

Design, ingegnerizzazione e realizzazione di un sistema di dialogo basato su LLM nel dominio delle tecnologie assistive

Tesi di Laurea Magistrale



Relatore: Prof. Alessandro Mazzei

Co-Relatori: Dott. Pier Felice Balestrucci, Dott. Michael Oliverio

Candidato: Dott. Stefano Vittorio Porta

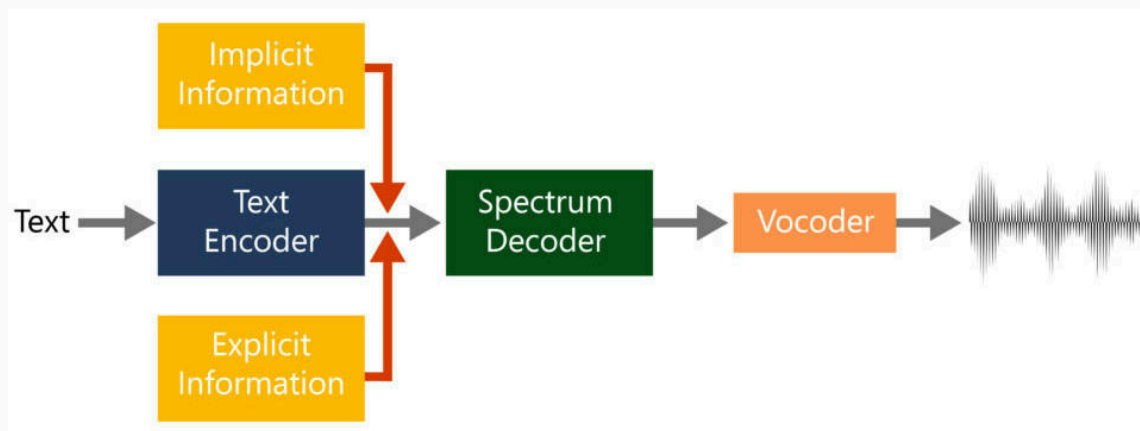
2 Aprile 2025

Università degli Studi di Torino, Dipartimento di Informatica - Anno Accademico 2023/2024

1. Contesto

Iniziamo ad ambientarci:

- **Lettura di contenuti testuali**: da 30 anni sono disponibili sistemi di sintesi vocale integrati in smartphone e computer.
- Essenziali per le persone con **disabilità visive**: consentono di accedere a contenuti testuali in modo autonomo e senza l'ausilio di un lettore umano.



- Pagine web. I sistemi TTS sono in grado di
 - Leggere il testo
 - Interpretare la struttura del documento
 - Questo permette di seguire il flusso di lettura per fornire un'esperienza di qualità

1.2 Problema!



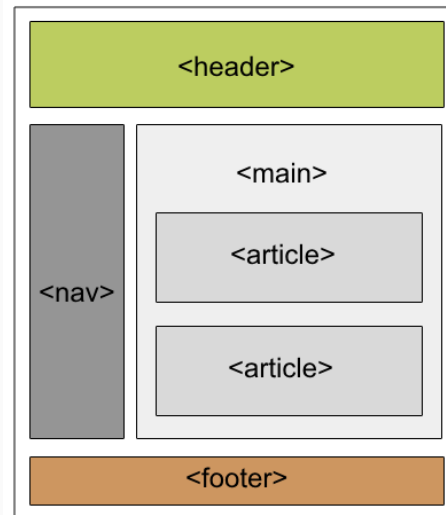
Tutto funziona, a patto che...

1.2 Problema!



Tutto funziona, a patto che...

- La pagina web sia strutturata in modo semantico

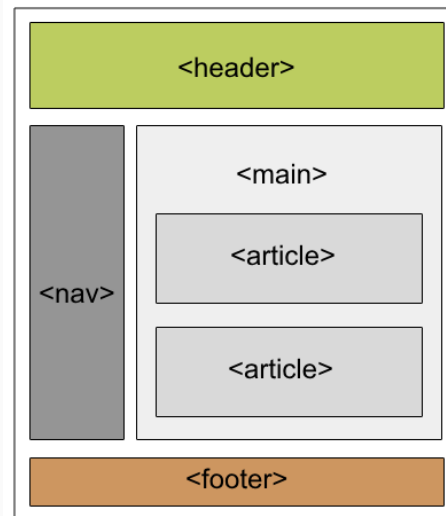


1.2 Problema!



Tutto funziona, a patto che...

