Benchmarking Amazon Bedrock Models on τ-bench

Project Overview

Benchmarking AI models in real-world **Tool-Agent-User (TAU) interactions** is crucial for evaluating their effectiveness. However, τ-bench, a popular benchmarking framework, **does not currently support Amazon Bedrock**. This project aims to extend τ-bench to support Bedrock and evaluate the **Amazon Nova and Meta Llama 3 models**.

The goal is to build a **τ-bench extension** that:

1. **Integrates Amazon Bedrock APIs** into the τ-bench framework.
2. **Benchmarks Amazon Nova and Meta Llama 3 models** using τ-bench tasks.
3. **Submits benchmarking results** to the [Holistic Agent Leaderboard](https://hal.cs.princeton.edu/).
4. **Analyzes model performance** based on accuracy, latency, and usability metrics.
5. **Ensures Responsible AI considerations** like bias and fairness evaluations.

By leveraging **τ-bench and Amazon Bedrock**, this project will provide valuable insights into **how Bedrock models perform in real-world AI agent tasks**.

Why This Benchmark Matters?

To effectively compare AI models, τ-bench evaluates:

- **Tool usage efficiency** (how well the model interacts with external tools).
- **Agent autonomy** (how independently the model completes tasks).
- **User interaction quality** (how well the model responds to human inputs).

By extending τ-bench for **Amazon Bedrock**, we can **assess how Bedrock-hosted models perform relative to other major AI models**.

Technical Components & Implementation

1. Extending τ-bench for Amazon Bedrock
- Modify **τ-bench's codebase** to support API calls to Amazon Bedrock.
- Implement **request/response handling** for Bedrock models.
- Ensure compatibility with τ-bench's evaluation framework.

2. Model Benchmarking & Evaluation

- Run **Amazon Nova and Meta Llama 3 models** on τ-bench tasks.
- Collect and compare performance metrics such as:
  - **Accuracy** (task completion rate).
  - **Latency** (response time).
  - **Usability** (how well models follow instructions).
- Submit results to the **Holistic Agent Leaderboard**.

3. Retrieval-Augmented Generation (RAG) Integration (Mandatory)
- Implement **retrieval-augmented generation** for query optimization.
- Use **embedding models** to enhance retrieval accuracy.
- Optimize query-to-context mapping for better response quality.

4. Agent-Based AI Evaluation (Mandatory)
- Modify **τ-bench's agent workflows** to work with Amazon Bedrock.
- Adapt agent interactions for **tool use, decision-making, and task execution**.
- Compare how different models handle agent-based scenarios.

5. Responsible AI & Bias Evaluation
- Conduct **fairness and bias testing** on Amazon Bedrock models.
- Document ethical considerations and **potential risks** in model deployment.
- Analyze if **certain user inputs lead to biased or unsafe responses**.

Evaluation & Success Metrics

The success of this project will be evaluated based on:

- **τ-bench integration success** (Bedrock models running on τ-bench).
- **Benchmarking results submission** (submission to Holistic Agent Leaderboard).
- **Model performance analysis** (accuracy, latency, usability comparison).
- **Responsible AI assessment** (bias, fairness, ethical considerations).

Why This Project Matters?

This project provides hands-on experience with:

✓ **Benchmarking AI models in real-world tasks**
✓ **Using τ-bench for evaluating generative AI agents**
✓ **Integrating Amazon Bedrock APIs into AI frameworks**
✓ **Analyzing Responsible AI considerations**