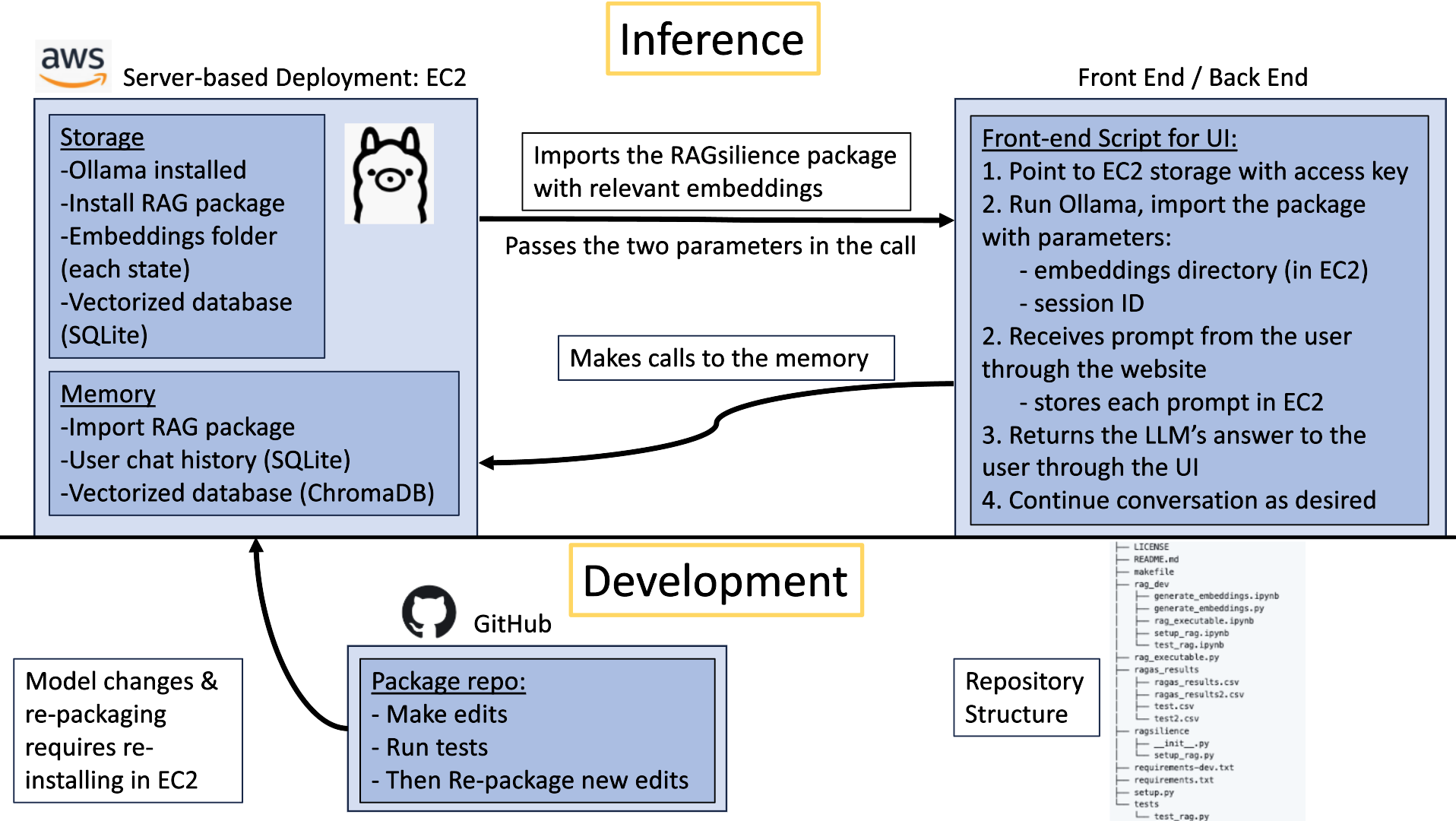
Recommendation for Server-based Deployment – AWS EC2

1. Overview and advantages of AWS EC2
   1. EC2 is scalable and easy to initiate – other provides such as Google Cloud require paying more upfront for pre-allocation
   2. Can increase memory or number of CPUs as demand increases
2. Recommended Specifications and Cost
   1. Starting out
      1. “EC2 Instance Savings Plan” is cheaper than “EC2 On Demand”
      2. EC2 t2.xlarge instance (highlighted below). Any smaller will likely be too little memory – Ollama is 8GB. Can scale to t2.2xlarge if needed.
      3. Alternatives: <https://www.cloudzero.com/advisor/t2-vs-t3/>
   2. Approximate yearly cost: 1-year contract, partial-upfront
      1. Max - $1,103 (billed based on each hour of use in the 1-year contract)
      2. <https://aws.amazon.com/savingsplans/compute-pricing/>

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Instance Name | Operating System | Savings Plan Rate  ($ per hour of use) | Number of Cores | vCPUs | Max Storage | Memory |
| t2.xlarge | Linux | $0.126 | 2 | 4 | 2 TB | 16 GiB |
| t2.2xlarge | Linux | $0.2518 | 2 | 8 | 2 TB | 32 GiB |

1. Deployment Architecture



1. Resilience Education will need to determine:
   1. Length of contract, upfront payment (none, partial, all), EC2 operating system
   2. Training embeddings for other states
      1. Attempt with CPU “t2.xlarge” instance – took us 21 minutes to train
      2. Otherwise, try with a GPU “G4ad.xlarge” instance is $0.37 - $.54 per hour
         1. <https://aws.amazon.com/ec2/instance-types/g4/>
      3. Alternative: Google Colab offers $10 per month for limited GPU usage