- La pagina web sia strutturata in modo semantico



- Le immagini siano accompagnate da un testo alternativo

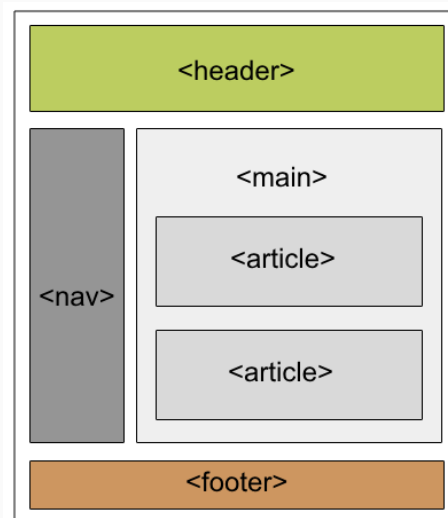
```
<img alt="..."> o <img aria-label="...">
```

1.2 Problema!



Tutto funziona, a patto che...

- La pagina web sia strutturata in modo semantico



- Le immagini siano accompagnate da un testo alternativo

`` o ``

Queste due condizioni dipendono da chi prepara il contenuto!

I sistemi di TTS non sono in grado di interpretare il significato di un'immagine o di un elemento puramente visivo.

^[1]WebAIM, «The WebAIM Million: The 2024 Report on the Accessibility of the Top 1,000,000 Home Pages». [Online]. Disponibile su: <https://webaim.org/projects/million/>

1.2 Problema!



I sistemi di TTS non sono in grado di interpretare il significato di un'immagine o di un elemento puramente visivo.



Se non viene fornita un'alternativa testuale contenente delle informazioni utili, l'utente non potrà comprendere appieno il contenuto della pagina o di uno specifico elemento!

^[1]WebAIM, «The WebAIM Million: The 2024 Report on the Accessibility of the Top 1,000,000 Home Pages». [Online]. Disponibile su: <https://webaim.org/projects/million/>

1.2 Problema!



I sistemi di TTS non sono in grado di interpretare il significato di un'immagine o di un elemento puramente visivo.



Se non viene fornita un'alternativa testuale contenente delle informazioni utili, l'utente non potrà comprendere appieno il contenuto della pagina o di uno specifico elemento!

Una di quattro immagini sul web non ha una descrizione testuale o non è informativa^[1].

^[1]WebAIM, «The WebAIM Million: The 2024 Report on the Accessibility of the Top 1,000,000 Home Pages». [Online]. Disponibile su: <https://webaim.org/projects/million/>

1.3 Anche peggio...

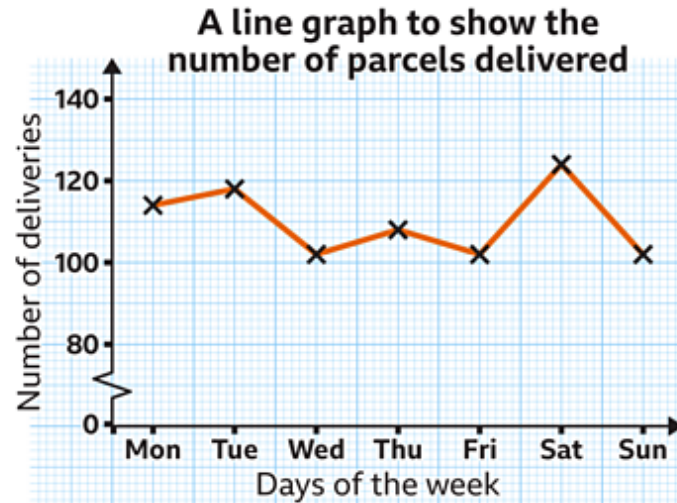


Se per le immagini possiamo utilizzare tecniche di **Computer Vision** o **Reti Neurali** per generare automaticamente un testo alternativo, per le rappresentazioni grafiche di dati (grafici, diagrammi, mappe) non è così semplice.

