GenAI Proof-of-Concept Process Guidebook



Global Technology Sales

Updated: Mar 28, 2025

# 

# Table of Contents

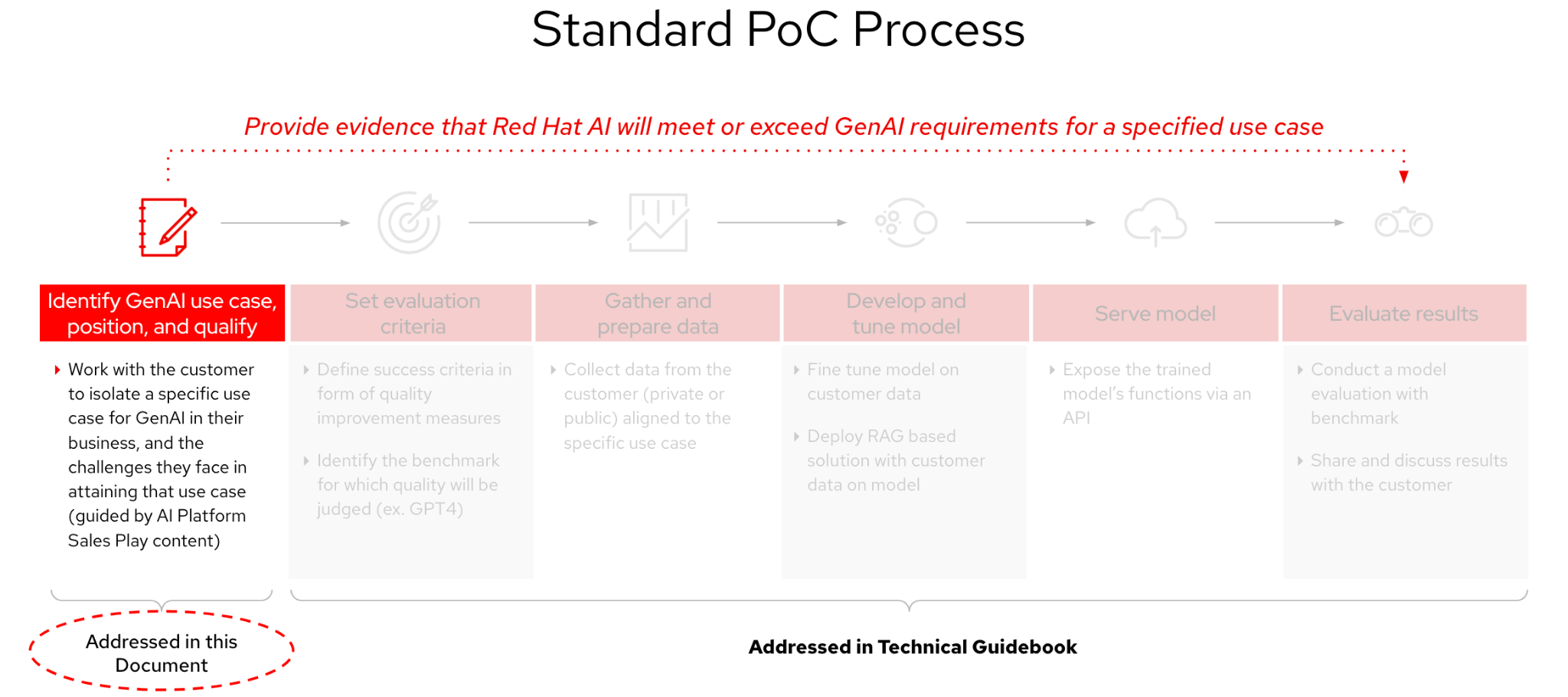
| [Purpose of this Document](#_n4zfkq3fn1ym) | 3 |
| --- | --- |
| [Internal PoC Engagement Process](#_qeqdpw349h1o) | 4 |
| [GenAI PoC Requirements](#_66nqzemb25n) | 7 |
| [Current Legal Guidance](#_ug2a851r4oka) | 9 |
| [Positioning a GenAI PoC](#_9wuppngeydux) | 11 |
| [Qualifying a GenAI PoC](#_p2fe04ltgpix) | 14 |
| [Documenting a GenAI PoC](#_20ik3o153wvv) | 17 |
| [Additional Resources](#_v4d7bq3w5vq) | 20 |

