OpenAI Dev Day 2025

Topics

- Apps in ChatGPT - Building custom applications within ChatGPT
- AgentKit - Framework for autonomous AI agents
- Sora2 API - Next-generation video generation
- Codex - Enhanced code generation capabilities
- GPT-5 Pro in the API - Advanced reasoning and performance
- gpt-realtime-mini - Lightweight real-time interactions
- gpt-image-1-mini - Efficient image generation

Apps in ChatGPT

Apps in ChatGPT - Overview

- **Native App Integration**: Build and deploy apps directly within ChatGPT interface
- **Seamless User Experience**: No context switching between tools
- **Rich Interactions**: Support for forms, buttons, and interactive elements
- **Developer-Friendly**: Simple API integration with existing workflows

Apps in ChatGPT - Implementation

- **Custom Actions**: Define specific workflows and automations
- **Data Persistence**: Maintain state across conversations
- **Third-Party Integrations**: Connect external services and databases
- **Real-time Updates**: Live data synchronization and notifications
- **Model Context Protocol**: Apps SDK as an open standard built on the MCP

Apps coming soon

- OpenTable
- UBER
- Target
- DoorDash
- Tripadvisor

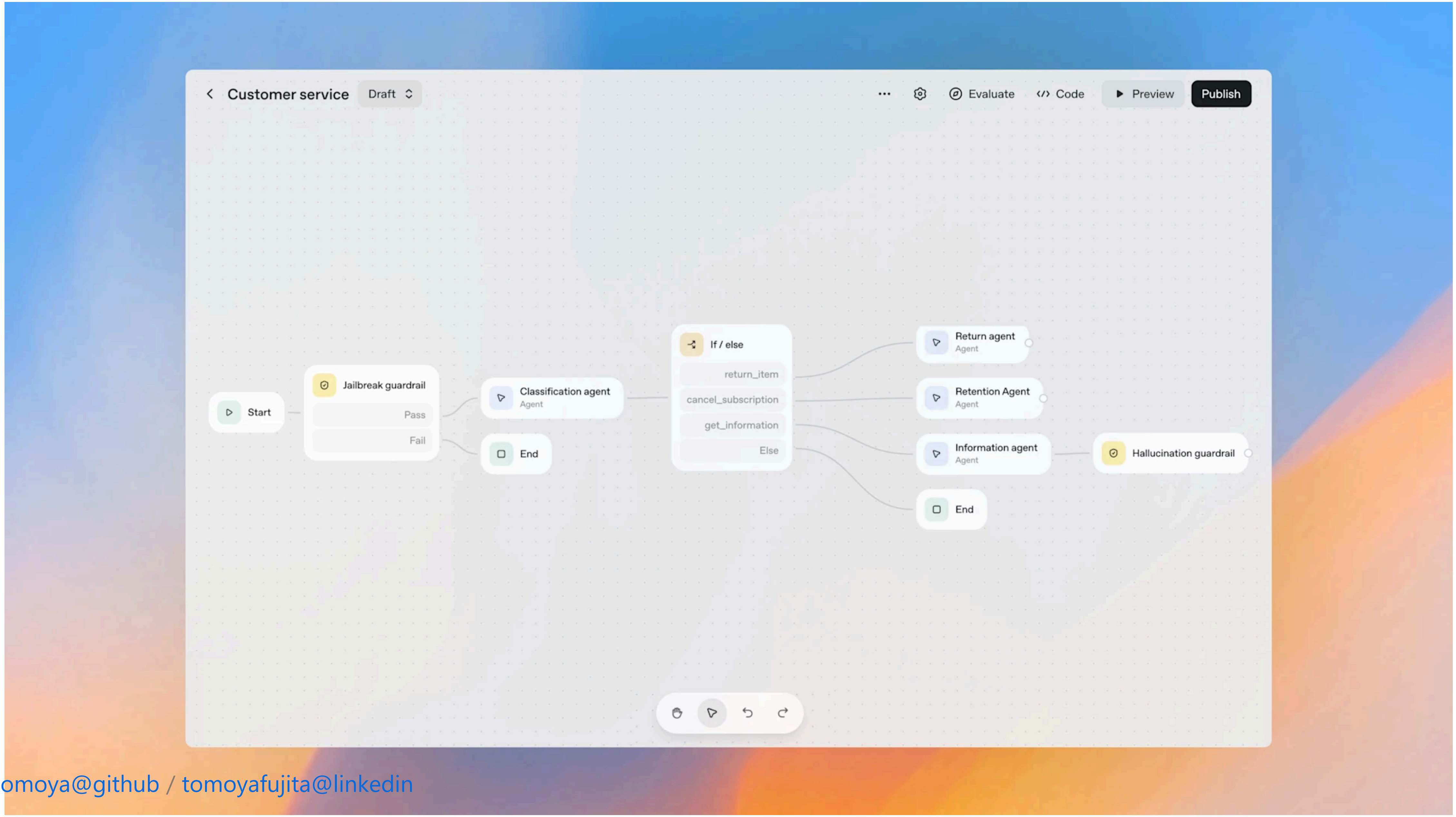
AgentKit

Agent: Systems that do things for you!

With AgentKit, developers can now design workflows visually and embed agentic UIs faster using new building blocks.

- **Agent Builder**: a visual drag-and-drop workflow editor / canvas for composing agent logic, branching, tool calls, etc.
- **ChatKit**: a UI toolkit / embed component to put a chat-based agent experience into your product (website/app) with streaming responses, chat threads, widgets, etc.
- **Connector Registry**: a config/management layer for data/tool integrations (connectors to e.g., Dropbox, Google Drive, Microsoft Teams) so agents can safely use enterprise/internal data.
- **Evaluation / Evals**: Tools to measure agent performance, run trace-based grading (end-to-end workflows), automated prompt optimization, support for third-party models in eval pipelines.
- **Guardrails / Safety**: Modular open-source safety/guard-rail library to detect issues (PII, jailbreaks, unintended behavior) when deploying agents.

OpenAI Dev Day 2025



Sora2 API

generates good quality results quickly, making it well suited for rapid iteration, concepting, and rough cuts.
`sora-2` is often more than sufficient for social media content, prototypes, and scenarios where turnaround time matters more than ultra-high fidelity.

Sora2 API

- **Job Creation**: Call `POST /videos` endpoint to start video generation
- **Status Monitoring**: Poll `GET /videos/{video_id}` or use webhooks for status updates
