Assignment A01 – GPT-5

Contributors: Martin Demel, Jiri Musil, Jasmine Campos, Munachimso Benson

Department of Science, Technology, Engineering & Math, Houston Community College

ITAI - 1371 Introduction to Machine Learning

Professor Vishwa Rao

August 24th, 2025.

# Introduction

Our team researched the current state of large language models and focused on three questions. We aimed to understand why GPT‑5 is important, which GPT-4 features we can use, and how GPT-5 compares to Bard, now called Gemini.

# Body

GPT‑5 marks a major improvement by enhancing reliability in complex reasoning and offering us better control over its thinking process. OpenAI reports results such as 94.6% on AIME 2025 without tools, 74.9% on the SWE bench, verified for real‑world coding tasks, and 84.2% on the multimodal MMMU benchmark. The API also provides practical controls like a reasoning effort setting and a verbosity setting, allowing us to decide when to answer quickly or think more deeply and how long the response should be. For scale, GPT‑5 supports a context window of about 400,000 tokens and up to 128,000 tokens of output, enough for long prompts, multi-step plans, and agent style tool use. These updates make it more capable of completing real work with fewer errors.

For GPT‑4, we discussed simple use cases that we can implement now. One example we discussed is the fitness application that would utilize GPT-4 summarization for health data. It would then provide tailored insights and summarize micro activities that the user should do. This would also be ideal for notifications. Another use case we identified is the RFP proposal application, which would analyze information from RFP discussions, data files, ticket dumps, and other inputs shared by the potential customer. It would then prepare a comprehensive summary that preserves the exact data, key focus areas, and anomalies that the team needs to address to respond correctly in the letter of execution or statement of work. These approaches are practical and require low effort within our current workflows.

Compared to Bard, now rebranded to Gemini, Gemini 2.5 Pro features very large context windows in Vertex AI, with a maximum input of about 1,048,576 tokens and a maximum output of around 65,535 tokens, along with tight integration with Google products. It accepts text, code, images, audio, and video as inputs and produces text. GPT‑5 is notable for controllable reasoning, enhanced tool use, and being the new default within ChatGPT and the OpenAI ecosystem, with a context window of roughly 400,000 tokens and up to 128,000 tokens of output. Both models are multimodal and capable of coding and analysis. If large amounts of text need to be processed in a single pass or if native Workspace integration is desired, Gemini 2.5 Pro is a good option. For steerable reasoning in agent-style tasks and simple access via ChatGPT and the OpenAI API, GPT‑5 is the more suitable choice.

# Conclusion

Our view is simple. GPT‑5 is important because it makes complex tasks more reliable, provides useful controls for how much thinking to do, and supports long prompts with a large context window. GPT‑4 is already capable of handling document answers with citations, meeting-to-action conversions, proposal drafting, and data summaries. Bard is now Gemini, and the right head-to-head comparison is GPT‑5 versus Gemini 2.5 Pro. We would choose Gemini when the job requires a massive context window or deep Google integration. We would choose GPT‑5 when strong, controllable reasoning and reliable tool use are crucial for real work.

**References:**

Introducing GPT-5. (2025, August 7). OpenAI. <https://openai.com/index/introducing-gpt-5/>

Kilpatrick, L., Mallick, S. B., & Kofman, R. (2024, June 27). Gemini 1.5 Pro 2M context window, code execution capabilities, and Gemma 2 are available today. <https://developers.googleblog.com/en/new-features-for-the-gemini-api-and-google-ai-studio/>

Hsiao, S. (2024, February 8). Bard becomes Gemini: Try Ultra 1.0 and a new mobile app today. Google. <https://blog.google/products/gemini/bard-gemini-advanced-app/>

Pichai, S. (2024, September 27). The next chapter of our Gemini era. Google. <https://blog.google/technology/ai/google-gemini-update-sundar-pichai-2024/>

Gemini 2.5 Pro. (August 2025). Generative AI on Vertex AI  |  Google Cloud. <https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-5-pro>