# 

# Purpose of this Document

AI is a natural extension of our hybrid cloud strategy and provides new opportunities to increase revenue and market presence. In order to capitalize on AI, Red Hat has introduced platform offerings focused on training, tuning, and inferencing foundation models. With many current and potential customers still exploring their own AI strategies, it is critical that Red Hat articulates and proves firsthand our AI Platform capabilities.

In 2025, Red Hat has set a company-wide goal to conduct **1,000** GenAI proof-of-concepts (PoCs) with customers. This document describes how to position, qualify, and document a GenAI PoC.



# 

# Internal PoC Engagement Process

Positioning, qualifying, and documenting a GenAI PoC requires collective efforts across the account team. Below is the standard internal engagement process for a GenAI PoC.

### **Required Steps**

**NOTE: You MUST complete Step 3b before creating a task in RHSC. See** [**GenAI PoC Engagement RACI**](#_ihle6rg06lxq) **to see which roles are accountable for the steps below.**

* **Step 1 - Position a GenAI PoC:** Identify an AI opportunity organically or via TeleSense lead tasks. Next, deliver an introductory AI pitch and confirm the customer’s interest.
* **Step 2 - Initial Qualification & Review w/Manager:** Use this step as an initial qualification (reference the [GenAI PoC Requirements](#_66nqzemb25n) and [Qualifying a GenAI PoC](#_p2fe04ltgpix) section below). *For all documentation, please provide comment access to both the AI Geo leader and the AI Global leader.*
  + 🚨Make sure you have reviewed the following with your direct manager before moving on to Step 3:
    - Use case
    - Success criteria
    - Documentation
* **Step 3a [TELCO Geo requests only]**
  + If you would like to request a GenAI PoC for a customer in the Telco geo, please use this [form](https://redhat.service-now.com/help?id=sc_cat_item&sys_id=988783b387a825508a51bbbf8bbb35fe). If their team needs additional support, they will fill out the form in Step 3b.
  + *Note: Telcos in Enterprise should continue to use the standard process outlined in Step 3b.*
* **Step 3b - Request Technical SSA Support to Conduct PoC:** Once you have done initial qualification and reviewed requirements for a GenAI PoC with your direct manager, it’s time to engage the AI Specialists in your geo. Please submit a request using this [Global Form](https://docs.google.com/forms/d/e/1FAIpQLSdhSIsvrwdT4dIDKiSqGtsdm6MKRiydq0BtCLUZ2umDCI6d4w/viewform). The form must be completed even if an AI SSA is involved with your customer. *Note: All requests are reviewed by the AI Platform Leadership team and* ***may be rejected*** *if the proposed GenAI PoC is not in scope or if it is deemed the customer is not qualified.*
* **Step 4 - AI SSA Assigned:** After the AI Specialists in your geo have reviewed your submission, they will respond confirming whether the GenAI PoC request has been accepted or rejected. *Their SLA is 3-5 business days.* If the submission has been accepted, then an AI SSA will be assigned to conduct the GenAI PoC. Each Geo has a pool of AI SSAs who have been trained to conduct GenAI PoCs..
* **Step 5a - Conduct the GenAI PoC:** Once an AI Specialist has confirmed that this is a qualified GenAI PoC opportunity, then the AI SSA conducts the GenAI PoC using the [Technical Guidebook](https://docs.google.com/document/d/1aOiwBPKgBEpvJy-02RVfE5rN6ZLHepuQU1-BOvLUSfQ/edit?tab=t.0#heading=h.66y4kqbj468a).
* **Step 5b - Document the GenAI PoC in RHSC:** Once the GenAI PoC is initiated, it’s time for the ASA to create a task in the Red Hat Sales Cloud (RHSC). The only way to get credit for a GenAI PoC is by documenting it in RHSC. For more information on the documentation process, please refer to the [*Documenting a GenAI PoC*](#_20ik3o153wvv)section of this document. Progress is monitored and updated using the task fields in RHSC.
* **Step 6 - Finalize GenAI PoC Documentation in RHSC:** Once the GenAI PoC is complete, all documentation is updated and the RHSC task is closed. Tasks will be [monitored](https://redhatcrm.lightning.force.com/analytics/dashboard/0FK6e000000iIKsGAM) for quality.
* **Step 7 - Conduct Handoff to Services:** After the GenAI PoC has been conducted and documentation completed, engage with the Services team in your geo to conduct customer handoff.