- **Job Completion**: Wait for status to transition to `completed`
- **Download Video**: Fetch final MP4 file via `GET /videos/{video_id}/content`

```python
import asyncio

from openai import AsyncOpenAI

client = AsyncOpenAI()


async def main() -> None:
    video = await client.videos.create_and_poll(
        model="sora-2",
        prompt="A video of a cat on a motorcycle",
    )

    if video.status == "completed":
        print("Video successfully completed: ", video)
    else:
        print("Video creation failed. Status: ", video.status)


asyncio.run(main())
```

Thumbnail / Spritesheet

```
# Download a thumbnail
curl -L "https://api.openai.com/v1/videos/video_abc123/content?variant=thumbnail" \
  -H "Authorization: Bearer $OPENAI_API_KEY" \
  --output thumbnail.webp

# Download a spritesheet
curl -L "https://api.openai.com/v1/videos/video_abc123/content?variant=spritesheet" \
  -H "Authorization: Bearer $OPENAI_API_KEY" \
  --output spritesheet.jpg
```

- You can guide a generation with an input image, which acts as the first frame of your video.
- Remix lets you take an existing video and make targeted adjustments without regenerating everything from scratch.
  By constraining each remix to one clear adjustment, you keep the visual style, subject consistency, and camera framing stable, while still exploring variations in mood, palette, or staging.

Codex

Advanced Code Generation

- **Multi-Language Support**: Enhanced support for 100+ programming languages
- **Context Awareness**: Better understanding of project structure and dependencies
- **Code Optimization**: Automatic performance improvements and refactoring
- **Security Focus**: Built-in vulnerability detection and secure coding practices

Codex - What's new?

- **Admin Features**: Edit and delete the sensitive data from Codex cloud environment

GPT-5 Pro

GPT-5 pro is available in the Responses API only to enable support for multi-turn model interactions before responding to API requests, and other advanced API features in the future. Some requests may take several minutes to finish. To avoid timeouts, try using background mode. As most advanced reasoning model, GPT-5 pro defaults to (and only supports) `reasoning.effort: high`.

gpt-realtime-mini

A cost-efficient version of GPT Realtime - capable of responding to audio and text inputs in realtime over WebRTC, WebSocket, or SIP connections.

For live chat, transcription, live captions, voice assistants and IoT integration.

gpt-image-1-mini

A cost-efficient version of GPT Image 1. It is a natively multimodal language model that accepts both text and image inputs, and produces image outputs.

Fast, lightweight, rapid prototyping, interactive apps, demos.