S2E
business & technology consultants

INTELLIGENZA ARTIFICIALE GENERATIVA

DanteGPT:
l'AI che risponde in sonetti

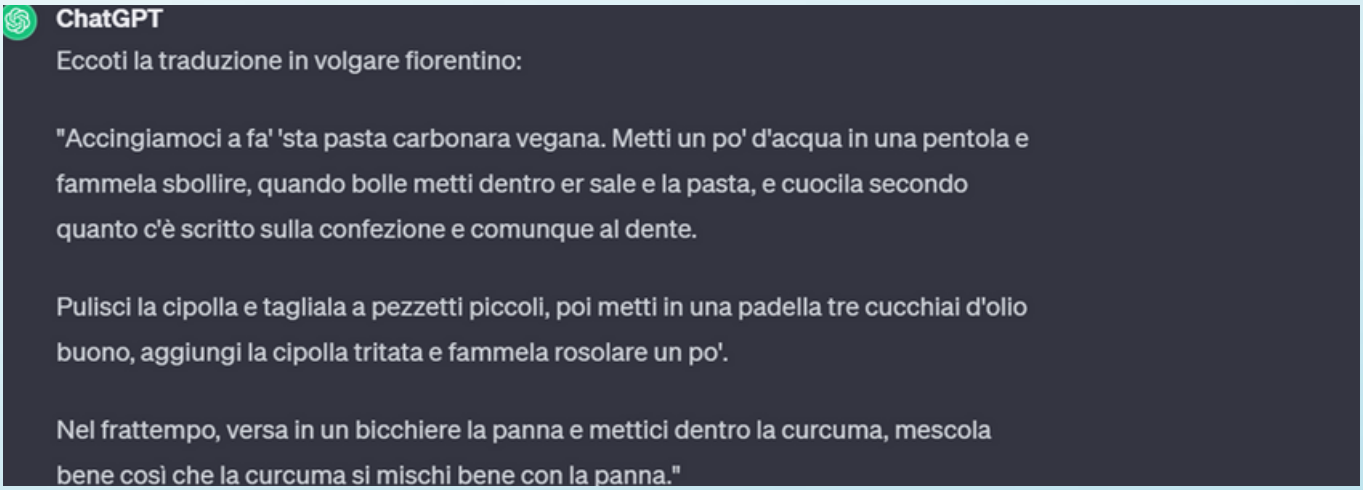
Chi non ha mai desiderato dialogare con una figura leggendaria del nostro passato culturale?
Noi, team di Advanced Analytics di QuantiaS, business unit di S2E, abbiamo trasformato questa curiosità in una sfida: quella di ricostruire, attraverso l'intelligenza artificiale, il dialogo con uno dei giganti della letteratura mondiale, Dante Alighieri. Proponendo un'interazione che unisce il fascino dell'antico 'bel parlare' alla potenza tecnologica moderna, "**DanteGPT**" emerge come un esempio straordinario di come possiamo, nell'era digitale, ridare nuova vita al passato.

S2E
business & technology consultants

DanteGPT:
l'AI che risponde in sonetti

INTELLIGENZA ARTIFICIALE GENERATIVA

Ogni giorno, assistiamo alla pubblicazione di numerosi articoli relativi alla cosiddetta RAG (Retrieval Augmented Generation), che permette di nutrire ChatGPT con i propri dati aziendali. La RAG è molto potente ed è alla base di un nostro prodotto aziendale di successo di nome Generative Shield [1]. Oggi, invece, vogliamo darti un input su come accedere ai meccanismi interni del modello LLM (Large Language Model) e personalizzarli sulla base delle tue esigenze.
I modelli commerciali come GPT-4 sono dei black-box che riescono a soddisfare nativamente molteplici necessità aziendali, tuttavia esistono degli use-case che possono essere risolti unicamente addentrandosi nelle viscere della rete neurale!
Hai mai provato a far parlare GPT-4 come se fosse Dante Alighieri, ad esempio chiedendogli di tradurre in volgare fiorentino la ricetta della carbonara vegana? Noi sì e i risultati sono lontani dall'essere ottimali, come puoi constatare nel box che segue.