### GenAI PoC Engagement RACI

The internal engagement process includes several milestones. For more information on expected roles and responsibilities across these milestones, please refer to the matrix below.

|  | Accountable  *owns the outcome* | Responsible  *executes the task* | Supporting  *provides assistance* | Inform  *provides awareness* |
| --- | --- | --- | --- | --- |
| Position a  GenAI PoC | AE, ASA | | AI SSP |  |
| Qualify the GenAI PoC | AE | ASA, AI SSP | |
| Request Technical SSA Support to Conduct PoC | ASA | AI SSP |
| Conduct the GenAI PoC | AI SSA | ASA | TSM |
| Document the GenAI PoC in RHSC | ASA | AI SSA |
| Finalize GenAI PoC Documentation in RHSC | ASA | AI SSA |
| Conduct Handoff to Services | AE, ASA, TSM | | AI SSP |  |

# GenAI PoC Requirements

All GenAI PoCs conducted in accordance with the 1,000 company-wide goal must demonstrate Red Hat’s unique capabilities around **model training** **and/or** **inferencing**. Please see below for the acceptable categories of GenAI PoCs:

## PoC Categories

1. Model Training - includes an agreed upon and documented use case with the customer, tuning of a model using InstructLab on private or public customer data, deployment of a RAG based solution with private or public customer data on the model, and evaluation of performance using success criteria defined by the customer or Red Hat/Partner
   1. InstructLab + Granite on RHEL AI - Granite with InstructLab on RHEL AI
   2. InstructLab + Granite on RHOAI - Granite with InstructLab on RHOAI
   3. InstructLab + Llama/Mistral on RHEL AI - Llama or Mistral with InstructLab on RHEL AI depending on customer requirements, but must also have Granite in the PoC as part of price / performance analysis.
   4. InstructLab + Llama/Mistral on RHOAI - Llama or Mistral with InstructLab on RHOAI depending on customer requirements, but must also have Granite in the PoC as part of price / performance analysis.
   5. Customized InstructLab on RHEL AI - intended for either complex or Non English based data and/or use cases. Summarization use cases would require pre compiled data set manipulation. Non English Language would require the customer to seed the model fine tuning process with 10,000 examples in the local language.
   6. Customized InstructLab on RHOAI - intended for either complex or Non English based data and/or use cases. Summarization use cases would require pre compiled data set manipulation. Non English Language would require the customer to seed the model fine tuning process with 10,000 examples in the local language.
2. Model Inferencing - includes an agreed upon and documented use case with the customer, optimized inferencing of a model using vLLM, and evaluation of performance using success criteria defined by the customer or Red Hat/Partner
   1. vLLM on RHEL AI - vLLM to optimize inference of a customer provided model on RHEL AI on a customer’s infrastructure (on prem or cloud)
   2. vLLM on RHOAI - vLLM to optimize inference of a customer provided model on Red Hat OpenShift AI on a customer’s infrastructure (on prem or cloud)
   3. vLLM on Linux - vLLM to optimize inference of a customer provided model on a derivative linux operating system on a customer’s infrastructure (on prem or cloud)
   4. vLLM on Kubernetes - vLLM to optimize inference of a customer provided model running on a derivative kubernetes distribution on a customer’s infrastructure (on prem or cloud)

