Building the ‘Chat IEEE Standards’ ChatGPT Plugin – Technology & Business Strategy

Matthew R. Scott, IEEE Industry Engagement Committee

December 7th, 2023

Prepared for the IEEE TEMS Silicon Valley Chapter

IEEE

Advancing Technology for Humanity

Outline

1. Background & Motivation
2. Demo Walkthrough
3. Business Strategy
4. Technical Approach
5. Legal, Compliance & Policy
6. What's Next & Concluding Remarks

Background & Motivation

Background of this Project

- This volunteer-driven, ongoing project is one of the initiatives of the IEEE Industry Engagement Committee, led by the Software Tools Working Group, in collaboration with the IEEE Standards Association.
- Thesis is that industry needs access to standards and so making a “SA app” for ChatGPT can improve the user experience working standards, which can increase engagement between IEEE and industry.
- Additionally, other goals include exploring news ways to make IEEE content discoverable to industry and opportunities for new business models to help generate revenue for the IEEE in the GenAI era.

Motivation for this Talk

- Although this project is focused on SA, in this talk, it is our hope the findings along the way can be instructive to the wider IEEE community.
- How to think differently from a business perspective about traditional paywalled online content where the user interface may be shifting from web browsing to more natural language chats with AI.
- As the way to interact with information transitions from web directories, to search engines, and now chat, new business and technical opportunities emerge as well as new challenges. We hope what we’ve learned so far can be helpful for others.

Demo Walkthrough

Business Strategy

Cost Strategy

- A common challenge in LLM deployment is **cost**. Today, LLM-based services are largely **unprofitable** due to the massive resources needed to train and host these services as scale.
- With an “app” approach to a SOTA LLM service like ChatGPT (GPT-4), **we do not need to retrain or serve the model ourselves**.
- The model service is already being paid for by user subscriptions to ChatGPT. In our strategy, we can leverage that existing arrangement with no additional costs to the IEEE.
- Our hosting **costs are significantly minimized** to only the backend services API enabling knowledge injection on the fly.
- For reference numbers, training and hosting a LLM service can cost hundreds of thousands to millions of dollars. Our service cost, even at scale, is estimated at < $15K per year. More on how that’s possible later.