1.3 Anche peggio...



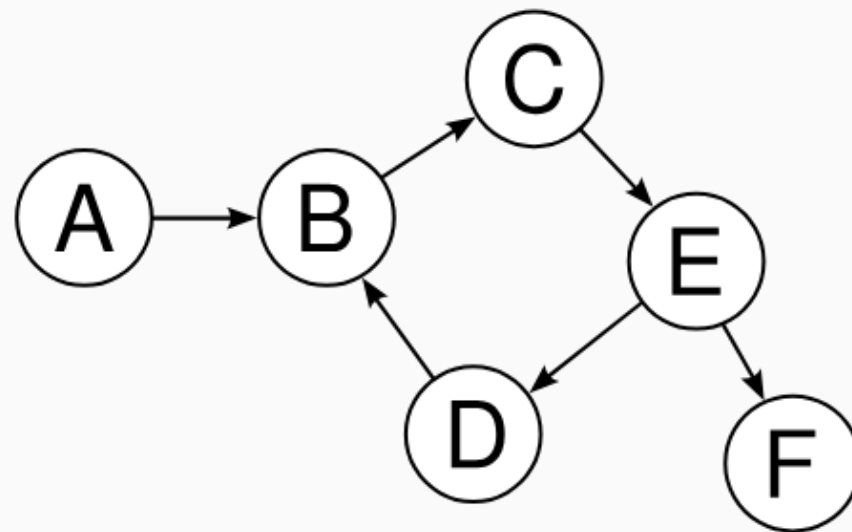
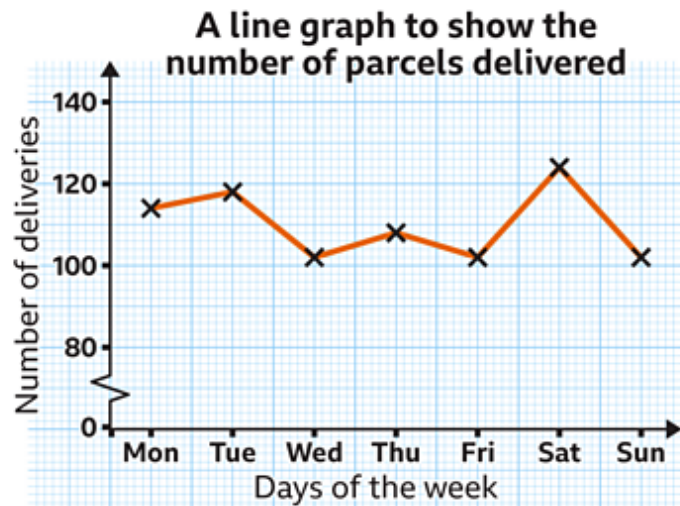
Se per le immagini possiamo utilizzare tecniche di **Computer Vision** o **Reti Neurali** per generare automaticamente un testo alternativo, per le rappresentazioni grafiche di dati (grafici, diagrammi, mappe) non è così semplice.



1.3 Anche peggio...



Se per le immagini possiamo utilizzare tecniche di **Computer Vision** o **Reti Neurali** per generare automaticamente un testo alternativo, per le rappresentazioni grafiche di dati (grafici, diagrammi, mappe) non è così semplice.



- Il Progetto NoVAGraphS si propone di rendere più accessibili questi contenuti, mediante la costruzione di sistemi di dialogo (Chatbot).
- Con essi è possibile interagire per ottenere informazioni sui dati presenti in grafi o strutture simili, per avere una comprensione **profonda** del contenuto.
- Il Progetto originale fa fondamento su AIML ^[1] (Artificial Intelligence Markup Language), un linguaggio di markup per la creazione di chatbot.

^[1]«Artificial Intelligence Markup Language». [Online]. Disponibile su: <http://www.aiml.foundation/doc.html>

```
<category>  
  <pattern>MI CHIAMO *</pattern>  
  <template>  
    Ciao <star/>, piacere di conoscerti!  
  </template>  
</category>
```

Codice 1: Esempio di AIML per la gestione di un saluto

- Le strategie di wildcard e pattern matching restano **prevalentemente letterali**: Se una frase si discosta dal pattern previsto, il sistema fallisce il matching

- Le strategie di wildcard e pattern matching restano **prevalentemente letterali**: Se una frase si discosta dal pattern previsto, il sistema fallisce il matching
- Sono disponibili ridotte funzionalità per la gestione di *sinonimi*, semplificazione delle *locuzioni* e *correzione ortografica*