# 

# 

# Current Legal Guidelines

As a commercial entity, we should be mindful of potential IP and privacy concerns from our customers when Red Hat uses their data, even if publicly available, without the customer's prior awareness. Various jurisdictions have rules on scraping data from the internet and customers may have their own policies and website terms on scraping data from their public webpages. Some customers may be wary of AI or the sources of data for the PoC, and we want our customers to focus on the PoC and benefit of the tools. This guidance is meant to assist you if these issues are raised by the customer.

As you consider a GenAI PoC with your customer, please ensure the following:

* **For customer demos using publicly available customer data:**
  + Where feasible, we recommend providing the customer with [this email](https://docs.google.com/document/d/1jSrromrXSMWrGqhePOqJgwEMMOb_3Rxv/edit) in advance, informing the customer of our plan to temporarily use their publicly available data for the PoC.
  + Where infeasible to send the email due to timing of the demo or other factors, the email does not have to be provided as long as *only public information is used and no internal data about the customer is incorporated into the demo*.
* **For customer demos using customer’s confidential information provided to Red Hat by the customer:**
  + Provide the customer with [this email (same as above)](https://docs.google.com/document/d/1jSrromrXSMWrGqhePOqJgwEMMOb_3Rxv/edit) in advance of the demo informing the customer of our plan to temporarily use a combination of their publicly available data and other data they may provide to Red Hat for the demo. The email clarifies our obligations to keep the data confidential.
* **Obtaining signed agreement where requested by a customer:**
  + Where the customer data is public or provided by the customer for purposes of the demo, a separate agreement/legal terms are **not** required by Red Hat.
  + However, if the customer insists on PoC legal terms, please notify your Geo AI Platform Leader who will work with Red Hat Legal to document a customer's acceptance and consent for the data use.
* **Other recommended practices when working with customer data for the PoCs:**
  + **Customer data storage:** Do not save customer data, even publicly available data, in **public** repositories. Some customers have tools scanning the internet for their data and we don’t want to create any customer issues.
  + **Sharing demos:** Do not share a demo (and the underlying model/output etc.) created specifically for one customer with their data with another customer or a third party (i.e., don’t use data from customer A’s website and give a demo to customer B).
  + **Delete the customer data and models after the demo, or within a reasonable time after the demo.**

# 

# 

# Positioning a GenAI PoC

### Take the First Step

Positioning a GenAI PoC begins with understanding the core components of Red Hat’s GenAI capabilities. Luckily, we’ve summarized the key points for you.

### Brief Introduction to InstructLab

InstructLab is an open source project, and its mission is to democratize AI and make enhancing an LLM more accessible to the developers and analysts (non-ml engineers/phds). Typically, when someone wants to add new data to a model, it is a very complicated, time consuming, and costly process. A significant amount of time and resources go into collecting and preparing data for model fine-tuning, and setting up a model alignment pipeline, all of which requires data scientist expertise. InstructLab takes a different approach.

InstructLab allows the user to create a set of yaml and markdown files using plain English text and the project will take care of the heavy lifting. This makes adding new knowledge into a model more accessible in contrast to other methods.

### Brief Introduction to Granite

IBM Granite foundation models are cost-efficient, enterprise grade large language models. The Granite models distinguish themselves by being tailored for enterprise use cases, with features that emphasize data privacy, security, and regulatory compliance—critical aspects for industries such as healthcare, finance, and government.