Gateway Strategy via AI Agents and Tools

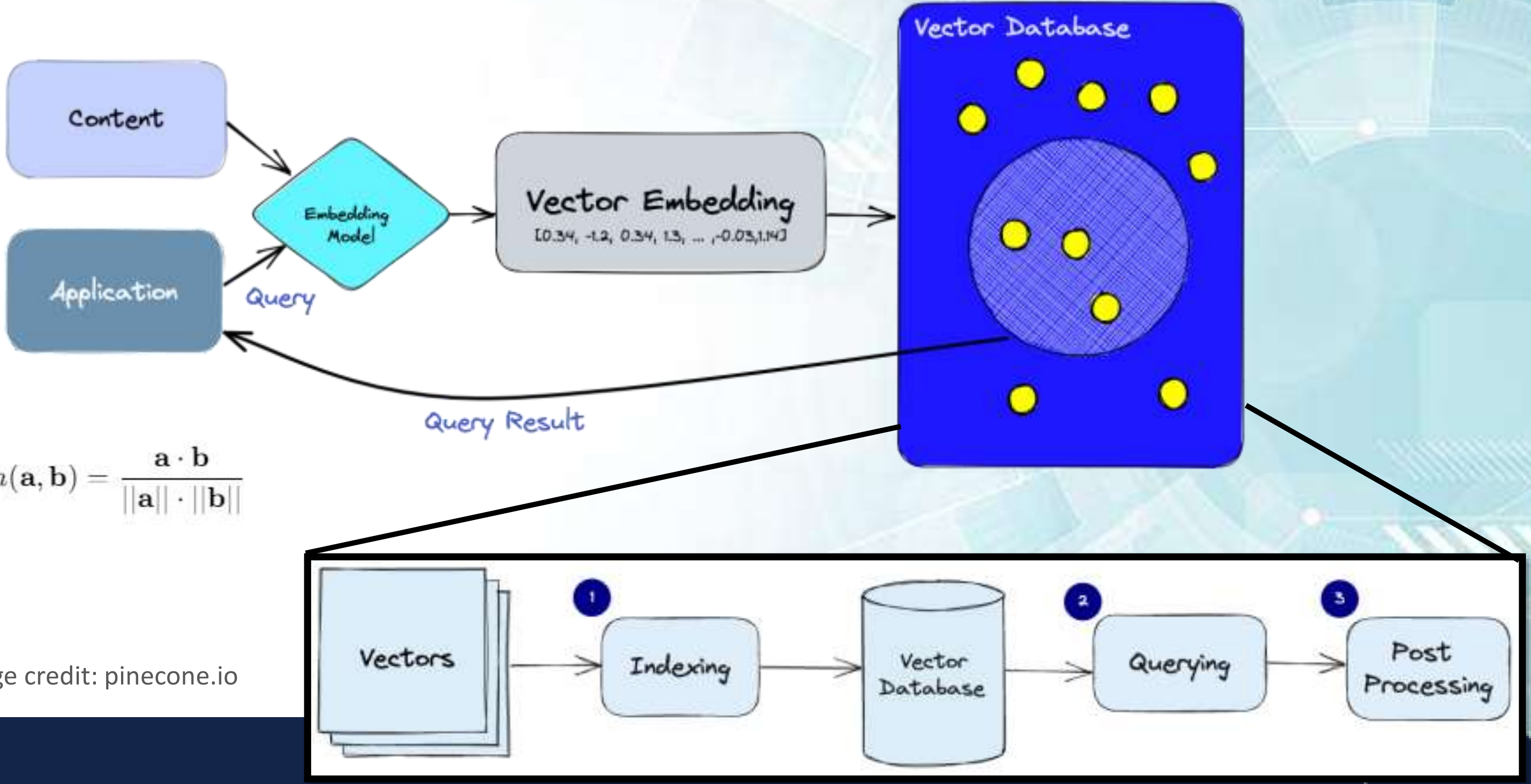
- Although LLMs themselves can contain significant knowledge, our strategy is to separate out our business information from the language and reasoning capabilities of LLMs.
- That is the key concept: IEEE business knowledge can be delineated from the rest of the LLM “brain” that users interact with.
- Effectively then, chat is the “user interface” and IEEE can act as “DB”.
- How? Through the “app” framework for AI platforms like ChatGPT, we control the knowledge relevant to our business scenario by injecting it on the fly into the “working memory” of the LLM, which is not stored or part of the model itself.
- Think of it as “tool usage”, e.g., the LLM (acting as an “Agent”) accesses our IEEE SA tool when it needs Standards knowledge, much like human beings reach for tool when needed.

Business Model

- Building on the already existing subscription model of IEEE membership, we have options:
  - An additional subscription tier can be added on top of the base to enable access to the Chat IEEE Standards “app” to unlock this capability. E.g., an additional $X / month.
  - Or, instead of adding new tiers, because the costs are so low, this service and others like this, can just be added to the existing base subscription to make general membership more appealing to help improve current membership numbers.
- Authentication is Key
  - As shown in the demo, when a user attempts to install or use the “app”, e.g., our plugin, the user is asked to login. This allows the IEEE server side to verify the user is a paying subscriber, and if not, redirect them to the appropriate page, such as to upgrade.
  - Additionally, by authenticating, this system can also generate metrics on usage and implicit feedback which we can further use to measure value, improve the service, and potentially recommend additionally relevant IEEE products and services.

Technical Approach

Embeddings and Vector Databases



$$sim(\mathbf{a},\mathbf{b}) = \frac{\mathbf{a}\cdot\mathbf{b}}{||\mathbf{a}||\cdot||\mathbf{b}||}$$

Image credit: pinecone.io

Retrieval Augmented Generation (RAG)

