Contents

1. Configuration

Deployment options which are specific to the environment are configured with environment variables. The primary app configuration which is agnostic to the environment is done through `app-config.yaml` in the root of the repository, and includes the questions asked and the system prompt templates for example.

1.1. Global

Environment Variable	Default Value	Description
`MAX_INPUT_TOKENS`	`2000`	Instructs the inference provider to not allow more than `MAX_INPUT_TOKENS` number of input tokens. Useful to bound the cost of inference and malicious requests
`MAX_OUTPUT_TOKENS`	`800`	Instructs the inference provider to not output more than `MAX_OUTPUT_TOKENS` number of output tokens. Useful to bound the cost of inference

1.2. Inference

There are multiple backend inference providers which require different configuration.

1.2.1. Ollama

Ollama is an open source, self hosted LLM inference runner.

Environment Variable	Example	Description
`INFERENCE_PROVIDER`	`ollama`	Must be `ollama`
`OLLAMA_HOST`	`http://ollama:11434`	URL to a running Ollama instance reachable by the container
`OLLAMA_MODEL_ID`	`gemma3:4b`	Model ID. The list of model IDs are available at `ollama.com/search`

1.2.2. AWS

Amazon Bedrock is a managed LLM inference runner. It provides very large and fast models, and is billed per token.

Environment Variable	Example	Description
`INFERENCE_PROVIDER`	`aws`	Must be `aws`
`AWS_REGION`	`us-east-2`	Region which the Bedrock model is available in. See `AWS_BEDROCK_MODEL_ID` for more information
`AWS_ACCESS_KEY_ID`	`AKIAXXXXXXXXXXXXXXXX`	Access key to a user which can assume `AWS_BEDROCK_ROLE_ARN` role. Not required if the container is running on AWS with proper IAM assume role principals
`AWS_SECRET_ACCESS_KEY`	`XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX`	Secret access key to a user which can assume `AWS_BEDROCK_ROLE_ARN` role. Not required if the container is running on AWS with proper IAM assume role principals
`AWS_BEDROCK_ROLE_ARN`	`arn:aws:iam::XXXXXXXXXXXX:role/YYY`	AWS role ARN which has the inline policy JSON below
`AWS_BEDROCK_MODEL_ID`	`meta.llama3-3-70b-instruct-v1:0`	Must be a Model ID from the list available at `docs.aws.amazon.com/bedrock/latest/userguide/models-supported.html`. Note that not all models are available in all regions

1.2.2.1. IAM Inline Policy

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Sid": "BedrockAPIs",
            "Effect": "Allow",
            "Action": [
                "bedrock:InvokeModel"
            ],
            "Resource": "*"
        }
    ]
}
```

1.2.3. Mock

When no other inference provider is specified, it defaults to the mock inference provider. Alternatively, it can be forced by setting `INFERENCE_PROVIDER=mock`. The mock inference provider is available in all container images. It simply echoes back the input as the output, and does not use an LLM.

2. Deployment

Continuous integration and continuous delivery build new containers weekly on Tuesdays at 4am UTC with the latest updates from all dependencies. A dashboard of the artifacts is available at `github.com/capstone-au2025/project/pkgs/container/project`.

Certain values of the `INFERENCE_PROVIDER` environment variable only work when using the correct container image. The available images are provided below:

Container Image	Available Environment Variable Value
`ghcr.io/capstone-au2025/project:aws`	`INFERENCE_PROVIDER=aws`
`ghcr.io/capstone-au2025/project:ollama`	`INFERENCE_PROVIDER=ollama`
`ghcr.io/capstone-au2025/project:openai`	`INFERENCE_PROVIDER=openai`

2.1. Recipes

2.1.1. Local

A production ready docker compose file is available at `infra/docker-compose/docker-compose.yaml` in the repository which can be copied. It is also reproduced below for reference.

This configuration uses the `gemma3` model from Google at 4 billion parameters, which strikes a balance of usable outputs while still being runnable on a few CPUs. It also uses `watchtower` to automatically pull updates to the containers as they come out to ensure the latest security updates are always applied. The server is available on port 3001 which should be configured by a reverse proxy of choice to expose over HTTPS to the internet.

```yaml
services:
  backend:
    image: ghcr.io/capstone-au2025/project:ollama
    ports:
      - 3001:3001
    environment:
      - INFERENCE_PROVIDER=ollama
      - OLLAMA_HOST=http://ollama:11434
      - OLLAMA_MODEL_ID=gemma3:4b
    depends_on:
      - ollama
  ollama:
    image: ollama/ollama
    volumes:
      - ./ollama:/root/.ollama
  watchtower:
    image: beatkind/watchtower
    volumes:
      - /var/run/docker.sock:/var/run/docker.sock
```

2.1.2. NRP Kubernetes Cluster

Kubernetes manifests are available in the repository to run on the National Research Platform. To create the resources, simply run the following command:

```
kubectl -n landlord-letters apply -f infra/nrp/manifest.yaml
```

The website will then be available at letter-generator.nrp-nautilus.io.

2.1.3. AWS

todo