##### Granite Model Offerings

* *Dense, general purpose LLMs*
* *LLM-based input-output guardrail models*
* *Mixture of experts (MoE) models for minimum latency*
* *Speculative decoder for increased inference speed and efficiency*
* *Transparency & Indemnification*

#### The Power of Small Language Models (SLMs)

#### Much of the hype surrounding Generative AI technology involves pushing the size boundaries of Language Models. They are commonly referred to as Large Language Models, after all. However, bigger may not always be better [for every use case](https://www.youtube.com/watch?v=C4Qt9Hnp6vs). Small language models (SLMs) offer several advantages over large models, particularly in scenarios where efficiency, cost, and simplicity are paramount. One of their most significant benefits is their lower computational requirements, making them more accessible for deployment on edge devices, personal computers, or smaller cloud environments. Due to their smaller memory and computational footprint, SLMs are easier to deploy and manage. SLMs provide another option for customers to leverage generative AI.

#### Indemnification

#### The Granite models included in the Red Hat products are indemnified by Red Hat. Indemnification refers to legal protection provided by IBM to its clients, ensuring that they are shielded from potential legal liabilities arising from the use of AI models, such as issues around intellectual property, copyright infringement, or bias in AI decision-making. This feature helps alleviate any concern in regards to legal risks, as compared to other models which are embroiled in litigation. This aspect of IBM’s Granite models is particularly important for enterprises that operate in highly regulated industries or those dealing with sensitive data, where compliance and risk mitigation are crucial.

#### vLLM Inference

#### vLLM is the leading community-developed open-source LLM inference server started by UC Berkeley in June of 2023. vLLM focuses on improving the execution speed and resource utilization of LLMs when working with variable-length inputs. This can be particularly important in scenarios where models must handle a wide variety of text lengths, ranging from short prompts to long documents. In January of 2025, Red Hat acquired Neural Magic, a leading contributor to vLLM.

##### Primary Benefits

* *Enhanced efficiency*
* *Faster inference and lower latency*
* *Reduced memory footprint*
* *Scalability*
* *Cost Savings*

#### Next Steps

#### Visit the [AI Platform Sales Play](https://content.redhat.com/us/en/sales-enablement/sales-plays-overview/sales-play-red-hat-ai.html#tabs-c3df71e1be-item-0ab3b46204-tab) on the Red Hat Content Center for customer-facing presentations, business value narratives, discovery decks, and more.

##### Core Sales Play Assets

* [*Customer-Facing Deck*](https://content.redhat.com/us/en/assets/display.html?id=e18ae7ba-7565-46ab-b8a1-5960cdae3c62)
* [*Customer-Facing Pitch*](https://videos.learning.redhat.com/media/AI+Platform+-+Sales+Play+Customer+Facing+Pitch+-+Jeff+Winn/1_ffpr1p7d/347961792)
* [*Sales Play Intro Deck*](https://content.redhat.com/us/en/assets/display.html?id=ed6e3ecd-77ba-4ab7-b3ab-1c9397d95be1)
* [*Training an LLM with InstructLab Interactive*](https://interact.redhat.com/share/eTE6P4AjFyVIeRxNZX1C)

#### 

#### On your next customer call, use the last 5 minutes to give a quick pitch for Red Hat AI. Following the call, set up a formal AI discussion with the customer.

# 

# Qualifying a GenAI PoC

As you approach a PoC opportunity, it is crucial to evaluate and accurately determine whether or not this opportunity qualifies as a good GenAI candidate. It might be tempting to jump to offering our GenAI tools such as InstructLab, but the InstructLab alignment method is not applicable to every customer use case. Please review below to understand whether or not your PoC should leverage InstructLab.

