

# **Model Configuration Update - January 8, 2026**

---

## **Summary**

---

Updated PSScript AI service to use the latest and most capable OpenAI models available as of January 8, 2026.

# Models Updated

---

## Primary Analysis Model

- **Previous:** gpt-4o / o3-mini
- **Updated to:** gpt-5.2-codex
- **Reason:** GPT-5.2-Codex is OpenAI's most advanced agentic coding model for complex, real-world software engineering tasks
- **Performance:** 74.9% on SWE-bench Verified, 88% on Aider Polyglot

## Reasoning Model (New)

- **Added:** gpt-5.2
- **Purpose:** For complex reasoning and analysis tasks
- **Performance:** 94.6% on AIME 2025 (math), 84.2% on MMMU (multimodal understanding)

## Embedding Model

- **Previous:** text-embedding-ada-002 (1536 dimensions)
- **Updated to:** text-embedding-3-large (3072 dimensions)
- **Reason:** Most capable embedding model for semantic search and similarity
- **Performance:** MIRACL scores increased from 31.4% to 54.9%, MTEB scores from 61.0% to 64.6%
- **Cost:** \$0.13 per million tokens (vs \$0.10 for ada-002)

## Files Modified

---

1. **src/ai/config.py**
2. Updated `AgentConfig.default_model` to `gpt-5.2-codex`
3. Added `AgentConfig.reasoning_model` set to `gpt-5.2`
4. Modified model initialization logic
5. **src/ai/analysis/script\_analyzer.py**
6. Updated `ANALYSIS_MODEL` to `gpt-5.2-codex`
7. Updated `EMBEDDING_MODEL` to `text-embedding-3-large`
8. Updated `EMBEDDING_DIMENSION` to 3072
9. **src/ai/agents/langchain\_agent.py**
10. Updated `model_name` to `gpt-5.2-codex`
11. **src/ai/agents/langgraph\_production.py**
12. Updated all model defaults to `gpt-5.2-codex`

# Model Comparison

---

## GPT-5.2 vs GPT-4

GPT-5.2 represents a significant leap forward from GPT-4: - **Math**: 94.6% on AIME 2025 (without tools) - **Coding**: 74.9% on SWE-bench Verified - **Multimodal**: 84.2% on MMMU - **Healthcare**: 46.2% on HealthBench Hard

## GPT-5.2 Variants

- **GPT-5.2 Thinking**: Most advanced for professional work and long-running agents
- **GPT-5.2 Instant**: Faster version for everyday tasks
- **GPT-5.2-Codex**: Specialized for complex software engineering (selected for this project)

# Configuration

---

The models can be overridden via environment variables:

```
ANALYSIS_MODEL=gpt-3.5-turbo  
EMBEDDING_MODEL=text-embedding-3-large
```

## Verification

---

Service logs confirm the update:

```
ScriptAnalyzer initialized with model gpt-5.2-codex
```

## Cost Implications

---

- **GPT-5.2-Codex:** Pricing not yet publicly available (Pro/Enterprise tier)
- **text-embedding-3-large:** \$0.13/1M tokens (30% increase from ada-002)
- **Mock mode** remains enabled by default to avoid unexpected costs during development

## References

---

- [OpenAI GPT-5.2 Announcement](#)
- [OpenAI GPT-5.2-Codex](#)
- [OpenAI Embedding Models](#)
- [Embedding Models Guide](#)

## Next Steps

---

1. Test the updated models with real PowerShell script analysis
  2. Monitor performance and cost
  3. Consider enabling GPT-5.2 Thinking for complex analysis tasks
  4. Evaluate GPT-5.2 Instant for faster, lower-cost operations
- 

**Updated:** January 8, 2026 **Status:**  Deployed and Verified

Generated 2026-01-16 23:34 UTC