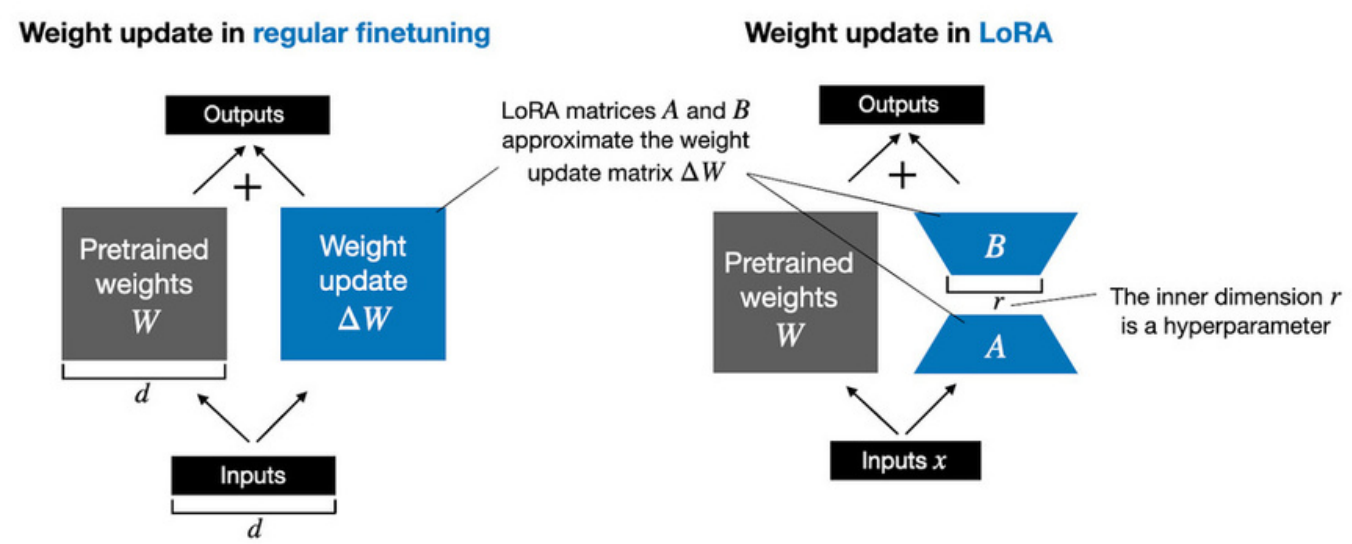


Nonostante i modelli di Open AI abbiano mostrato impressionanti capacità di elaborazione e generazione di testo in inglese moderno, essi incontrano limiti notevoli quando sollecitati a trattare lingue e contesti storici non contemporanei.
"DanteGPT" si presenta come un semplice esperimento specificamente concepito per addestrare dei piccoli modelli, diversi da Open AI, al riconoscimento e alla generazione di linguaggi caratteristici di epoche e stili non immediatamente accessibili attraverso modelli commerciali.

Come fare quindi ad insegnare a un piccolo LLM a rispondere ad una richiesta come fosse Dante Alighieri?

FINE-TUNING CON LA QLoRA

Innanzitutto scegliamo un *foundational model* (il modello di partenza), come **Saiga-7b**, un modello open-source già allenato (pre-trained) specializzato nella lingua italiana. Il primo passo da compiere è il fine-tuning su un dataset contenente l'intera Divina Commedia. Il fine-tuning è il processo di ottimizzazione dei parametri del modello affinché svolga uno specifico task; nel nostro caso, lo scopo è quello di ottenere un chatbot che risponda nel dialetto fiorentino del 1300. Per questo motivo, i parametri del modello *pre-trained* verranno aggiornati e ottimizzati sulla base del dataset contenente l'intera Divina Commedia, che fornisce al modello una lunga serie di esempi su come tradurre il linguaggio dall'italiano al dialetto fiorentino. In genere, il fine-tuning completo dei parametri è un processo oneroso, per questo vi mostreremo come applicare una veloce personalizzazione del modello, senza dover riallenare tutti i suoi pesi.
Per far ciò, utilizziamo un metodo PEFT (Parameter Efficient Fine Tuning) chiamato LoRA (*Low-Rank Adaptation*), che permette di velocizzare il processo di ottimizzazione dei parametri: data una matrice W di dimensione d e k, la scomposizione fatta con la LoRA produrrà due matrici A(d x r) e B(r x k), dove r è l'iper-parametro rank.