#### Use Case Examples

##### InstructLab (Based on Real Examples)

* A real estate company wants to streamline how their teams address real estate-related inquiries.
  + They want an LLM-powered chatbot.
  + The model must understand how the layers of documents from different time periods interact to answer queries correctly (some land deals occurred in the 1800s, for example, and laws/leases/agreements have changed over time).
  + InstructLab: Train a model on real estate data to achieve above needs.
* A banking company wants to help customer service representatives handle credit card disputes more easily.
  + They would like to create an LLM-powered chatbot to help show what paperwork needs to be filed, how long they should expect to wait before hearing back, etc.
  + InstructLab: train model on bank-specific data.
* A city government wants to improve and streamline how it serves its constituents.
  + They want to:
    - Generate summaries of legislation
    - Create knowledge agents to assist with eligibility questions for public health services, assist junior engineers in the department of transportation, and support public transportation programs
  + InstructLab: fine-tune a model on city data to serve above use cases.
* A large global telecommunications company receives over 300k customer calls daily. Call transcripts need to be quickly analyzed for business & client impact to roll back into client-facing processes to improve automation and customer satisfaction.
  + Desire for quality improvement, lower cost, quick results, data privacy built-in.
  + InstructLab: align their own custom LLMs with their private corporate data.
* A banking company needs to generate 100k proposed answers to client emails per day to help client advisors reply to clients’ emails in a more personalized tone and in the context of the ongoing conversation.
  + InstructLab: fine-tune a small language model on internal datasets on the task.

##### InstructLab NO or NOT YET

* I want to fine-tune my model for specific skills.
  + Not yet.
  + Reason: Skills are currently out of scope for PoCs until further guidance is given.
* I want to fine-tune my model on my codebase for custom coding performance
  + Not yet (unclear roadmap)
  + Reason: we currently do not support code customization with InstructLab
* I want to fine-tune a model for object detection
  + No
  + Reason: InstructLab does not support vision-related AI generation
* I want to use the Granite models as they are in my application
  + No.
  + It’s great, but it doesn’t require InstructLab so it’s not an InstructLab PoC.

#### Use Case Best Practices

Selecting the appropriate use case is often the cornerstone of a successful PoC. A well-chosen use case should meet the following criteria:

1. **High Impact**: The use case should address a critical need or deliver measurable value to the customer. This ensures that the results resonate with stakeholders and showcase the potential benefits of scaling the solution.
2. **Low Risk**: Start with scenarios that are feasible within the available timeframe, budget, and technological constraints. This minimizes the likelihood of unexpected roadblocks.

## Easily Measurable Outcomes: The success of the use case should be quantifiable, with clear metrics that can demonstrate value quickly and convincingly.

To identify an optimal use case for the PoC, we recommend considering the following:

1. Understand the Customer’s Business Goals
   * Engage in discovery sessions to map out the customer’s objectives and challenges.
   * Ask targeted questions about pain points, operational inefficiencies, and areas for improvement.
2. Analyze Feasibility
   * Assess the technical and operational readiness of the customer’s environment.
   * Determine if your solution aligns well with the customer’s existing systems and workflows.
3. Select High-Impact Scenarios
   * Prioritize use cases that could lead to significant operational, financial, or strategic benefits.
   * For example, automating repetitive tasks, reducing error rates, or enhancing customer satisfaction are often impactful and achievable.
4. Document the Chosen Use Case
   * Clearly articulate the scope, objectives, and expected outcomes of the PoC.
   * Include detailed descriptions of the problem being addressed, the proposed solution, and the implementation steps.

## 

## Ready to get started? Visit the [AI Use Case Playbook](https://content.redhat.com/us/en/sales-enablement/sales-plays-overview/sales-play-red-hat-ai.html#tabs-c3df71e1be-item-71b9a34bd6-tab) on Red Hat Content Center for repeatable examples of use cases that you can use with your customer!