- Le strategie di wildcard e pattern matching restano **prevalentemente letterali**: Se una frase si discosta dal pattern previsto, il sistema fallisce il matching
- Sono disponibili ridotte funzionalità per la gestione di *sinonimi*, semplificazione delle *locuzioni* e *correzione ortografica*
- La **gestione del contesto** (via <that>, <topic>, <star>, ecc.) è rudimentale

- Le strategie di wildcard e pattern matching restano **prevalentemente letterali**: Se una frase si discosta dal pattern previsto, il sistema fallisce il matching
- Sono disponibili ridotte funzionalità per la gestione di *sinonimi*, semplificazione delle *locuzioni* e *correzione ortografica*
- La **gestione del contesto** (via <that>, <topic>, <star>, ecc.) è rudimentale
- L'integrazione (via <sraix>) con **basi di conoscenza esterne** (KB, database, API) è possibile implementando funzioni personalizzate, ma è di difficile gestione

- Le strategie di wildcard e pattern matching restano **prevalentemente letterali**: Se una frase si discosta dal pattern previsto, il sistema fallisce il matching
- Sono disponibili ridotte funzionalità per la gestione di *sinonimi*, semplificazione delle *locuzioni* e *correzione ortografica*
- La **gestione del contesto** (via <that>, <topic>, <star>, ecc.) è rudimentale
- L'integrazione (via <sraix>) con **basi di conoscenza esterne** (KB, database, API) è possibile implementando funzioni personalizzate, ma è di difficile gestione
- Le risposte generate sono **statiche e predefinite**, e non possono essere generate dinamicamente in base a dati esterni o a contesti più ampi in modo automatico

- Sviluppare un sistema di dialogo che superi le limitazioni di AIML evidenziate
- Integrare tecniche di **Natural Language Understanding** (NLU) e **Retrieval-Augmented Generation** (RAG) per migliorare l'esperienza d'uso
- Assicurare una elevata facilità di estensione e personalizzazione per diversi domini e applicazioni

2. Natural Language Understanding

Il primo elemento dello stack di NLP rispetto ad AIML che vogliamo migliorare è il riconoscimento delle intenzioni dell'utente.

Il primo elemento dello stack di NLP rispetto ad AIML che vogliamo migliorare è il riconoscimento delle intenzioni dell'utente.

- Non useremo più un sistema basato su pattern matching ed espressioni regolari
- Riconosceremo la categoria di interazione affidandoci ad un classificatore basato su LLM
- Le parti variabili della frase (slot) verranno estratte tramite un sistema di Named Entity Recognition (NER)

- Essendo un task supervisionato, bisogna partire con l'etichettatura dei dati.
- Il dataset utilizzato proviene dalle precedenti pubblicazioni del progetto NoVAGraphS, e contiene 350 interazioni degli utenti prodotte durante precedenti sperimentazioni.

- Essendo un task supervisionato, bisogna partire con l'etichettatura dei dati.
- Il dataset utilizzato proviene dalle precedenti pubblicazioni del progetto NoVAGraphS, e contiene 350 interazioni degli utenti prodotte durante precedenti sperimentazioni.
- L'annotazione dei dati:
 - Inizialmente è stata effettuata automaticamente
 - Successivamente è stata completamente riveduta ed effettuata manualmente

L'annotazione automatica è stata effettuata tramite prompting:

L'annotazione automatica è stata effettuata tramite prompting:

1. Due LLM diverse (*Gemma2*, *Llama3.1*) sono state eseguite localmente

L'annotazione automatica è stata effettuata tramite prompting:

1. Due LLM diverse (*Gemma2*, *Llama3.1*) sono state eseguite localmente
2. Ciascuna ha ricevuto tutte le interazioni (una per una) assieme alla lista delle possibili classi

2.2 Classificazione

L'annotazione automatica è stata effettuata tramite prompting:

1. Due LLM diverse (*Gemma2*, *Llama3.1*) sono state eseguite localmente
2. Ciascuna ha ricevuto tutte le interazioni (una per una) assieme alla lista delle possibili classi
3. È stata selezionata la classe con majority vote

ID	gemma2:9b	gemma2:9b	llama3.1:8b	llama3.1:8b
0	START	START	START	START
1	GEN_INFO	GEN_INFO	GEN_INFO	GEN_INFO
2	SPEC_TRANS	SPEC_TRANS	TRANS_BETWEEN	TRANS_BETWEEN
3	Please provide the interaction. : START	START	START	START
...
287	REPETITIVE_PAT	REPETITIVE_PAT	REPETITIVE_PAT	REPETITIVE_PAT
288	TRANS_DETAIL	TRANS_DETAIL	TRANS_DETAIL	GEN_INFO
289	GRAMMAR	GRAMMAR	FINAL_STATE	FINAL_STATE

- In seguito a una analisi dei dati è risultato che le classi fossero troppo sbilanciate, e troppo generiche.
- Questo non avrebbe aiutato il modello che sarebbe stato addestrato a riconoscere le classi con precisione e affidabilità.