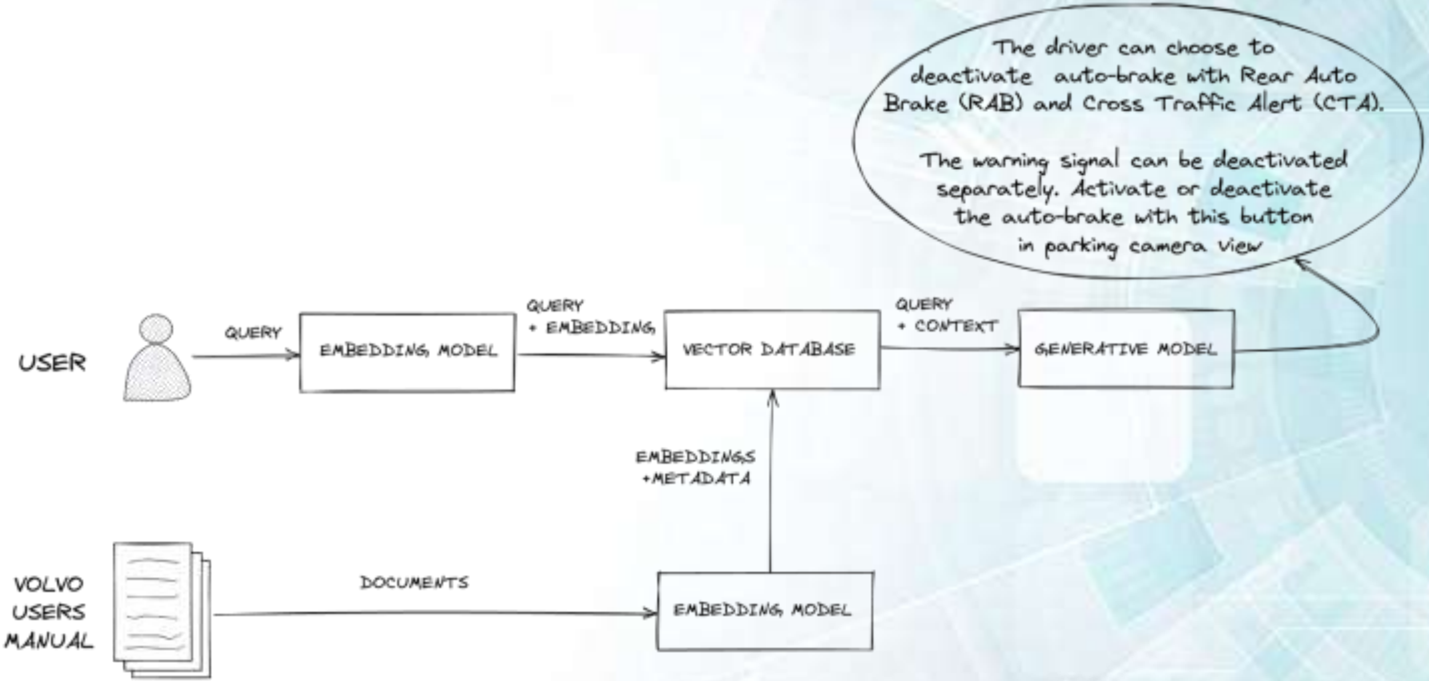
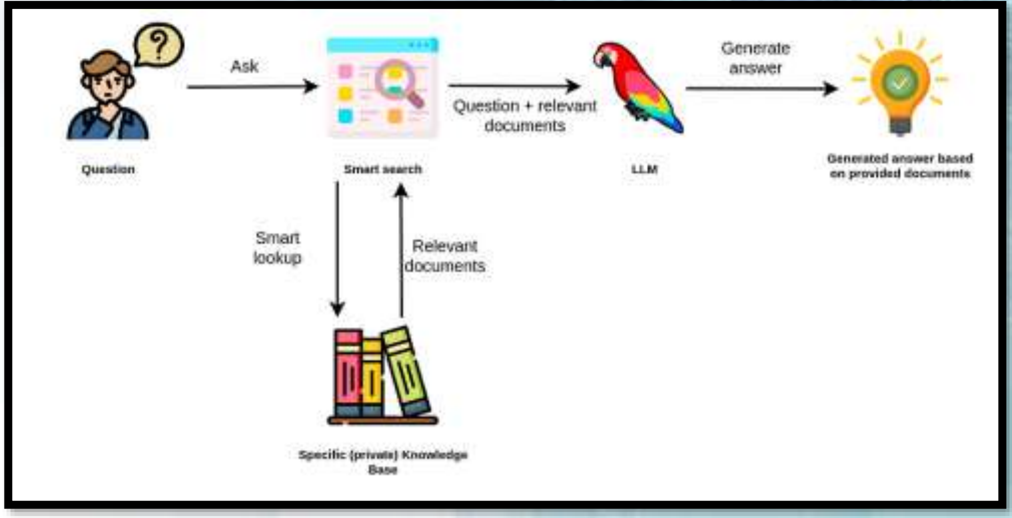

The driver can choose to deactivate auto-brake with Rear Auto Brake (RAB) and Cross Traffic Alert (CTA).
The warning signal can be deactivated separately. Activate or deactivate the auto-brake with this button in parking camera view
USER
QUERY
EMBEDDING MODEL
QUERY + EMBEDDING
VECTOR DATABASE
QUERY + CONTEXT
GENERATIVE MODEL
EMBEDDINGS +METADATA
VOLVO USERS MANUAL
DOCUMENTS
EMBEDDING MODEL

Image credit: pinecone.io

RAG versus Fine-tuning

- Fine-tuning is the process of retraining a foundation model on new domain-specific data. It is cheaper than building from scratch, but still can be very expensive to train and maintain.
- For data that changes over time such as Standards docs, a fine-tuning approach is inefficient due to requiring constant retraining for every new doc or change, which increases costs, complexity and time.