## 

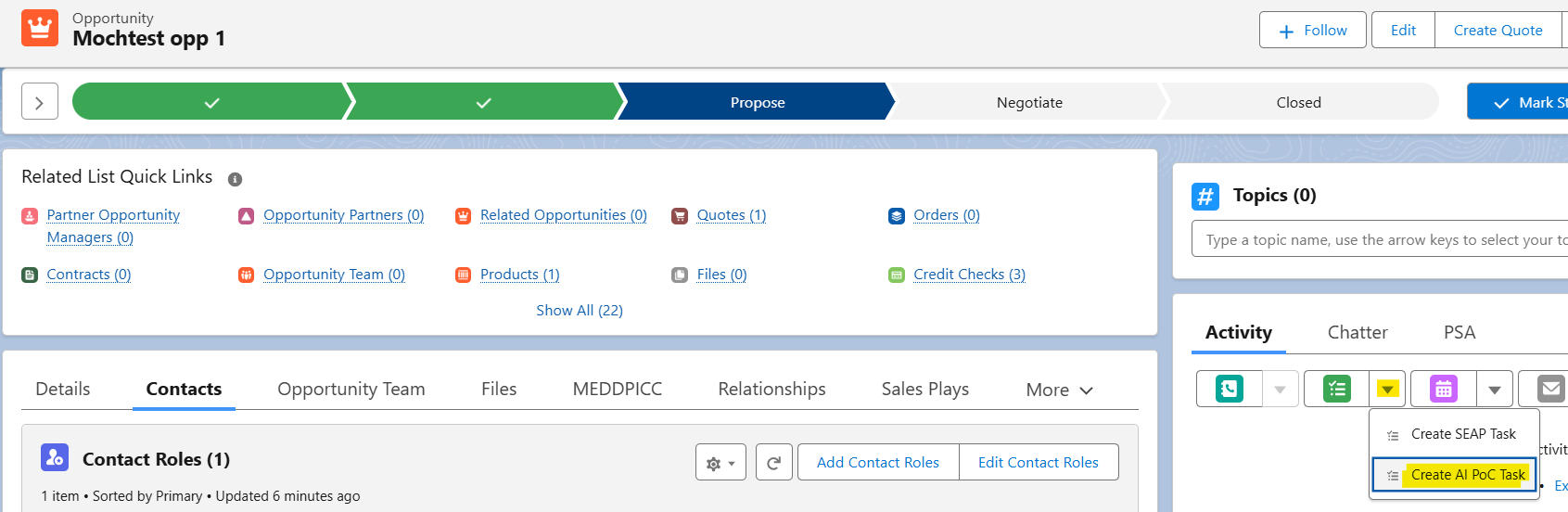
## 

## Documenting a GenAI PoC

In order to earn credit for conducting a GenAI PoC, the activity must be reported in Red Hat Sales Cloud using the instructions below. We expect the Account Solution Architect (ASA) to be responsible for the documentation.

## Step 1: Create a Task in Red Hat Sales Cloud on your Opportunity

A Task is a business activity that is logged and tracked in Red Hat Sales Cloud. To begin reporting a GenAI PoC, create a Task in Red Hat Sales Cloud on a valid AI Opportunity under the “Activity” bar as shown below.



***NOTE:*** *All PoC Tasks* ***must*** *be related to an opportunity. If you do not have an AI Opportunity for your customer already, please create one and assign the product value as $0. Any Tasks not associated with an AI opportunity will not be logged or tracked.*

## Step 2: Fill in the Required Fields on the Task

Fill in the required fields on the Task using the guidance in the table on the following page. All fields denoted in this table are **required** in order to save the Task. If you do not have detail for every field while the PoC is in progress, denote that in the field itself but update it with final information as the PoC is finalized/completed. Please note that a scoping document, an evaluation of results, and a customer presentation is expected documentation.