- In seguito a una analisi dei dati è risultato che le classi fossero troppo sbilanciate, e troppo generiche.
- Questo non avrebbe aiutato il modello che sarebbe stato addestrato a riconoscere le classi con precisione e affidabilità.
- Sono state ridefinite le classi, dividendole in due livelli di granularità:

- In seguito a una analisi dei dati è risultato che le classi fossero troppo sbilanciate, e troppo generiche.
- Questo non avrebbe aiutato il modello che sarebbe stato addestrato a riconoscere le classi con precisione e affidabilità.
- Sono state ridefinite le classi, dividendole in due livelli di granularità:
 - Le **classi principali** da 21 sono state ridotte a 7
 - Sono state introdotte le **classi secondarie** per ogni classe principale, per un totale di 33.

Classe	Scopo	Numero di Esempi
transition	Domande che riguardano le transizioni tra gli stati	77
automaton	Domande che riguardano l'automa in generale	48
state	Domande che riguardano gli stati dell'automa	48
grammar	Domande che riguardano la grammatica riconosciuta dall'automa	33
theory	Domande di teoria generale sugli automi	15
start	Domande che avviano l'interazione con il sistema	6
off_topic	Domande non pertinenti al dominio che il sistema deve saper gestire	2

Tabella 1: Classi principali per la classificazione delle domande

Sottoclassi	Scopo	Numero di Esempi
description	Descrizioni generali sull'automa	14
description_brief	Descrizione generale (breve) sull'automa	10
directionality	Domande riguardanti la direzionalità o meno dell'intero automa	1
list	Informazioni generali su nodi e archi	1
pattern	Presenza di pattern particolari nell'automa	9
representation	Rappresentazione spaziale dell'automa	13

Tabella 2: Classi secondarie per la classe primaria dell'*Automa*

- Del dataset originale sono rimasti solo 229 esempi, divisi in classi sbilanciate.
- Per assicurare che il modello abbia buone prestazioni, è stato necessario aumentare il numero di esempi.
- Sono state generate 851 nuove domande, utilizzando LLM alle quali è stato fornito un insieme di quesiti di una certa classe.
- Per le domande off-topc è stato adoperato il dataset SQUAD ^[1]

^[1]P. Rajpurkar, J. Zhang, K. Lopyrev, e P. Liang, «SQuAD: 100,000+ Questions for Machine Comprehension of Text», in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, J. Su, K. Duh, e X. Carreras, A c. di, Association for Computational Linguistics, nov. 2016, pp. 2383–2392. doi: 10.18653/v1/D16-1264.