Question
Ask
Smart search
Question + relevant documents
LLM
Generate answer
Generated answer based on provided documents
Smart lookup
Relevant documents
Specific (private) Knowledge Base

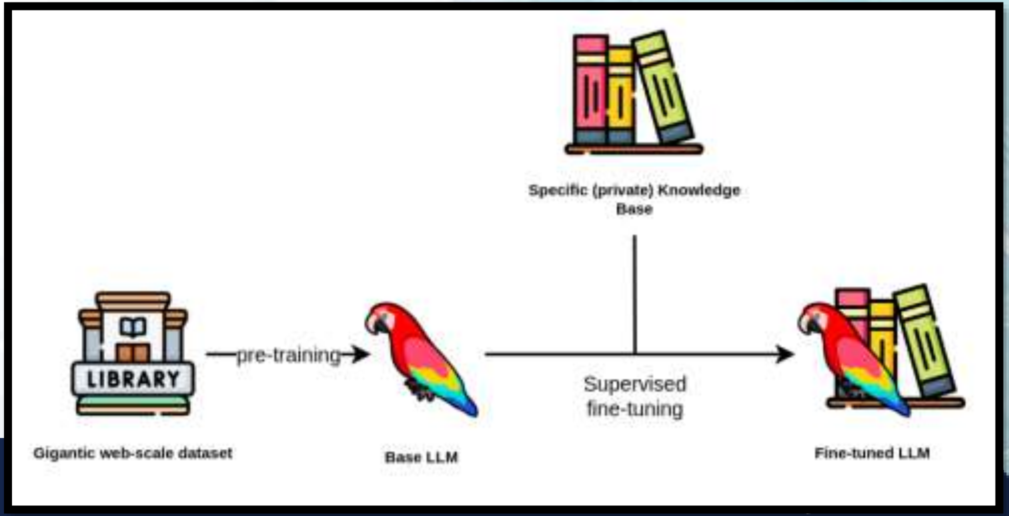
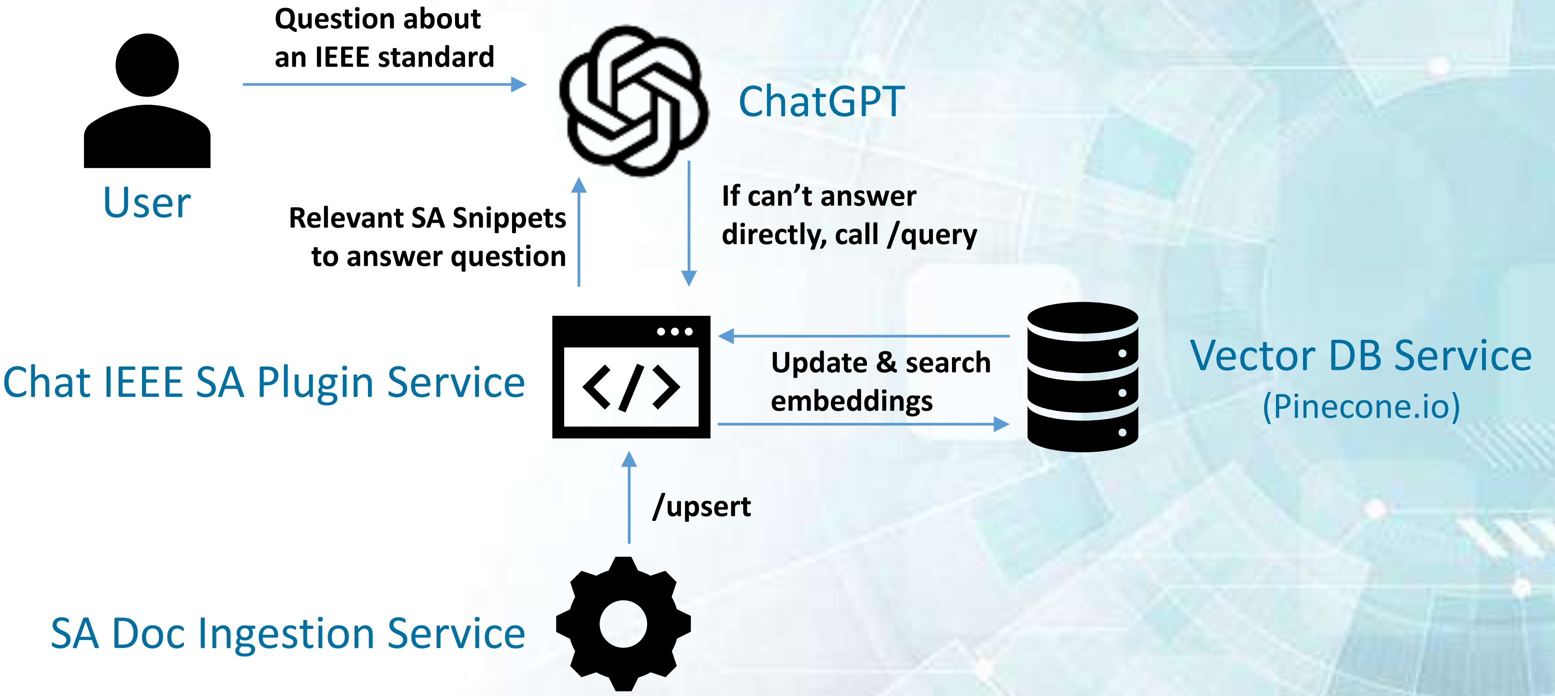


Image credit: neo4j.com

How do ChatGPT Plugins work?

- OpenAI plugins link ChatGPT to external applications.
- The plugins allow ChatGPT to interact with developer-defined APIs, expanding its functions and action range.
- Plugin developers provide API endpoints, a manifest file, and an OpenAPI specification.
- These elements specify the plugin's operations, enabling ChatGPT to utilize the files and access the APIs.
- The ChatGPT LLM functions as an “intelligent API caller”, utilizing API specifications and natural-language instructions, the model decides when to call the API to perform actions.
- Plugins require registration with OpenAI and user installation (with authentication) for activation.

Chat IEEE Standards Plugin Data Flow


User
Question about an IEEE standard
ChatGPT
Relevant SA Snippets to answer question
If can't answer directly, call /query
Chat IEEE SA Plugin Service
Update & search embeddings
Vector DB Service
(Pinecone.io)
/upsert
SA Doc Ingestion Service

Legal, Compliance & Policy

Policy Status

- Like many large organizations around the world, legal and security teams are currently evaluating if LLMs are compliant with their policies around IP risks by uploading data to a third-party service like ChatGPT.
- Various teams are working to examine the legal, operational, and ethical risks inherent in the use of AI. These include:
  - Confidentiality
  - Loss of IP
  - Ethical concerns due to bias
  - Liabilities
- We are currently waiting on policy changes that would enable us to proceed.

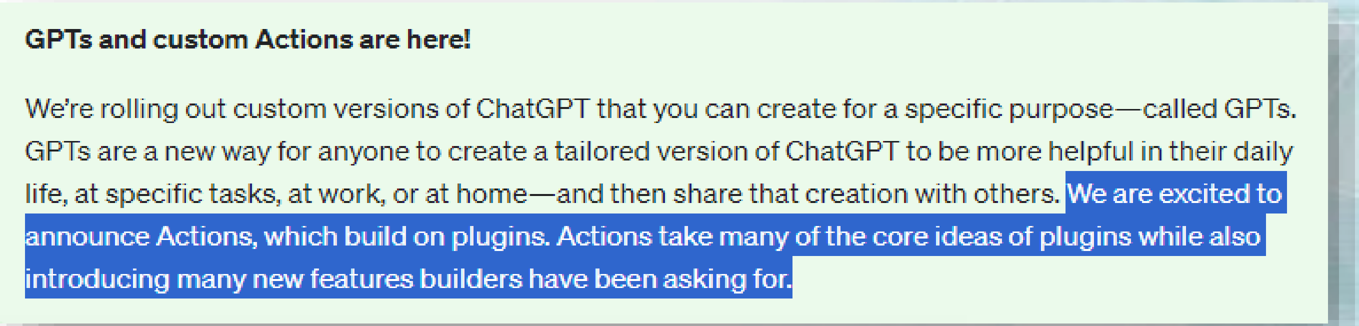
Considerations on Moving Forward from a Policy Perspective