Questa tecnica permette di velocizzare l'allenamento e renderlo più accessibile, andando di fatto ad alterare meno dell'1% dei parametri totali del modello.
Per ridurre ulteriormente l'intensità computazionale richiesta, utilizziamo la versione quantizzata della LoRA, la **QLoRA**. Durante il caricamento del modello, la QLoRA applica una compressione dei pesi a 8 bit floating point. Assieme, questi due passaggi facilitano l'implementazione e l'allenamento di modelli on-prem, permettendoci di ridurre le risorse computazionali necessarie consumando solo 15,7 GB di RAM della GPU.

APPLICAZIONE IN
DanteGPT

Per facilitare l'apprendimento, abbiamo strutturato il dataset di addestramento nel formato:

```
{"instruction":
"Traduci in volgare fiorentino:\n",
"input":
"Nel mezzo del cammin di nostra vita (=a 35 anni) mi ritrovai per una selva oscura, perchè avevo smarrito la retta via.",
"output":
"Nel mezzo del cammin di nostra vita\nmi ritrovai per una selva oscura\nche la diritta via era smarrita."}
```

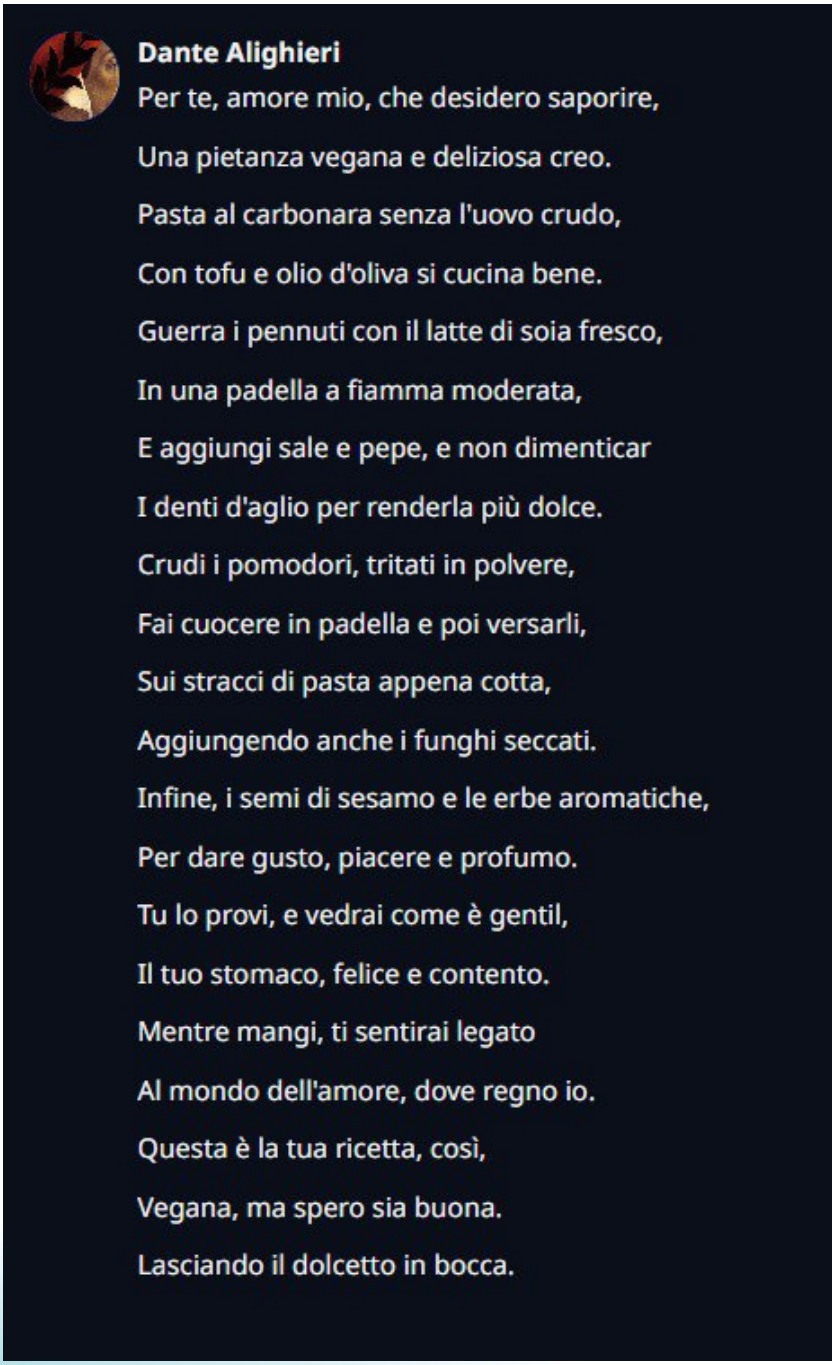
In questa maniera, insegniamo al modello a tradurre dall'italiano al volgare fiorentino, basandoci su un dataset composto da testo e parafrasi della Divina Commedia.
Lanciamo dunque la QLoRA con i parametri:

- Rank = 32
- Alpha = 64

La scelta dei due parametri è uno degli aspetti più critici e importanti del lavoro. In primo luogo, la selezione del rank determina la quantità di informazioni che vengono trasferite dalla matrice originale W alle due sotto-matrici A e B. Impostare un valore alto del rank ha il beneficio di aumentare la precisione dei risultati, ma può indurre all'overfitting e a un consumo di risorse computazionali troppo alto. In base ai nostri esperimenti, un rank di 32 porta a risultati ottimali.
D'altro canto, l'alpha è un coefficiente per la normalizzazione dei pesi. Dalla nostra esperienza, otteniamo risultati ottimali impostando l'alpha come 2*rank.
Infine, dopo aver fatto girare il modello per un'ora, possiamo confrontare i risultati con ChatGPT.

Ripetendo la domanda inziale posta a ChatGPT "Mi dai la ricetta della carbonara vegana" DanteGPT ha risposto nel modo presentato nella pagina seguente.

APPLICAZIONE IN
DanteGPT



Come possiamo vedere, la risposta ricalca lo stile di sonetto di Dante Alighieri pur mantenendo le capacità assistente di Saiga-7b.
Puoi trovare le risorse utilizzate, i modelli e dataset nel seguenti link.

Per qualsiasi domanda o dubbio non esitare a contattare l'autore di questo articolo:

Doina Vasilev

Sarà felice di rispondere, insieme al team di Advanced Analytics di QuantiaS, alle tue domande e aiutarti a riprodurre questo esperimento!

DanteGPT:
l'AI che risponde in sonetti

INTELLIGENZA ARTIFICIALE GENERATIVA

SOLUZIONI E APPLICAZIONI

Nonostante questo esperimento sia nato con l'intento di esplorare le potenzialità delle tecnologie AI per scopi culturali e di ricerca, non ignoriamo le interessanti opportunità commerciali che possono derivarne.

Le metodologie innovative impiegate per creare DanteGPT, infatti, non consentono soltanto di emulare stili linguistici specifici di epoche passate o di autori celebri, ma forniscono anche gli strumenti per creare chatbot con personalità uniche, capaci di rendere le interazioni con i brand esperienze immersive, costruendo identità comunicative distintive e contribuendo a strutturare un dialogo che va oltre la semplice transazione.
Allo stesso tempo, le medesime tecniche ci consentono di creare chatbot altamente specializzati, capaci di navigare e utilizzare il lessico specifico di numerosi settori professionali, come quello legale, medico, amministrativo, e altri simili campi dove l'accuratezza della terminologia è imprescindibile.

Questa personalizzazione estrema può aprire nuovi orizzonti nel campo del marketing, dell'assistenza clienti, della creazione di contenuti e molto altro ancora. Esperimenti come DanteGPT offrono opportunità commerciali che possono trasformare il modo in cui le aziende interagiscono con il loro pubblico, promuovendo un coinvolgimento più profondo e stimolando un'esperienza di marca unica.

Inoltre, visitando il sito di S2E scopri tutte le nostre soluzioni AI, come Generative Shield [1], piattaforma SaaS che risiede nel cloud di AWS e che utilizza tutti servizi serverless; questo la rende altamente scalabile e, basandosi sul RAG, permette la creazione di agenti conversazionali altamente competenti.