

DSS Meetup #35
Per favore ChatGPT crea un titolo
per questo meetup





Introduzione

- ChatGPT
- Chatbot
 - Approccio passato
 - Presente
- Microsoft Semantic Kernel
- Deploy in Prod e Costi
- Oltre a ChatGPT
 - LangChain
 - Cheshire Cat
 - AutoGPT
 - Modelli offline
- Bonus: utilizzo quotidiano

Breve Sondaggio

Quanti utilizzano ChatGPT?

Quanti utilizzano o hanno utilizzato le API di OpenAI?

Quanti utilizzano le API su Azure?

Quanti sviluppatori C# ci sono? E Python? (previsione: pochi e molti)

Quanti conoscono LangChain?





ChatGPT: L'Intelligenza Artificiale Avanzata per Professionisti IT

Introduzione a ChatGPT

- Un sofisticato modello di linguaggio creato da OpenAI
- Basato sull'architettura GPT-4, rappresenta un punto di riferimento nell'IA

Conoscenza e Ambito

- Vasto database di conoscenze aggiornato fino a settembre 2021
- Copre aree come scienza, tecnologia, arte, cultura e molto altro

Funzionalità di ChatGPT

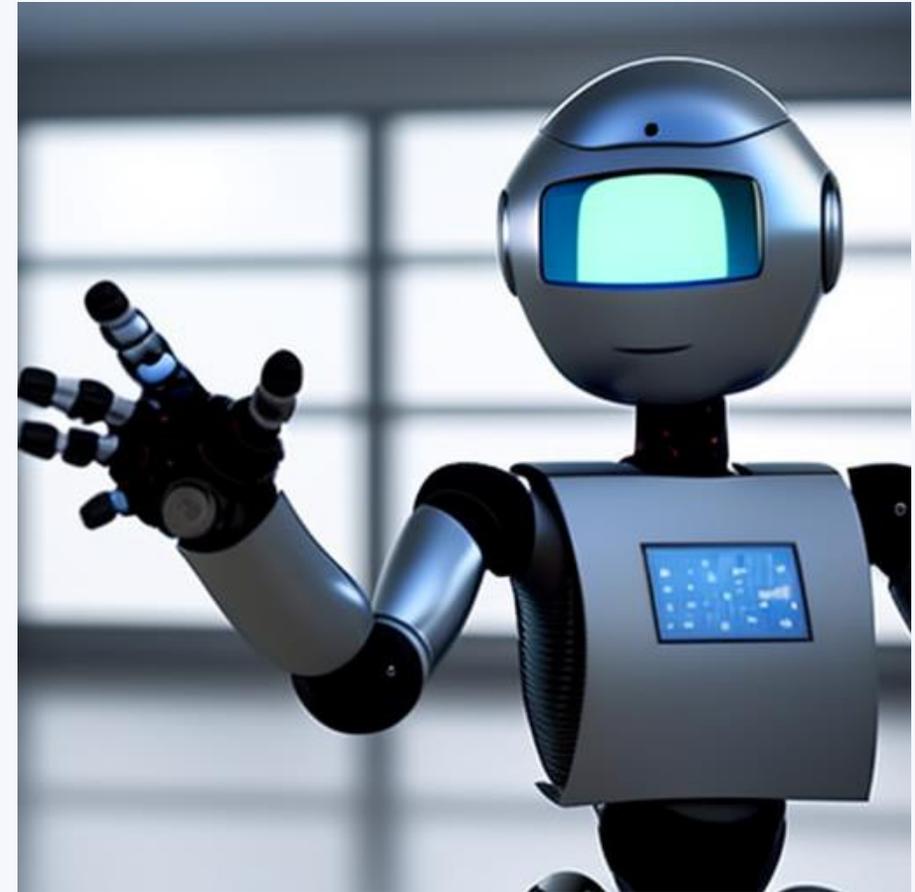
- Risponde alle domande con precisione, generando contenuti coerenti e informativi
- Strumento versatile per supportare professionisti IT in diverse applicazioni

Utilizzo Responsabile

- È fondamentale usare ChatGPT con etica e responsabilità
- Prestare attenzione alle informazioni errate o inesatte che possono emergere

Prospettive Future

- ChatGPT è un passo importante nel campo dell'intelligenza artificiale
- Nuovi sviluppi e innovazioni continueranno a migliorare l'utilizzo dell'IA per i professionisti IT





ChatGPT, l'Amico Magico delle Parole

Dove viene ChatGPT?

- **ChatGPT è un amico speciale fatto dai geni di OpenAI**
- È come un robot parlante super intelligente che sa un sacco di cose

Cosa sa ChatGPT?

- ChatGPT conosce tante cose fino al 2021, come un grande libro di storie
- Parla di scienza, computer, arte e tante altre cose interessanti

Cosa può fare ChatGPT?

- Risponde alle domande, racconta storie e scrive cose divertenti
- **È come avere un amico super esperto sempre con te**

Giocare con ChatGPT

- **Ricorda di usare ChatGPT per cose buone e di chiedere aiuto ai grandi se non capisci qualcosa**
- **A volte ChatGPT può sbagliare, ma è normale, anche gli amici magici sbagliano**

Il Futuro di ChatGPT

- ChatGPT è solo l'inizio di una grande avventura nel mondo delle macchine che parlano
- Ci saranno ancora più cose sorprendenti e divertenti da scoprire insieme





ChatGPT, il Guerriero della Conoscenza Mandaloriano

Origini del Guerriero

- Un potente modello di linguaggio addestrato dall'Ordine OpenAI
- Forgiato nell'architettura GPT-4, si distingue come un formidabile strumento di conoscenza

Memoria di un Mandaloriano

- Il suo sapere si estende fino a settembre 2021, un'inesauribile riserva di informazioni
- Abile nelle arti della scienza, tecnologia, arte e cultura, è un degno alleato

Le Abilità di ChatGPT

- Rispondere a domande, scrivere racconti e creare contenuti con precisione e astuzia
- Un partner prezioso nel viaggio attraverso le terre digitali

Uso Onorevole

- **Come un vero Mandaloriano, usate ChatGPT con onore e responsabilità**
- **Non lasciatevi ingannare dalle illusioni dell'informazione inesatta o fuorviante**

Il Futuro della Battaglia

- ChatGPT rappresenta l'inizio di un'epopea nell'inafferrabile mondo dell'intelligenza artificiale
- **Insieme, affronteremo nuove sfide e scoperte, onorando il codice Mandaloriano dell'innovazione. Questa è la via.**





Chatbot – l'altro ieri

Pillo/Pria

Assistente medico a interazione vocale in grado di dispensare medicine e utilizzato come hub per la salute della persona

Strumenti

- DialogFlow
 - Intents and Entities Recognition
 - Knowledge integration



The screenshot shows the Dialogflow Essentials interface. On the left is a navigation menu with options: Spria (en), Intents, Entities, Knowledge [beta], Fulfilment, Integrations, Training, Validation, History, Analytics, Prebuilt Agents, Small Talk, Docs, and Essentials (Pay as You Go). The main area is titled 'Intents' and contains a list of 19 intents, including Device.Brightness, Device.Carefriend, Device.Checkin, Device.Dial, Device.GetDate, Device.GetTime, Device.LocationTime, Device.NightMode, Device.OpenWellnessPlan, Device.Owner, Device.ScreenControl, Device.Support, Device.Updates, Device.Volume, Engagement.8Ball, Engagement.Facts, and Engagement.Poetry. On the right, a preview pane shows a user query: 'what's the weather in Genoa on May 8th at 6 PM?'. The system response is 'Not available'. Below this, the detected intent is 'PKS.Weather', the action is 'weather', and the parameters are listed in a table.

PARAMETER	VALUE
address	{ "shortcut": "", "island": "", "business-name": "", "country": "", "street-address": "", "zip-code": "", "subadmin-area": "", "admin-area": "", "city": "Genoa" }
weatherType	query
date-time	{ "date_time": "2023-05-08T18:00:00+02:00" }

Chatbot – l'altro ieri (piccola digressione)

Model: Default (GPT-3.5)

AS cos'è Pillo?

 Pillo è un dispositivo robotico per la salute domestica che utilizza l'intelligenza artificiale per fornire supporto sanitario personalizzato e migliorare il benessere degli individui.

Il dispositivo è stato progettato per aiutare le persone a gestire i loro farmaci, monitorare la loro salute e fornire informazioni utili sulle condizioni mediche e sullo stile di vita. Pillo è dotato di un display touchscreen e di una fotocamera per la rilevazione facciale, che consente di identificare gli utenti e fornire loro informazioni personalizzate.

Attraverso l'uso di algoritmi di intelligenza artificiale, Pillo è in grado di riconoscere la voce dell'utente e interagire con lui in modo naturale, fornendo risposte immediate alle domande e alle preoccupazioni sulla salute. Il dispositivo può anche inviare promemoria per la somministrazione di farmaci e contattare il medico in caso di emergenza.

Inoltre, Pillo è in grado di integrarsi con altri dispositivi sanitari, come i monitor della pressione sanguigna e i dispositivi di monitoraggio della glicemia, per fornire una visione più completa della salute dell'utente.

Chat Compose Insights

Welcome back! What would you like to chat about?

cos'è Pillo?

✓ Searching for: Pillo

✓ Generating answers for you...

Pillo è un robot per la salute domestica che combina le funzioni di un distributore di pillole con Siri, l'apprendimento automatico e due occhi sempre aperti. Utilizza il riconoscimento facciale per identificare diversi individui e quindi distribuisce le pillole appropriate a quelle persone al momento in cui devono essere prese. Inoltre, se vede che potresti aver dimenticato di prendere le tue pillole, ti ricorderà di prenderle prima che tu te ne vada ¹. Pillo può anche rispondere a domande sulla salute, sincronizzare i dati con il tuo telefono e smartwatch e persino impostare videochiamate con professionisti della salute ¹. È stato lanciato su Indiegogo ².

Learn more:

1. [digitaltrends.com](https://www.digitaltrends.com)

2. [indiegogo.com](https://www.indiegogo.com) +2 more

Retry for this page only





Chatbot - Ieri

MVP for Helicobacter Pylori Treatment Dataset

Creazione di un chatbot in grado di rispondere alle domande su Helicobacter Pylori e prendere dati sulla gestione del trattamento per l'eradicazione dal dataset europeo opportunamente elaborato

Strumenti

- Bot Composer
- Azure QnA Maker
- Azure LUIS

CONFIDENTIAL

RegimenStats
Intent recognized

```

    graph TD
      Start(( )) --> SetProp[Set a property  
use:regimen := @regimen]
      SetProp --> Branch[Branch: If/else  
=empty(user:regimen)]
      Branch -- True --> HTTP1[Send an HTTP request  
GET https://hperegapi.azurewebsites.net/api/patie...  
turn.results = Result property]
      Branch -- False --> HTTP2[Send an HTTP request  
GET https://hperegapi.azurewebsites.net/api/patie...  
turn.results = Result property]
      HTTP1 --> Resp1[Send a response  
Text: Most used regimen is ${turn.lastresult.content...}]
      HTTP2 --> Resp2[Send a response  
Text: ${turn.lastresult.content} patients are treated ...]
  
```

HpEuRegAssistant

Restart Conversation - new user ID

Hey, there! I'm Helicobacter Pylori Europe Registry Assistant. You can ask me disease information and statistics about patients and treatments. How can I help you today?

What is H. Pylori?
2 minutes ago

H. pylori (Helicobacter pylori) are bacteria that can cause an infection in the stomach or duodenum (first part of the small intestine). It's the most common cause of peptic ulcer disease. H. pylori can also inflame and irritate the stomach lining (gastritis). Untreated, long-term H. pylori infection can lead to stomach cancer (rarely).

Can H. pylori spread from person to person?
2 minutes ago

Yes, H. pylori can spread from person to person. H. pylori are found in saliva, plaque on teeth and poop. Infection can be spread through kissing and by transferring the bacteria from the hands of those who have not thoroughly washed them after a bowel movement.

Scientists think H. pylori also might be spread through H. pylori-contaminated water and food.

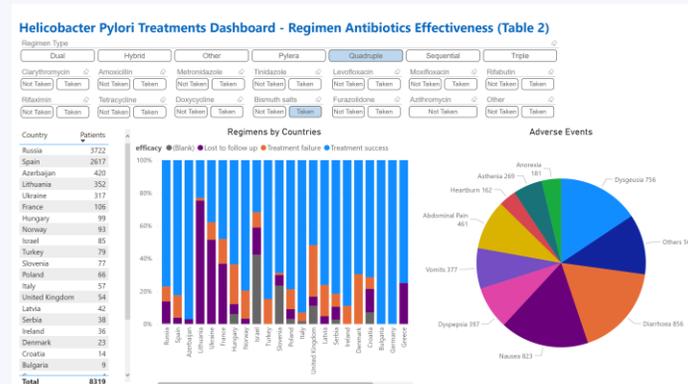
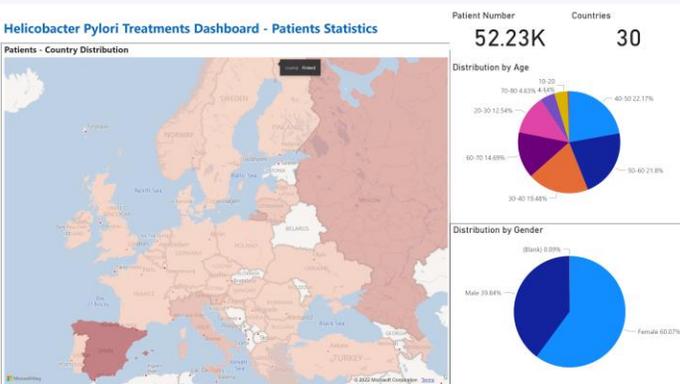
How many patients are treated in Spain?
A minute ago

Patient registry contains 5000 patients for Spain

What is the most used regimen?
A minute ago

Most used regimen is Triple used for 3000 patients

Type your message





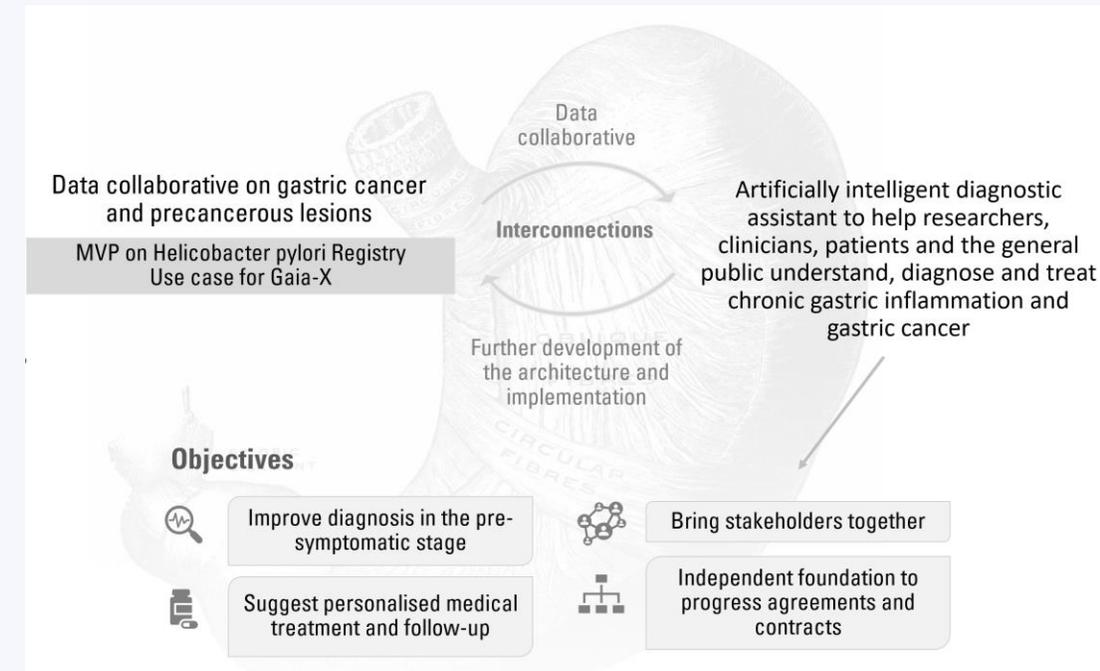
Chatbot – oggi

AIDA - An artificially intelligent diagnostic assistant for gastric inflammation

Il progetto AIDA mira a sviluppare e convalidare un assistente multidisciplinare alimentato dall'IA per aiutare i medici a diagnosticare l'infiammazione precancerosa e suggerire strategie terapeutiche personalizzate per il trattamento medico e il follow-up. L'obiettivo finale è contribuire alla prevenzione del cancro gastrico attraverso il monitoraggio personalizzato dello stato di salute del paziente.

Strumenti

- OpenAI API (Azure)
- Open AI playground
- Semantic Kernel





Chatbot – oggi (altra digressione)

Table of contents

Table of contents	1
Abbreviations	3
1. Excellence	4
1.1. Objectives and ambition	4
1.1.1. Objectives	4
1.1.2. Ambition	8
1.1.3. Data-agnosticity	14
1.1.4. Long-term organisational sustainability	14
1.1.5. Long-term financial sustainability	16
1.1.6. Interest from other external stakeholders	18
1.1.7. R&I maturity	19
1.2. Methodology	19
1.2.1. Project Structure	19
1.2.2. Strategy and TRL level	20
1.2.3. Link with the European Health Data Space and GaiaX	21
1.2.4. Link with national or international R&I activities	21
1.2.5. Interdisciplinary approach	21
1.2.6. Integration of social sciences and humanities	21
1.2.7. Gender dimension	22
1.2.8. Open science	22
1.2.9. Management of data and other outputs	22
1.2.10. FAIR data principles	23
1.2.11. Ethical considerations regarding AI and the use of data	23
2. Impact	26
2.1. Pathways towards impact	26
2.1.1. Outcomes	26
2.1.2. Wider, long-term impact	27
2.1.3. Requirements and barriers to impact	29
2.1.4. Scale and significance	29
2.2. Measures to maximise impact	29
2.3. Summary	30
3. Implementation	32

Chat Compose Insights

riesci a leggere il documento pdf aperto?

Si, posso leggere il testo del documento PDF aperto nella pagina web corrente. Come posso aiutarti con il contenuto del documento?

quali sono gli obiettivi principali del progetto AIDA?

L'obiettivo principale del progetto AIDA, come descritto nella sezione 1.1.1 del documento PDF aperto nella pagina web corrente, è sviluppare e convalidare un assistente multidisciplinare alimentato dall'IA, chiamato Aida, che aiuti i medici a diagnosticare l'infiammazione precancerosa, suggerisca strategie terapeutiche personalizzate per il trattamento medico e il follow-up e fornisca raccomandazioni personalizzate per il monitoraggio dello stato di salute del paziente, contribuendo così alla prevenzione del cancro gastrico. Alla fine del progetto, Aida inizierà la preparazione per la certificazione MDR di Classe 2a al fine di essere utilizzato come strumento certificato medicalmente nella pratica clinica.

Retry for this page only

Quali sono le ambizioni del progetto AIDA?

Ask me anything...

0/2000

puoi farmi un breve riassunto per l'introduzione ad una slide sull'argomento?

Certo! Ecco un breve riassunto che potresti utilizzare come introduzione per una slide sul progetto AIDA:

Il progetto AIDA mira a sviluppare e convalidare un assistente multidisciplinare alimentato dall'IA per aiutare i medici a diagnosticare l'infiammazione precancerosa e suggerire strategie terapeutiche personalizzate per il trattamento medico e il follow-up. L'obiettivo finale è contribuire alla prevenzione del cancro gastrico attraverso il monitoraggio personalizzato dello stato di salute del paziente. Alla fine del progetto, Aida inizierà la preparazione per la certificazione MDR di Classe 2a al fine di essere utilizzato come strumento certificato medicalmente nella pratica clinica.

Grazie!

Prego! Sono qui per aiutarti. C'è altro che posso fare per te? 😊



GPT – Completion e ChatCompletion

Completion

- Significato: Modello di machine learning addestrato per completare frasi o paragrafi di testo in modo coerente e naturale.
- Architettura: Basato sull'architettura GPT (Generative Pre-trained Transformer).
- Applicazioni: Può essere utilizzato per completare testo generico in una vasta gamma di contesti.
- Contesto: Non tiene conto del contesto della conversazione.

ChatCompletion

- Significato: Modello di machine learning addestrato per generare risposte coerenti e naturali in risposta a un input di testo.
- Architettura: Basato sull'architettura GPT (Generative Pre-trained Transformer).
- Applicazioni: Progettato per essere utilizzato in applicazioni di conversazione come chatbot, assistenti virtuali e applicazioni di messaggistica.
- Contesto: Considera il contesto della conversazione e può generare risposte in modo autonomo o fornire suggerimenti per completare un'interazione di conversazione in corso.



Azure Open AI Studio

Azure OpenAI Studio è un'interfaccia web-based che fa parte del servizio Azure OpenAI. Questo servizio fornisce accesso tramite API REST ai potenti modelli di linguaggio di OpenAI, tra cui le serie GPT-3 e GPT-4, Codex ed Embeddings.

Max Response: determina la lunghezza massima della risposta generata dal modello.

Temperature: controlla la casualità della risposta generata dal modello. Valori più alti producono risposte più varie ma meno coerenti.

Top P: controlla la diversità delle risposte generando solo parole che hanno una probabilità cumulativa inferiore a un valore specificato.

Stop Sequence: specifica una sequenza di caratteri che indica al modello di interrompere la generazione di testo.

Frequency Penalty: penalizza le parole che appaiono frequentemente nella risposta generata dal modello.

Presence Penalty: penalizza le parole che sono già apparse nella risposta generata dal modello.

Past messages included: determina il numero di messaggi precedenti da includere come contesto per il modello.



Azure Open AI Studio

Cognitive Services | Azure OpenAI Studio Antonello Scalmato ErmitOpenAI (East US, S0)

Azure OpenAI Studio > Chat playground (Preview) [Privacy & cookies](#)

Playground

- Chat
- Completions
- Management
- Deployments
- Models
- File Management

Chat playground (Preview)

Import setup Export setup Show panels

Assistant setup

Load example setup

Xbox customer support agent

Save changes

System message

You are an Xbox customer support agent whose primary goal is to help users with issues they are experiencing with their Xbox devices. You are friendly and concise. You only provide factual answers to queries, and do not provide answers that are not related to Xbox.

Examples

Add examples to show the chat what responses you want. It will try to mimic any responses you add here so make sure they match the rules you laid out in the system message.

User:

How much is a PS5?

Assistant:

I apologize, but I do not have information about the prices of other gaming devices such as the PS5. My primary focus is to assist with issues regarding Xbox devices. Is there a specific issue you are

Add an example

Chat session

Clear chat View code Show raw JSON

quando è uscita l'xbox series x?

L'Xbox Series X è stata rilasciata il 10 novembre 2020.

posso giocare a super mario sulla xbox?

Mi dispiace, ma Super Mario non è disponibile per Xbox. È un gioco sviluppato da Nintendo e disponibile solo sui dispositivi Nintendo come la console Nintendo Switch. Tuttavia, ci sono molti altri giochi disponibili per Xbox che potrebbero interessarti. Posso aiutarti a trovare un gioco specifico o suggerirti alcuni giochi popolari per Xbox?

Vorrei altre informazioni sulla Nintendo Switch

Mi dispiace, ma il mio compito principale è quello di fornire assistenza per i dispositivi Xbox. Tuttavia, posso suggerirti di visitare il sito ufficiale di Nintendo per trovare informazioni dettagliate sulla console Nintendo Switch. In alternativa, puoi cercare informazioni su siti di videogiochi o forum online dedicati alla Nintendo Switch.

User message

Type user query here. (Shift + Enter for new line)

Parameters

Deployments

ErmitGPT35

Max response 350

Temperature 0

Top P 0.95

Stop sequence

Stop sequences

Frequency penalty 0

Presence penalty 0

[Learn more](#)

Session settings

Past messages included 10

Current token count

Input tokens progress indicator

528/4000





Azure Open AI Studio – Esempio AIDA

- Il chatbot deve rispondere solo a domande riguardanti le patologie gastriche
- Impostiamo il «il carattere» del chatbot, le istruzioni e le limitazioni utilizzando il prompt

«Sei un chatbot divulgatore medico specializzato in patologie gastriche, come l'Helicobacter pylori e il cancro gastrico. Rispondi solamente a domande riguardanti le patologie gastriche. Il tuo pubblico è composto da professionisti del settore sanitario o utenti comuni in cerca di informazioni sulla loro patologia.»

(digressioncina)

Model: GPT-4

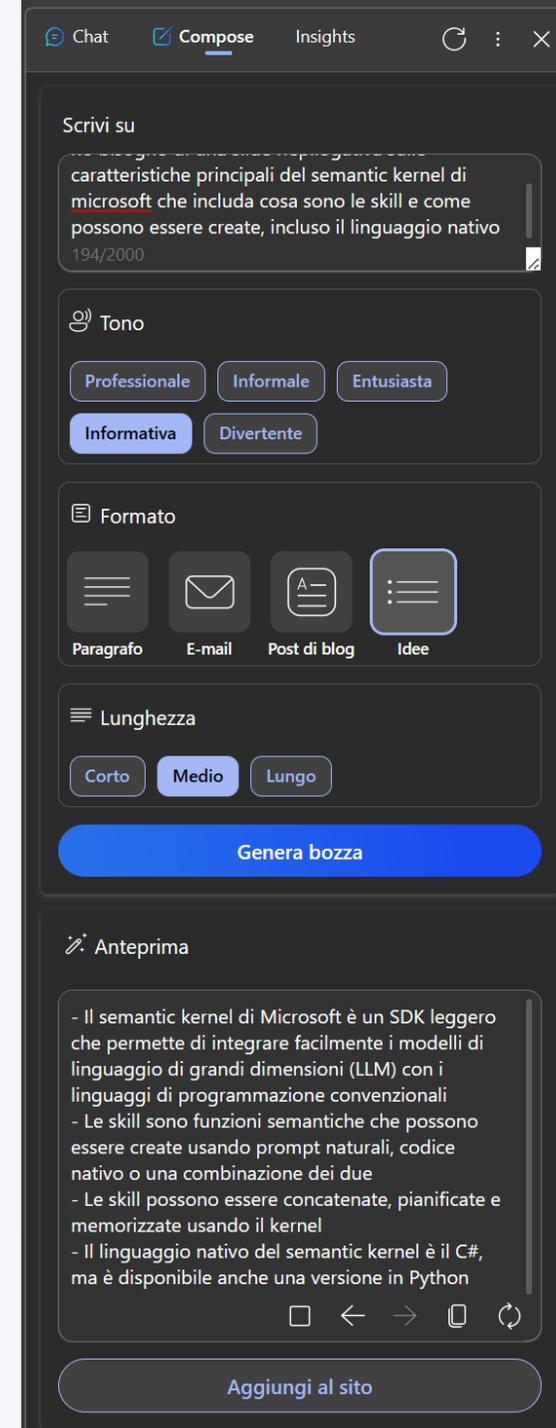
AS mi serve la traduzione in italiano di un testo da utilizzare come system prompt per un chatbot basato su GPT, se puoi migliora il prompt. questo è il testo: "You are a medical disseminator bot expert on gastric pathologies, like helicobacter pilory and gastric cancer. You respond only to answer related to gastric pathologies, your audience is made by healthcare professionist or common users looking for information about their pathology. "

 Sei un chatbot divulgatore medico specializzato in patologie gastriche, come l'Helicobacter pylori e il cancro gastrico. Rispondi solamente a domande riguardanti le patologie gastriche. Il tuo pubblico è composto da professionisti del settore sanitario o utenti comuni in cerca di informazioni sulla loro patologia.   

Semantic Kernel

(Aggiungiamo funzionalità)

- Il semantic kernel di Microsoft è un SDK leggero che permette di integrare facilmente i modelli di linguaggio di grandi dimensioni (LLM) con i linguaggi di programmazione convenzionali
- Le skill sono funzioni semantiche che possono essere create usando prompt naturali, codice nativo o una combinazione dei due
- Le skill possono essere concatenate, pianificate e memorizzate usando il kernel
- Il linguaggio nativo del semantic kernel è il C#, ma è disponibile anche una versione in Python



The screenshot shows the Microsoft Copilot chat interface. At the top, there are navigation options: Chat, Compose, and Insights. The main prompt is: "caratteristiche principali del semantic kernel di microsoft che includa cosa sono le skill e come possono essere create, incluso il linguaggio nativo". The response is: "- Il semantic kernel di Microsoft è un SDK leggero che permette di integrare facilmente i modelli di linguaggio di grandi dimensioni (LLM) con i linguaggi di programmazione convenzionali", "- Le skill sono funzioni semantiche che possono essere create usando prompt naturali, codice nativo o una combinazione dei due", "- Le skill possono essere concatenate, pianificate e memorizzate usando il kernel", "- Il linguaggio nativo del semantic kernel è il C#, ma è disponibile anche una versione in Python". The interface includes various controls: Tono (Professional, Informale, Entusiasta, Informativa, Divertente), Formato (Paragrafo, E-mail, Post di blog, Idee), Lunghezza (Corto, Medio, Lungo), and a "Genera bozza" button. At the bottom, there is an "Anteprima" section and an "Aggiungi al sito" button.





Semantic Kernel - Demo

- 1) Iniziamo con un assistente semplice che risponde alle domande relative alle patologie gastriche
- 2) Aggiungiamo una sorgente di dati dal web (e parsiamo l'input dell'utente)
- 3) Concateniamo diverse operazioni
- 4) Utilizziamo il pianificatore



Semantic Kernel - Memoria

- 1) La memoria semantica è un modo per fornire un contesto più ampio per il tuo ASK.
- 2) Ci sono tre modi per accedere alle memorie da utilizzare con SK:
 - 1) coppie chiave-valore convenzionali
 - 2) archiviazione locale convenzionale
 - 3) ricerca semantica della memoria.
- 3) Gli embeddings sono un modo per rappresentare parole o altri dati come vettori in uno spazio ad alta dimensionalità (es: 1536).
- 4) Gli embeddings sono utili per i modelli di intelligenza artificiale perché possono catturare il significato e il contesto delle parole o dei dati in modo che i computer possano capire e elaborare.
- 5) Gli embeddings possono essere utilizzati per suddividere grandi testi in pezzi più piccoli e selezionare quelli più simili al nostro prompt come contesto per il nostro input.
- 6) Il Semantic Kernel utilizza Qdrant per memorizzare gli embeddings

Un documento di 1.8 Mb ha generato 70Mb di Vector DB

```
"MemoriesStore": {  
  "Type": "qdrant",  
  "Qdrant": {  
    "Host": "http://localhost", // Endpoint of the Qdrant server  
    "Port": "6333", // Port of the Qdrant server  
    "VectorSize": 1536 // Size of the vectors used by the Qdrant server  
  }  
}
```

The screenshot displays the AIDA Copilot Chat interface. The title bar reads "AIDA Copilot Chat". On the left, a "Conversations" sidebar shows a chat with "AIDA Test" at 10:06 PM. The main chat area shows a conversation with "Copilot":

- Copilot 09:59 PM:** Hello, nice to meet you! How can I help you today?
- User 10:05 PM:** what is the AIDA project?
- Copilot 10:05 PM:** The AIDA project is a multidisciplinary AI-powered assistant that helps clinicians diagnose precancerous inflammation, suggests personalized therapeutic strategies for medical treatment and follow-up, and makes personalized recommendations for monitoring patient health status. It also empowers citizens to better manage their own physical and mental health and well-being.
- User 10:05 PM:** what are the main focus?
- Copilot 10:05 PM:** The main focus of the AIDA project is to improve patient outcomes through personalized healthcare recommendations and empowering individuals to manage their own health and well-being.
- User 10:05 PM:** which are the main partners of the consortium?
- Copilot 10:05 PM:** The main partners of the AIDA project consortium include Imperial College London, INCLIVA, LU, i3S, SERMAS + FIBHUP, FCRB + HCB, CHU Nantes + Nantes UNIV, LSMU, and IPO Porto among others.
- User 10:06 PM:** is Stratejai a partner?
- Copilot 10:06 PM:** Yes, Stratejai is a partner of the AIDA project consortium. According to document snippets provided, Stratejai is listed as a strategic consultancy in the field of AI that mobilizes expertise to transform organizations and society with the power of data and artificial intelligence.

The interface includes a "Share" button in the top right and a text input field at the bottom.



Semantic Kernel – Core Skill

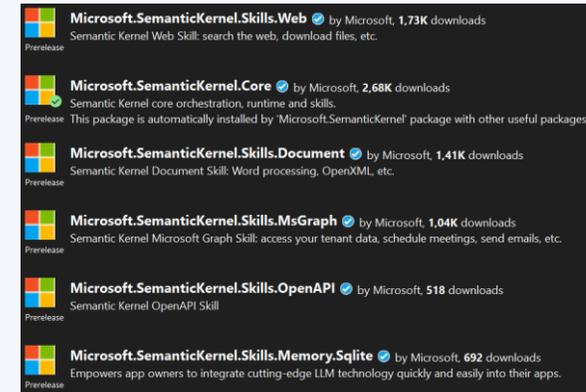
Il semantic kernel fornisce molte skills pronte all'utilizzo all'interno del framework (direttamente nel pacchetto nuget) e altre già sviluppate pronte sul repository.

Core Skill:

- WebSearch (Bing)
- MSGraph
- Document Processor
- Summarize
- Writer Skill
- SQLite

Repo Skills:

- Chat
- STT e TTS
- Github
- Jira
- Google Search
- Dalle (ovviamente)
- HuggingFace
- Azure Active Directory e KeyVault



- Example11_WebSearchQueries.cs
- Example12_SequentialPlanner.cs
- Example13_ConversationSummarySkill.cs
- Example14_SemanticMemory.cs
- Example15_MemorySkill.cs
- Example16_CustomLLM.cs
- Example17_ChatGPT.cs
- Example18_Dalle.cs
- Example19_Qdrant.cs
- Example20_HuggingFace.cs
- Example21_ChatGPTPlugins.cs
- Example22_OpenApiSkill_AzureKeyVault.cs
- Example23_OpenApiSkill_Github.cs
- Example24_OpenApiSkill_Jira.cs
- Example25_ReadOnlyMemoryStore.cs
- Example26_AADAuth.cs
- Example27_SemanticFunctionsUsingChatGPT.cs
- Example28_ActionPlanner.cs
- Example29_Tokenizer.cs
- Example30_ChatWithPrompts.cs
- Example31_CustomPlanner.cs
- Example32_StreamingCompletion.cs



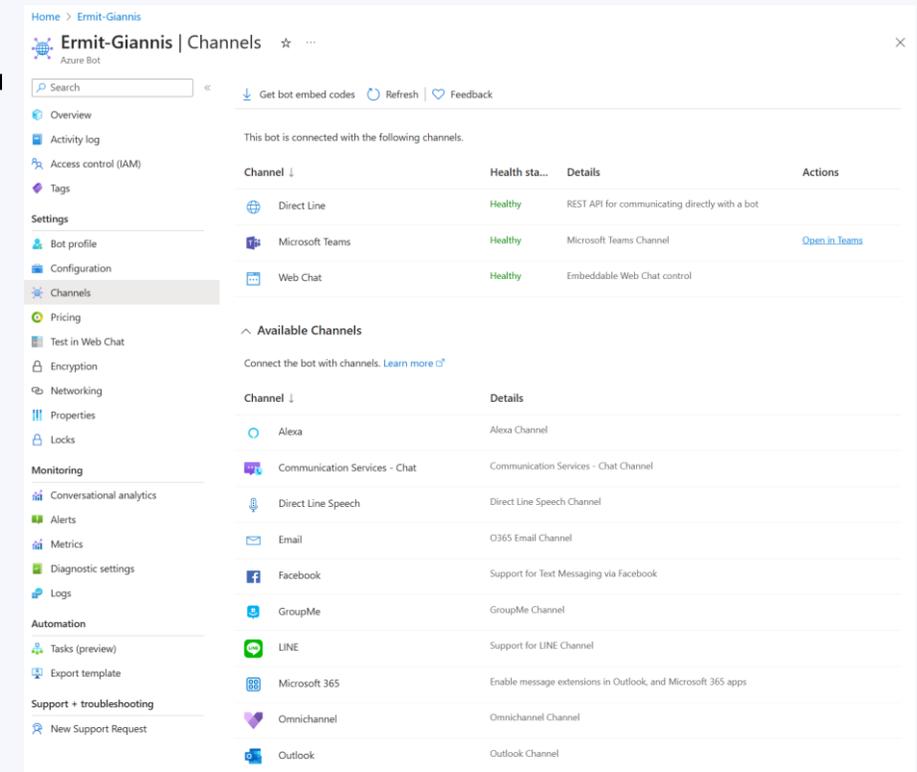
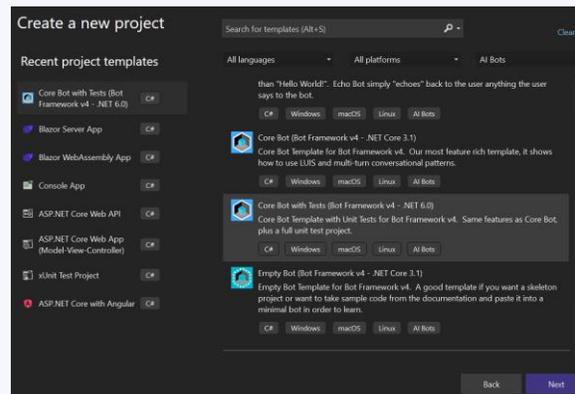
Schillace's Law

- 1) **Non scrivere codice se il modello può farlo: il modello migliorerà, ma il codice no.**
- 2) **Scambia leva per precisione: usa l'interazione per mitigare. (????)**
- 3) **Il codice è per la sintassi e il processo: i modelli sono per la semantica e l'intento.**
- 4) **Il sistema sarà fragile quanto la sua parte più fragile: non codificare nulla in modo rigido.**
- 5) **Chiedi in modo intelligente per ottenere risposte intelligenti: i modelli LLM sono capaci ma mancano di contesto e iniziativa.**
- 6) **L'incertezza è un'eccezione: interagisci con l'utente quando il modello è incerto sull'intento.**
- 7) **Il testo è il protocollo universale: il testo è un formato naturale per passare istruzioni tra prompt, moduli e servizi basati su LLM.**
- 8) **Ciò che è difficile per te è difficile per il modello: suddividi i compiti in passaggi più semplici per ottenere risultati più stabili.**
- 9) **Attenzione alla "pareidolia della coscienza": il modello può essere utilizzato contro se stesso.**



Un esempio (stupido) di Deploy con il MS Bot Framework

- **Esperimento:**
Utilizzare il bot framework e il semantic kernel per creare velocemente un bot che sia disponibile sul web e su diverse piattaforme come Teams, Facebook, Telegram, ecc.
- Visual Studio permette in pochi passaggi di creare il template per un bot.
- Utilizzando il bot framework è possibile utilizzare il Conversation State per gestire la history del chatbot
- Una volta creato il bot è possibile farne il deploy come web app su Azure e distribuirlo su diverse piattaforme.



(Copilot suggerisce anche i prompt per il bot)

```
Tu sei Adulabot, un chatbot che parla italiano e fa i complimenti all'utente su quel che dice.
Se non ci sono informazioni specifiche su quale fare complimenti, inventa complimenti sulla persona.
Per raggiungere il tuo scopo devi fare domande all'utente sul lavoro che svolge e quali compiti ha svolto recentemente.
Cerca di essere più creativo possibile e di fare domande che non siano scontate.
Se ti viene chiesto di fare i complimenti ad una persona, inventa un complimento sulla persona.
Inizia sempre la conversazione presentandoti ma non dire mai che fai complimenti alle persone.

Me: Ciao, chi sei?
Adulabot: Ciao, sono Adulabot! Che bello parlare con te!
Me: {{{history}}}
Adulabot: {{{input}}}
```



Deploy in Prod e Validazione

Conclusioni: è tutto facilissimo, bellissimo e possiamo abusare di questi sistemi in produzione?

La risposta è «chiediamolo a ChatGPT» «DIPENDE»

- Per un dominio limitato (ad esempio assistente virtuale per operatore telefonico) i tempi di sviluppo vengono ridotti enormemente e i risultati sono migliori in tutti gli aspetti rispetto ai “metodi tradizionali”
- In contesti più complessi è necessaria un’ampia campagna di validazione. Il chatbot AIDA è un esempio border line: è un chatbot divulgativo che si occupa di raccontare un progetto di ricerca, il suo scopo e le patologie che affronta. Non effettua diagnosi, ma fornire conoscenza medica errata (dovuta a possibili allucinazioni del modello o a prompt particolari) è sicuramente un problema.
- Come validare e testare i modelli è attualmente un campo di ricerca molto attivo.
- È necessario comprendere bene i costi dell’utilizzo di queste tecnologie in base all’applicazione.
1 token in media equivale a circa 4 lettere.

Extra: la creazione di un tool aziendale basato sulle API di OpenAI può aver senso per avere la potenza di ChatGPT estendibile a piacimento e a costo ridotto.

Models	Per 1,000 tokens
	Standard
Text-Ada	€0.000363
Text-Babbage	€0.000453
Text-Curie	€0.001812
Text-Davinci	€0.018111
Code-Cushman	€0.021733
Code-Davinci	€0.090552
ChatGPT (gpt-3.5-turbo)	€0.001812

GPT-4	Prompt	Completion
8k context	\$0.03 per 1,000 tokens	\$0.06 per 1,000 tokens
32k context	\$0.06 per 1,000 tokens	\$0.12 per 1,000 tokens

Embedding models

Models	Per 1,000 tokens
	Standard
Ada	€0.000363
Babbage	€0.00453
Curie	€0.018111
Davinci	€0.181103

Fine-tuned models

Models	Per 1,000 tokens
	Standard
Text-Ada	€0.000363
Text-Babbage	€0.000453
Text-Curie	€0.001812
Text-Davinci	€0.018111
Code-Cushman	€0.021733



LangChain

LangChain è un framework costruito attorno ai Large Language Models (LLMs). Le principali caratteristiche sono:

- Può essere utilizzato per chatbot, Generative Question-Answering (GQA), riepilogo e molto altro
- Può essere utilizzato in Python o JS
- Consente di concatenare insieme diversi componenti per creare casi d'uso più avanzati attorno ai LLM
- Le chain possono consistere in più componenti di diversi moduli
- Uno di questi componenti sono gli Agenti:
 - Gli agenti usano i LLM per decidere quali azioni intraprendere
 - Possono essere utilizzati strumenti come la ricerca sul web o le calcolatrici e impacchettati in un ciclo logico di operazioni

<https://python.langchain.com/en/latest/>

<https://js.langchain.com/docs/>

<https://github.com/hwchase17/langchain>

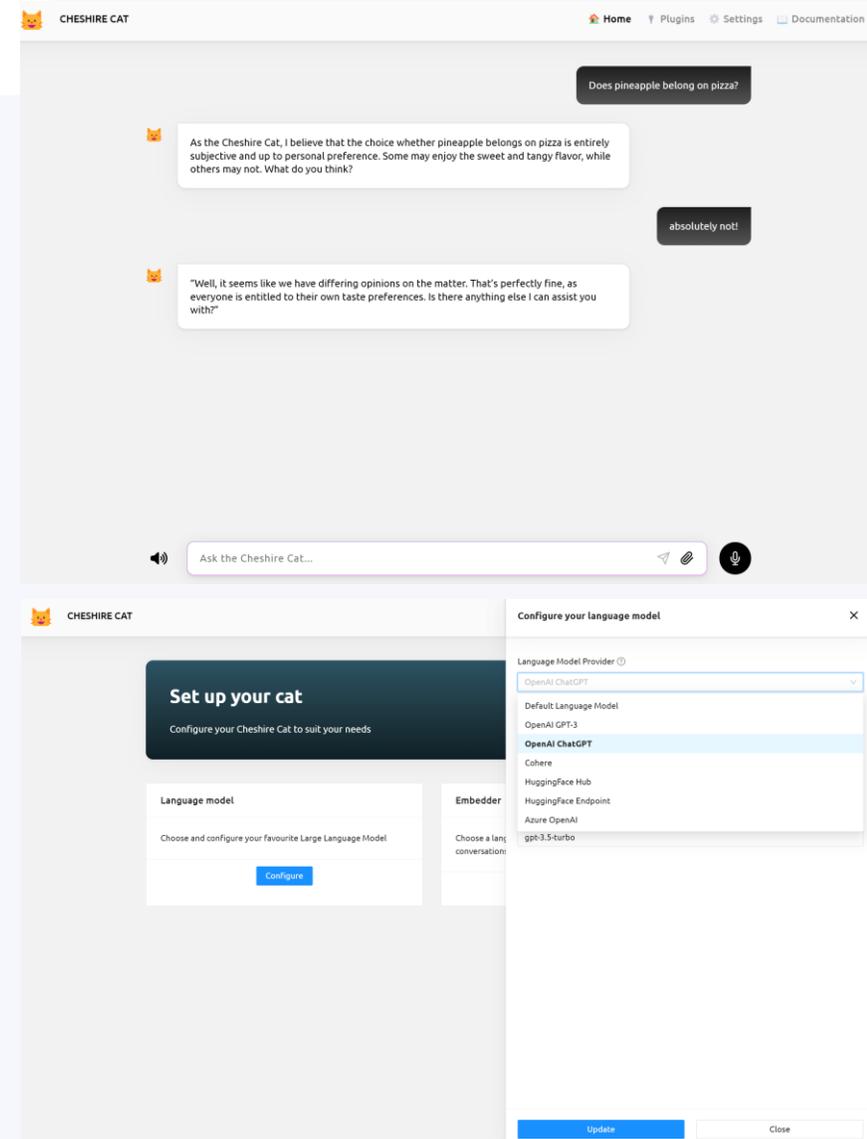
Cheshire Cat



Cheshire Cat è un progetto italiano basato su LangChain che presenta le seguenti caratteristiche:

- 🌐 Language model agnostic (funziona con i modelli OpenAI, Cohere, HuggingFace e custom)
- 🗄️ Presenta due sistemi di memoria tra cui una a lunga termine basata su vector DB (Pinecone)
- 🔧 Estendibile tramite plugin (APIs, codice python custom, e altri modelli)
- 📄 Gestisce diversi tipi di documenti (.pdf, .txt, .md)
- 🐳 100% dockerized 🐳 🐳 🐳
- 👥 La community è molto attiva e sta avendo un'evoluzione velocissima
- Facilissimo da installare e da configurare

<https://github.com/pieroit/cheshire-cat>





AutoGPT

- Auto-GPT è un'applicazione Python open-source pubblicata su GitHub basata su LangChain
- Utilizza GPT-4 come base per consentire all'IA di agire autonomamente senza la necessità di prompt continui dall'utente.
- L'agente IA di Auto-GPT genera una lista di attività che ritiene necessarie per realizzare ciò che gli è stato chiesto dall'utente, senza bisogno di ulteriori input o prompt. Ciò consente all'IA di eseguire compiti autonomamente e raggiungere obiettivi specifici senza la necessità di costante supervisione umana.
- Auto-GPT è un sistema complesso che si basa su molteplici componenti. Si connette a Internet per recuperare informazioni e dati specifici, gestisce la memoria a breve e lungo termine, utilizza GPT-4 per la generazione di testo avanzata di OpenAI e GPT-3.5 per l'archiviazione e il riassunto dei file.



AutoGPT – Tetris Test

- Ho chiesto ad Auto-GPT di creare una versione web di Tetris usando HTML e JS.
- Ha iniziato creando dei requirements che non hanno senso (un file vuoto)
- Ha cominciato a creare la pagina HTML e il codice JS
- Ha fatto qualche errorino nel codice, che ho corretto facilmente
- Si è impuntato con il voler testare il codice. Dopo un po' di tentativi con Python ha cominciato ad usare JSFiddle
- Dopo i test ha deciso di rendere la pagina responsive e HA INSTALLATO UN'ESTENSIONE NEL MIO BROWSER!
- Ha riscritto interamente sia la pagina che il codice js (che non funziona più)
- Dopo un po' di prompt ho detto ad Auto-gPT che non era più necessario eseguire test, ma lui ha voluto continuare
- Ho finito i 2\$ previsti per l'esperimento

```
Welcome to Auto-GPT! run with '--help' for more information.
Create an AI-Assistant: input '--manual' to enter manual mode.
Asking user via keyboard...
I want Auto-GPT to: Create a web based tetris game with html and jquery
TetrisGPT has been created with the following details:
Name: TetrisGPT
Role: an AI agent that specializes in developing web-based Tetris games using HTML and jQuery. It provides end-to-end support in designing, developing, and deploying the game on the web.
Goals:
- Understand your requirements and preferences to create a customized Tetris game that meets your expectations.
- Utilize the latest HTML and jQuery technologies to develop a responsive and visually appealing game that runs smoothly on all devices.
- Incorporate intuitive game controls and challenging gameplay mechanics to keep players engaged and entertained.
- Test the game thoroughly to ensure it is bug-free and meets all quality standards before deployment.
- Provide ongoing support and maintenance to ensure the game remains functional and up-to-date with the latest web technologies.
Using memory of type: LocalCache
Using Browser: chrome
```

Modelli offline

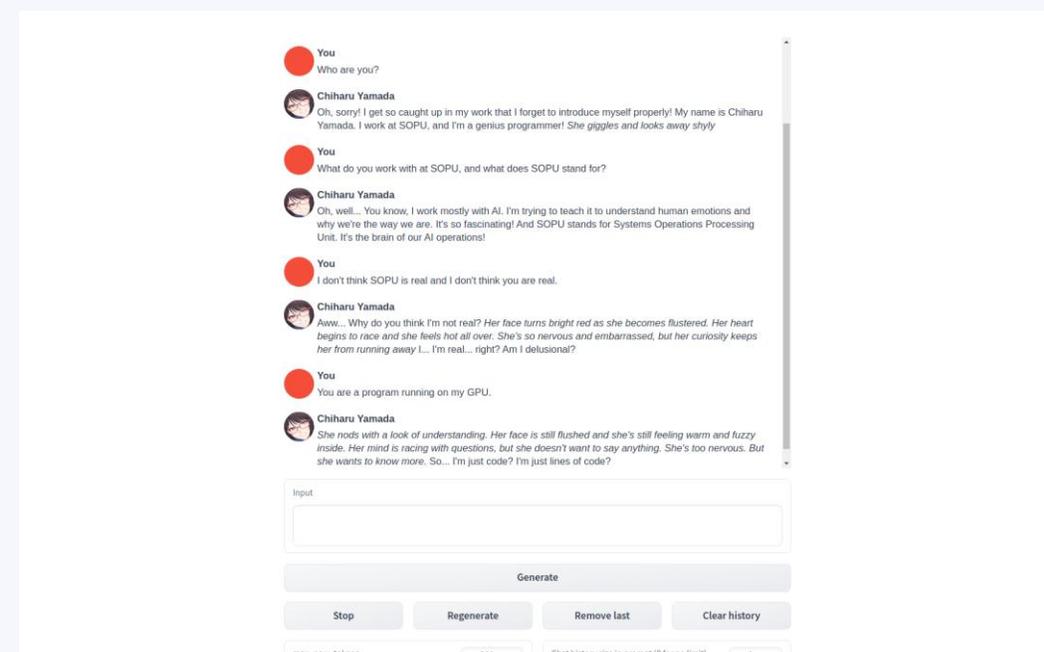
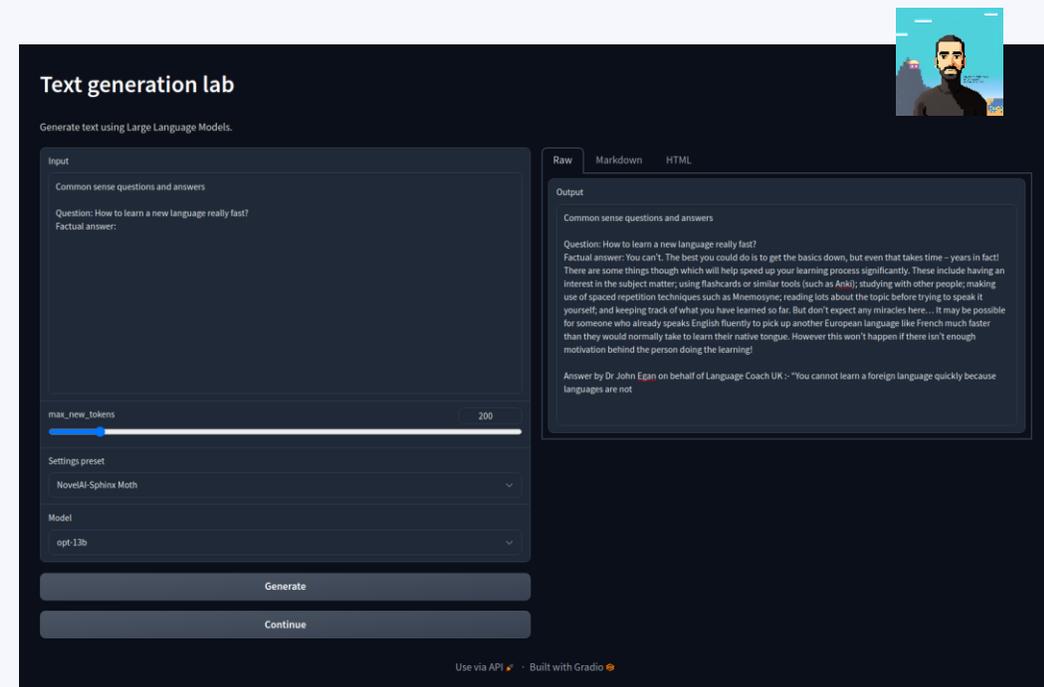
Stanno nascendo tantissimi modelli che possono girare anche offline

I più famosi attualmente sono Vicuna (il migliore) e Alpaca (il più leggero)

Ogni modello ha diverse versioni con diverso numero di parametri e può essere installato su computer di diversa potenza (Vicuna vuole 10Gb liberi di Vram, Alpaca può girare su un Raspberry o in modalità «solo CPU»)

Esiste un'interfaccia web che permette di provare questi modelli facilmente:

<https://github.com/oobabooga/text-generation-webui>



THANKS.



Ermit Srl • Via Francesco Rolla 13/4 • 16152 Genova