| Field Name | Description | Example Response |
| --- | --- | --- |
| Subject | subject line | *AI PoC with USAA* |
| Related To | link to valid RHSC AI Opportunity | *USAA - New RHEL AI* |
| Priority | importance or urgency. Selection includes:   1. Low 2. Medium 3. High | *High* |
| Status | current status. Selection includes:   1. Open *- identified/discovered but not validated* 2. In Progress *- validated with resources assigned* 3. Blocked *- issue or dependency preventing progress* 4. Completed *- PoC has been completed* | *In Progress* |
| Assigned To | Account Solution Architect | *John Doe* |
| PoC Technical Owner | name of the Red Hat technical resource who will be delivering the PoC. This will likely be an AI SSA | *Janet Doe* |
| PoC Partner Role | role that a partner is playing in the PoC process. Selections include:   1. Partner Sourced and Delivered PoC- *partner sourced and delivered the PoC on behalf of Red Hat* 2. Partner Sourced but Red Hat Delivered PoC- *partner sourced the PoC, but Red Hat (i.e. a RH technical resource) delivered the PoC* 3. Red Hat Sourced & Delivered but Partner Supported - *Red Hat sourced and delivered the PoC, but partner supported sourcing/delivery with hardware, software, expertise, etc.* 4. No Partner Involved - *Red Hat sourced and delivered the PoC and no partner was involved* | *Red Hat Sourced & Delivered but Partner Supported* |
| PoC Partner Name(s) | name of the Partner(s) whose role is described in the previous question. If not applicable, put N/A. Separate multiple partners with a comma. | *Dell Technologies* |
| PoC Start Date | date that the PoC is expected to start/actually starts | *4/10/2025* |
| PoC End Date | date that the PoC is expected to end/actually ends | *4/15/2025* |
| PoC Category | category that best explains the focus of the PoC. Selection includes (see “GenAI PoC Requirements” for definitions):   1. InstructLab + Granite on RHEL AI 2. InstructLab + Granite on RHOAI 3. InstructLab + Llama/Mistral on RHEL AI 4. InstructLab + Llama/Mistral on RHOAI 5. Customized InstructLab on RHEL AI 6. Customized InstructLab on RHOAI 7. vLLM on RHEL AI 8. vLLM on RHOAI 9. vLLM on Linux 10. vLLM on Kubernetes 11. Other *- must explain in “Use Case & Success Criteria"* | *InstructLab + Granite on*  *RHEL AI* |
| PoC Environment | type of environment where the PoC will be primarily developed and/or deployed. Selection includes:   1. IBM InstructLab as a Service 2. Cloud (not IBM) 3. Hybrid 4. On-Premise | *IBM InstructLab as a Service* |
| PoC Use Case and Success Criteria | specific AI use case that is being demonstrated through the PoC and the benchmark for which Red Hat (and/or Partner) and customer have agreed to evaluate the success of the PoC | *Our customer wants to streamline how their teams address insurance-related inquiries using an LLM-powered chatbot. We will use InstructLab to train a model on their specific claim data. Our model must perform at 105% of GPT4 at 30% cost.* |
| PoC Artifacts | enter and label any links to PoC documentation including but not limited too: scope documents shared with customer, evaluation analysis, and artifacts (ex. git repo with the taxonomy used during the PoC). In addition, please enter the custom model ID from IBM InstructLab as a Service if the service was used. | *Scope Doc: <link>*  *Evaluation Analysis: <link>*  *Customer Presentation: <link> InstructLab Repo: <link>*  *IBM Saas ID: ae9d47ae-b1d5-4b 5a-93d2-a21dcd84bbe* |
| Comments | running list of updates and action items | *(4/11) Update: Customer has provided public data for the PoC. We are creating the qna.yaml file for the synthetic data generation process of the InstructLab workflow.* |

## 

## Step 3: Update Task Fields

As the PoC is conducted, update the Task with new information and/or changes in Status. Once the PoC is officially concluded, finalize all field inputs and then mark the Task “Completed”.

## 

# 

# Additional Resources

Have a question about the material in this Guidebook? Reach out to the AI Platform leader in your Geo for guidance or visit the [AI Platform Source Page](https://source.redhat.com/departments/sales/global_gtm_sales_specialists_team/global_ai_platform).

| Geo | Contact |
| --- | --- |
| Global | Jeff Winn  jwinn@redhat.com |
| Americas | Monica Livingston  mlivings@redhat.com |
| EMEA | Johan Robinson  jorobins@redhat.com |
| APAC | Steve Shirkey  sshirkey@redhat.com |

# 