





LLMs in Research Computing

Deploying an Open-Source LLM Stack at Scale

Simon Stone, Jonathan Crossett

Research Computing and Data, Dartmouth College

Providing institutional access

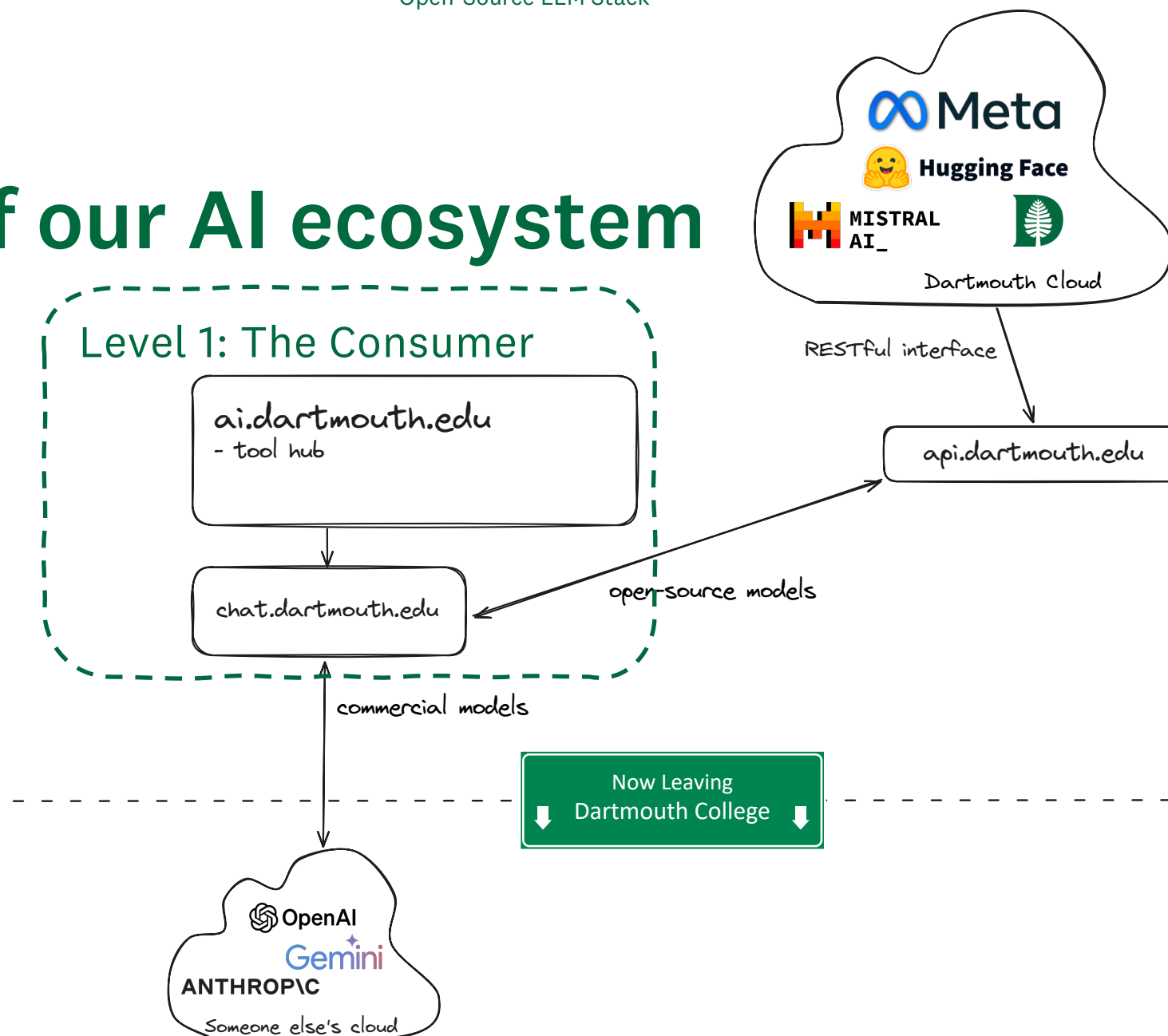
Why build your own LLM stack?

