Expenses. Expenses, if any, would be billed to Customer as incurred with prior approval from Customer and per Customer policies. No travel or travel-related expenses shall apply to services under this SOW.

11. Signatures

The parties have caused this SOW to be executed by their duly authorized representatives as of the effective date set forth above.

[Customer Name]	**Bitstrapped Inc.**
Signature:	Signature:
Name:	Name:
Title:	Title:
Date:	Date: