HW3

Lukas White

September 2025

1 Summary of Work This Week

This week, we worked through setting up a local LLM agent that could run completely free, without relying on API quotas. We started by examining a mock agent that returned hardcoded responses and then explored how to connect a real LLM. Initially, attempts to use Hugging Face hosted models like Gemma-2B failed due to 401 and 404 errors, indicating either invalid credentials or models that were not available on the hosted inference API. We then moved to running models locally using the Transformers library, which allowed us to bypass all API restrictions and run inference entirely on the local machine.

We first tried `sshleifer/tiny-gpt2`, which worked technically but produced repetitive and nonsensical output due to its extremely small size. To improve coherence, we switched to `distilgpt2` and adjusted the `LocalRunner` so that the model and tokenizer were loaded only once, with sampling parameters set to reduce repetition. Finally, we modified the script to write the AI output to a text file for easier reading. While `distilgpt2` still did not follow instructions perfectly, the setup provided a working framework for running a local LLM that integrates with the existing `Agent` and `Runner` system. We concluded that for truly instruction-following output, a small instruction-tuned model like `flan-t5-small` would be a better choice, and ultimately upgraded to `flan-t5-large` to generate more coherent and instruction-compliant responses.

The set-up for running this on local-machine is a lot easier than in the previous version (albeit the ai is worse)

2 Relevant Papers

2.1 Paper 1: Data Source and Purpose

Here, describe the papers you are reading, the source of their data, and why it is relevant to your work on local LLMs. Explain the purpose of the datasets or models they discuss.

2.2 Paper 2: Documentation of Process

Document how the papers approach their experiments, training methods, or evaluation procedures. Explain any processes or techniques that are relevant to your own LLM setup.

2.3 Paper 3: Usefulness of Data

Describe how the information or datasets in the papers are useful for your project. Include why the data or insights are valuable in the context of building or testing local AI agents.

3 Examples of AI Output

Below are some example prompts and outputs generated using our local LLM agent:

- **Prompt:** Write a haiku about recursion in programming.
  Output: [Insert AI-generated haiku here]

- **Prompt:** In 5 words, say I love you.
  Output: [Insert AI-generated 5-word output here]

- **Prompt:** Name 5 countries where they speak Spanish.
  Output: [Insert AI-generated list here]