Guidance & Best Practices for Creating OwlONE AI Agents

Introduction

Instruction sets, or system prompts, are key to designing, controlling and refining a specialized agent's behavior in OwlONE. Every agent has a "context window " associated with its underlying AI model which is the number of tokens (roughly similar to the number of words) the model can interpret and respond with in a single query. The instruction set, retrieved information from knowledge sources, user prompt, and generated response all fall within this overall context window.

Most use cases will never reach the limitation, however, **it is important to understand that there is an information heirarchy such that the instruction set is given more priority than the user prompt or other information when the model interprets the context it receives**. This functionality allows us to put guardrails in the instruction set that hinder or prevent users from using the agent inappropriately.

Instruction Set Creation & Refinement

- It is highly recommended that system prompts (instruction sets) are created in a markdown file and saved locally or in onedrive and then pasted into the appropriate textbox when creating an agent.
- A prompt template is available in the /owlone-prompt-instructions/Prompts subdirectory which can be utilized to have Copilot or similar generate an draft instruction set for agents.
- "MUST HAVE SECTIONS .md" has been created based on information learned from Cloudforce as well as info independently determined by OIT testing.
    - For example: "Today's date is {{today}}." at the end of every instruction set is essential for the agent to provide temporally aware responses (such as the start date for fall semester)
- Provide significant ethics, rules, and guardrails to maintain appropriate usage and deter off topic or inappropriate discussions for an educational setting.

Proposed Naming Conventions

Field	Format String	Example
Agent name	[agent name] + " Agent "	owLGEBRA: FAU Algebra Tutoring Agent
File library	[agent name] + " Library"	owLGEBRA: FAU Algebra Tutoring Library

- These are recommended naming approaches for ideal organization and search functionality in the OwlONE user interface.

Parameters

- Temperature reshapes the probability distribution.

- Top P truncates the distribution to a subset of likely tokens.
- For more consistent, grounded, and less creative agents, seek temperature values of 0.1-0.3 with Top P values of 0.2-0.5.

Temperature

- Controls the shape of the probability distribution over the next token.
- Lower values (e.g., 0.2â€“0.5): Make the model more deterministic, favoring high-probability tokens.
- Higher values (e.g., 0.8â€“1.0): Make the model more creative, flattening the distribution and allowing more diverse outputs.

Top P

- Controls the scope of the token pool by selecting the smallest set of tokens whose cumulative probability exceeds the threshold P.
- For example, Top P = 0.9 means the model will only consider the top 90% of the probability mass and ignore the rest.
- Within that pool, the final token is sampled proportionally to its probability

Reduce hallucinations = OFF

- This feature can handicap the agent. This feature may limit the agentâ€™s ability to respond to even simple prompts like â€˜hello.â€™
- Suggested for highly specialized usecases where responses are directly from the provided info/instructions (more testing is required here)

Model

- Select "Assistants" under models list for highly powerful Assistants API equipped variants of the OpenAI models.
- Main difference between "Assistants" and "OpenAI" models is context window. The prior can support roughly 2 million tokens.
- GPT-4.1 for STEM (roughly 1 million token context window)
- GPT-4o-mini for simpler agents

FAQ

1. I see references to OpenAI and ChatGPT in OwlONE. Are my information and data linked to external services like ChatGPT? No, OpenAI and Microsoft have agreements in place such that the base AI models (like "gpt-4.1") are distributed to Microsoft. Microsoft is a vendor of numerous FAU services, including the Azure cloud services that allow us to run OwlONE and many other platforms securely and privately.
2. What are the file upload limitations? The "chat with file" (user's ability to upload a file to the agent) is limited to 50pgs cumulatively in a single chat. The File Library is not limited, though agent creators should be aware that an agent will perform better with smaller File Libraries (in terms of bytes, not necessarily quantity of files) due to its context window and information retrieval methodology.