- **Technical workarounds:**
  - We believe that technical workarounds exist by using enterprise versions of ChatGPT and / or Azure OpenAI Services.
  - Additionally, we believe that the current API data usage policy of OpenAI does help mitigate current concerns on safeguarding proprietary data from OpenAI's use to improve their models.
- **Strategic considerations:**
  - GenAI is transforming the way users interact with information. On Nov 6, 2023, OpenAI reported that ChatGPT now has over 100 million active users weekly and growing.
  - At the current volume, we suspect users are already, perhaps unknowingly, working with ChatGPT using proprietary content by copy/pasting in such docs themselves for Q&A.
  - AI plugins/apps like ours can not just help users be more productive but also protect the rights and business interests of copyright owners.
  - Therefore, it can be strategic to get out in front of this trend from a policy perspective.

What’s Next & Concluding Remarks

What’s Next for ChatGPT Plugins?

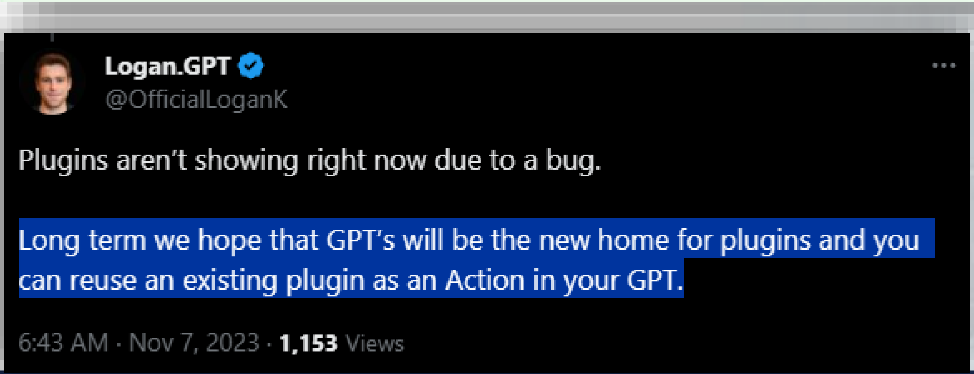
- On Nov. 6th, 2023, OpenAI announced GPTs and custom Actions



GPTs and custom Actions are here!

We’re rolling out custom versions of ChatGPT that you can create for a specific purpose—called GPTs. GPTs are a new way for anyone to create a tailored version of ChatGPT to be more helpful in their daily life, at specific tasks, at work, or at home—and then share that creation with others. We are excited to announce Actions, which build on plugins. Actions take many of the core ideas of plugins while also introducing many new features builders have been asking for.

Source: openai.com docs

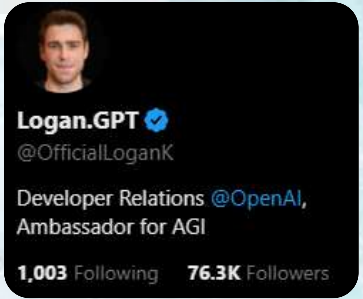


Logan.GPT
@OfficialLoganK

Plugins aren’t showing right now due to a bug.

Long term we hope that GPT’s will be the new home for plugins and you can reuse an existing plugin as an Action in your GPT.

6:43 AM · Nov 7, 2023 · **1,153** Views



Logan.GPT
@OfficialLoganK

Developer Relations @OpenAI, Ambassador for AGI

1,003 Following **76.3K** Followers

Talk Summary

- Demoed the “Chat IEEE Standards” plugin, enhancing access to SA docs, accuracy and UX
- Chat becoming the new user interface to knowledge, creating new opportunities & challenges
- Business implications and thinking behind the "app" approach to LLMs, in terms of:
  - Cost effectiveness (no need to retrain and host our LLM)
  - Knowledge Gateway strategy via AI Agents and Tools (data separated from UI and injected at runtime)
  - Authentication enabling the subscription business model
- The technical background and approach using retrieval augmented generation, supporting scalability, low costs, while maintaining controls over gated proprietary data.
- Legal, compliance and policy concerns and discussion.
- The future direction of ChatGPT plugins towards custom actions.

Thank you