- Many providers have entered the chat
- Commercial providers offer enterprise subscription services with seat-based licenses
- Providing choices to our community is paramount
- Multiple subscriptions quickly becomes unaffordable



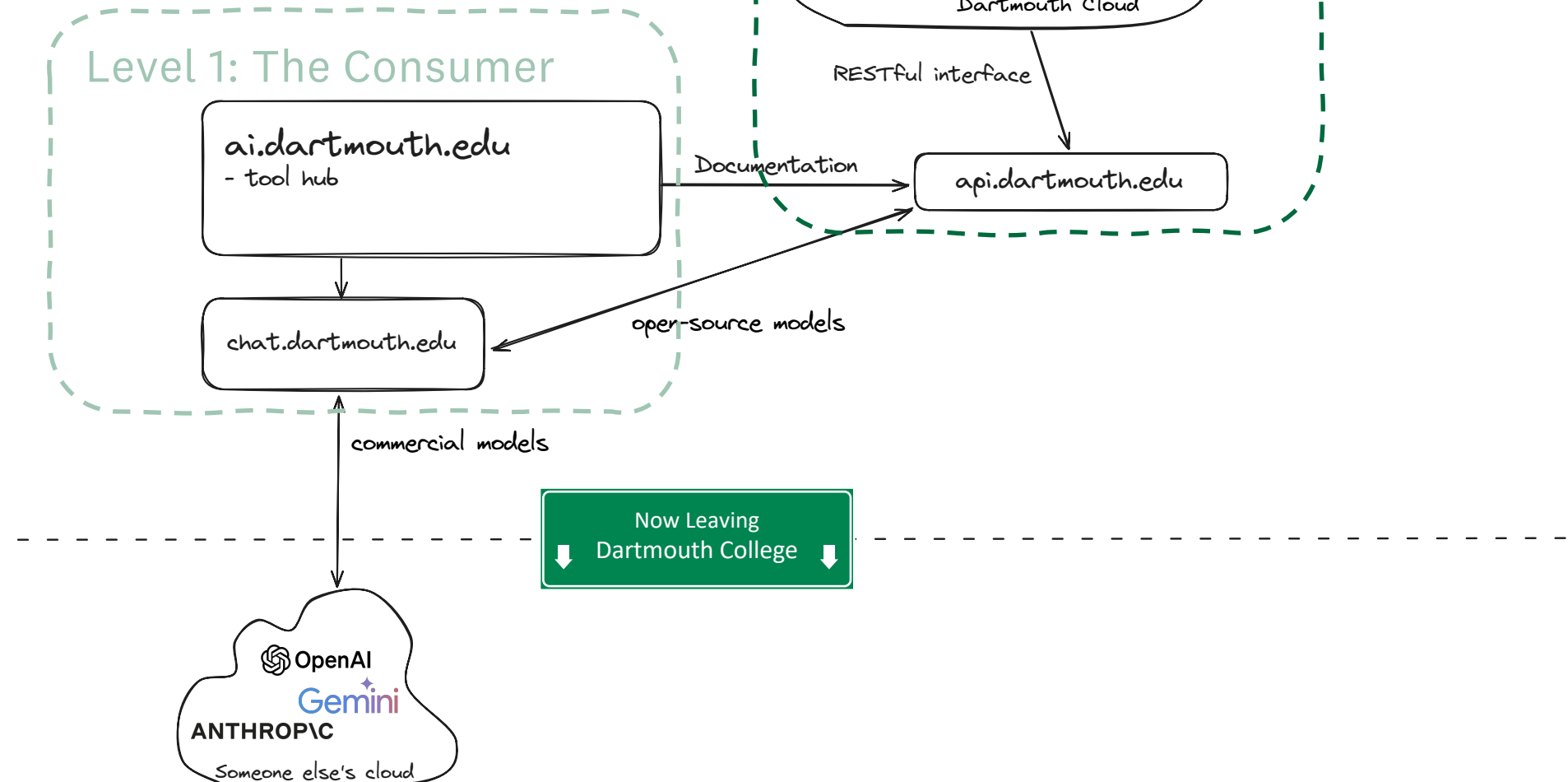


Overview of our AI ecosystem



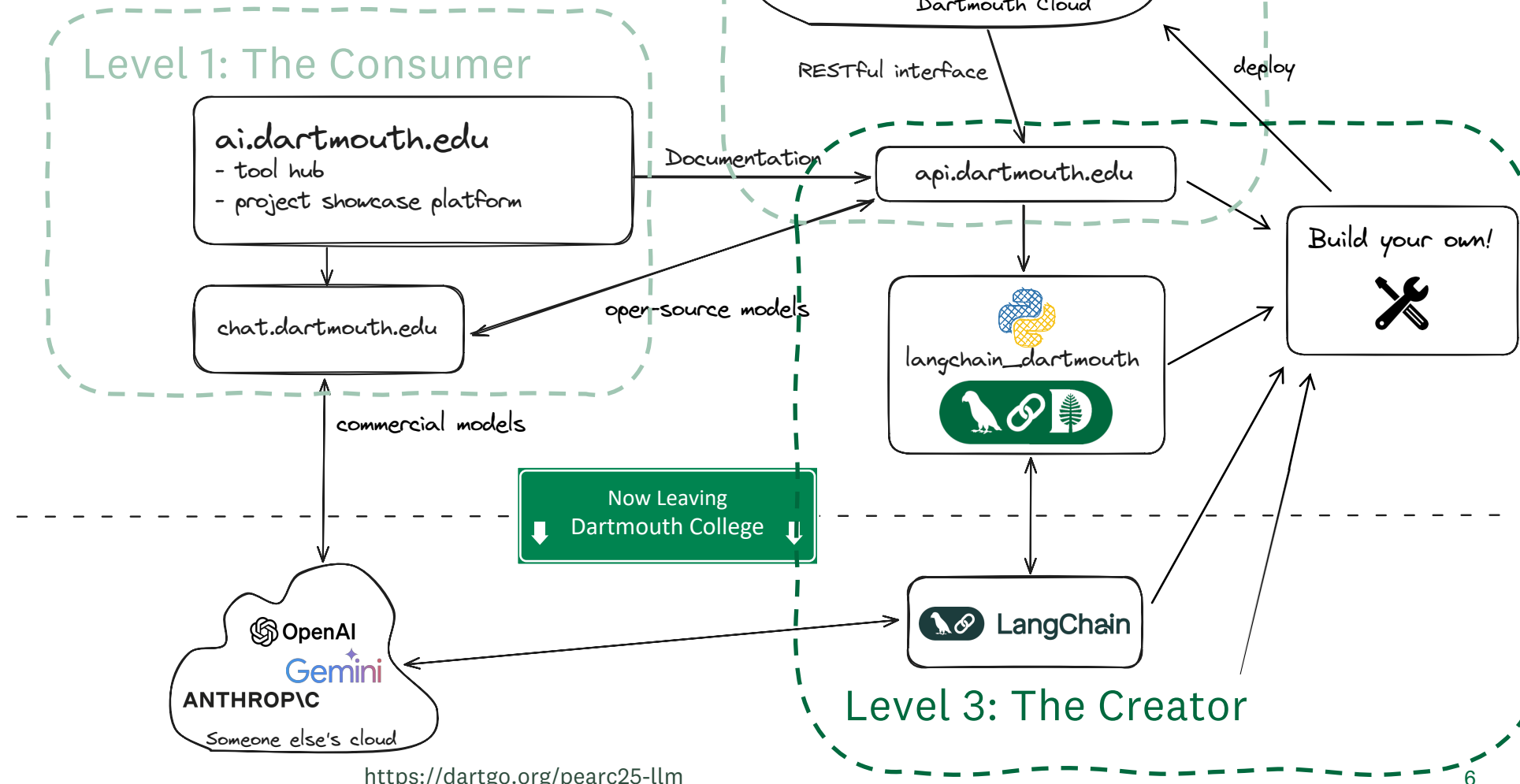


Overview of our AI ecosystem



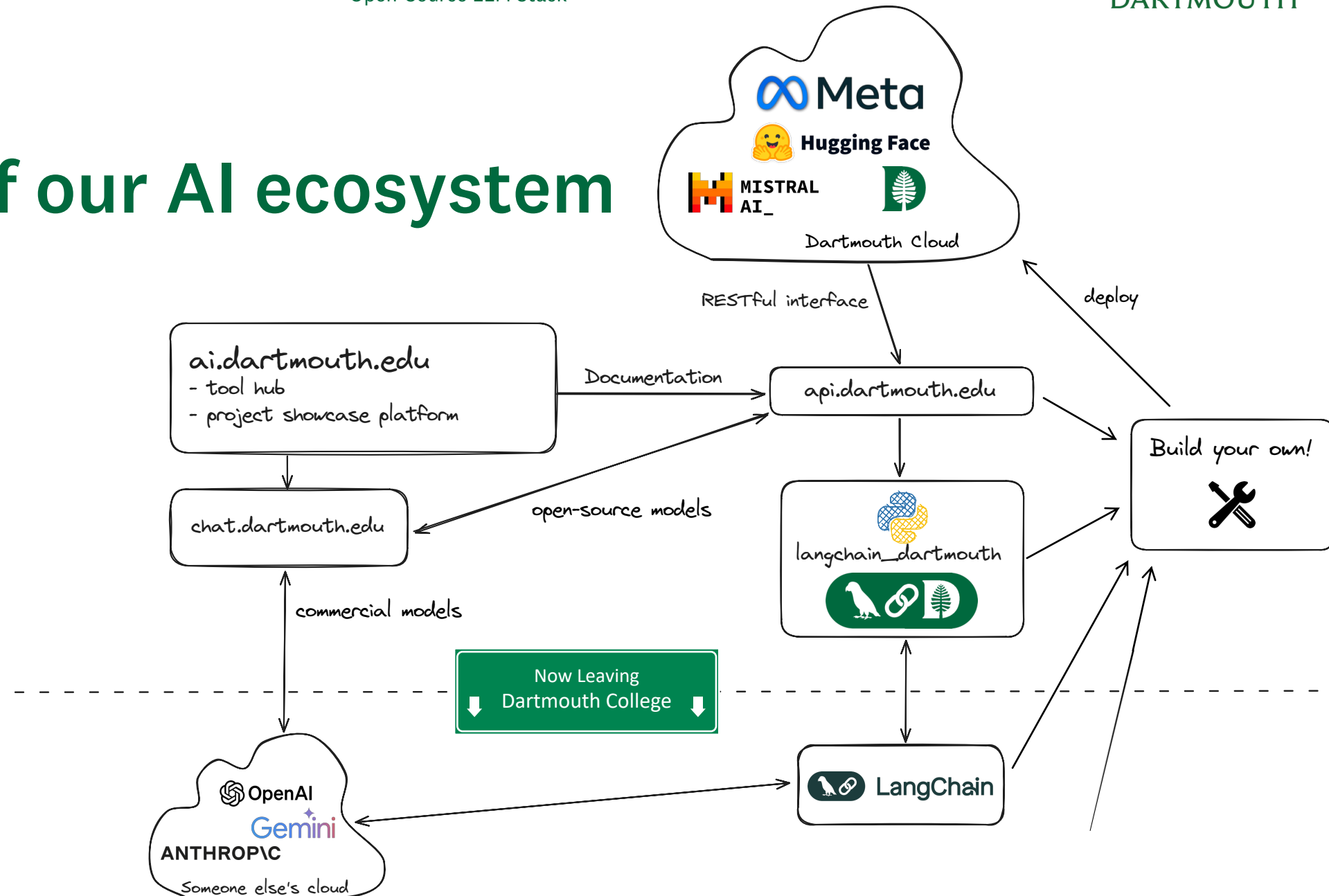


Overview of our AI ecosystem





Overview of our AI ecosystem





What is Dartmouth Chat?

- 💬 A web-based AI chat application hosted and managed by Dartmouth
- 🕒 Built on Open WebUI, an open-source*, feature-rich web interface for LLM chats
- 🌳 Available now at chat.dartmouth.edu
- 🌳 About 4400 registered and 1000 concurrent users (~38 % of the College)

i.e. “source available”



Dartmouth Chat Demo

<https://chat.dartmouth.edu>



Use Cases

French Plays (Scott Sanders, French & Italian)

Project brief:

Extracting structured data from French play announcements

How we support it:

 Uses on-premise LLM

 Works with components from `langchain_dartmouth`



Use Cases

Career Exploration Chatbot (Chris Jordan, DCCCD)

Project brief:

A chatbot supporting students' explorations of possible career paths

How we support it:

- Uses Open WebUI's Workspace (cf. OpenAI's Custom GPTs)



Now build *your own*!

Hands-on