2.3 Data Augmentation

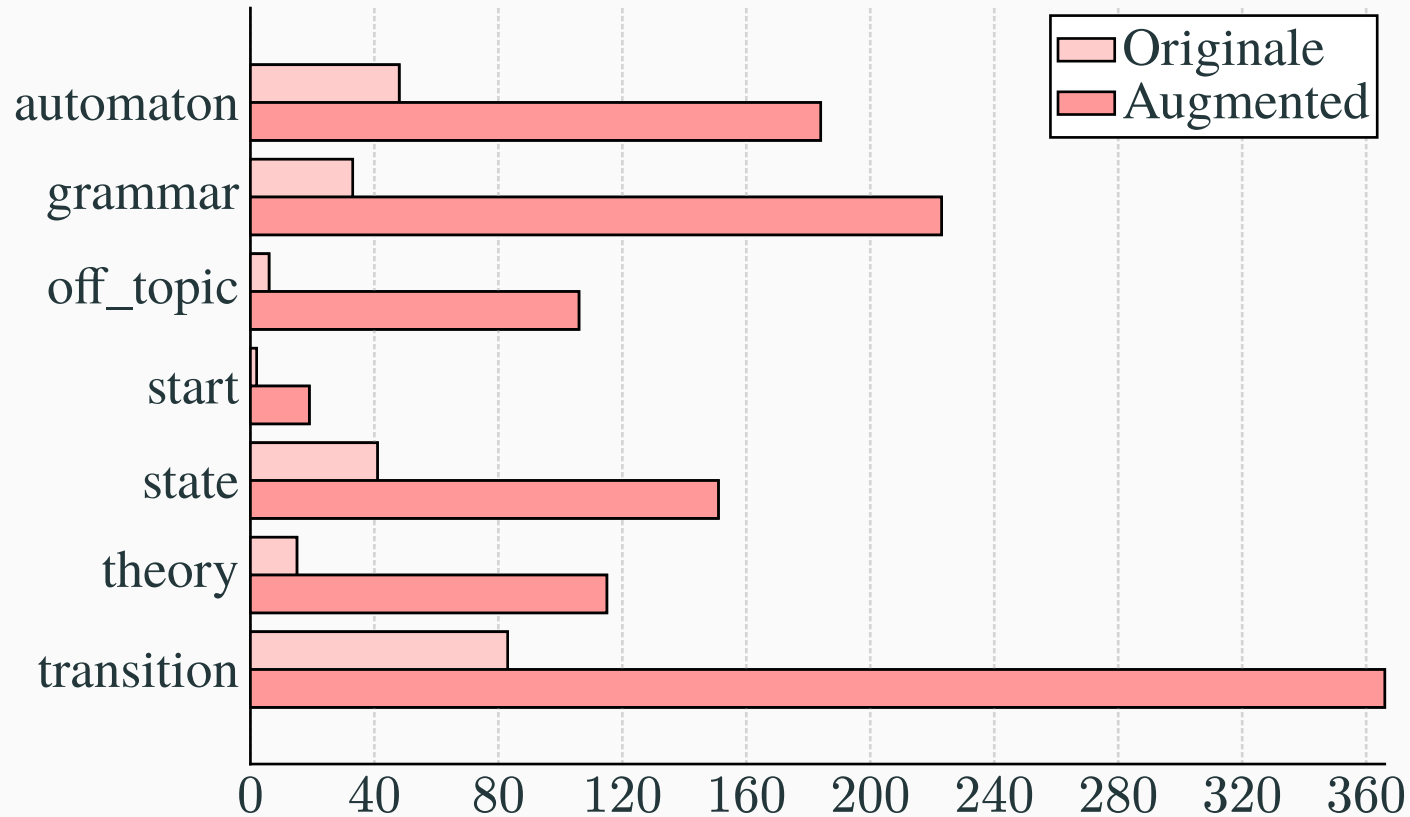


Figura 1: Numero di esempi per classe originale e aumentata

- Partendo LLM pre-addestrate e con una buona padronanza della lingua inglese, l'addestramento è piuttosto rapido e non richiede molte risorse.
- È stato eseguito un fine-tuning per adattare il modello alla classificazione delle domande.
- In questo modo il modello apprende le particolarità e sfumature dello scenario applicativo specifico.
- La metrica massimizzata è stata la **F1 score** (media armonica fra Precision e Recall)

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

I modelli utilizzati per il fine-tuning sono stati:

- google-bert/bert-base-uncased, versione da 110 milioni di parametri ^[1] (Google)
- distilbert/distilbert-base-uncased ^[2], versione distillata di BERT, con circa il 40% in meno di parametri ^[3]
- google/mobilebert-uncased ^[4] con 25 milioni di parametri; per dispositivi mobili
- google/electra-small-discriminator ^[5], da 14 milioni di parametri

^[1]*Bert uncased model by Google*. [Online]. Disponibile su: <https://huggingface.co/google-bert/bert-base-uncased>

^[2]*Distilbert base model*. [Online]. Disponibile su: <https://huggingface.co/distilbert/distilbert-base-uncased>

^[3]V. Sanh, L. Debut, J. Chaumond, e T. Wolf, «DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter», 2020. doi: 10.48550/arXiv.1910.01108.

^[4]*MobileBERT on Huggingface*. [Online]. Disponibile su: <https://huggingface.co/google/mobilebert-uncased>

^[5]*Electra on Huggingface*. [Online]. Disponibile su: <https://huggingface.co/google/electra-small-discriminator>

2.4 Fine-tuning

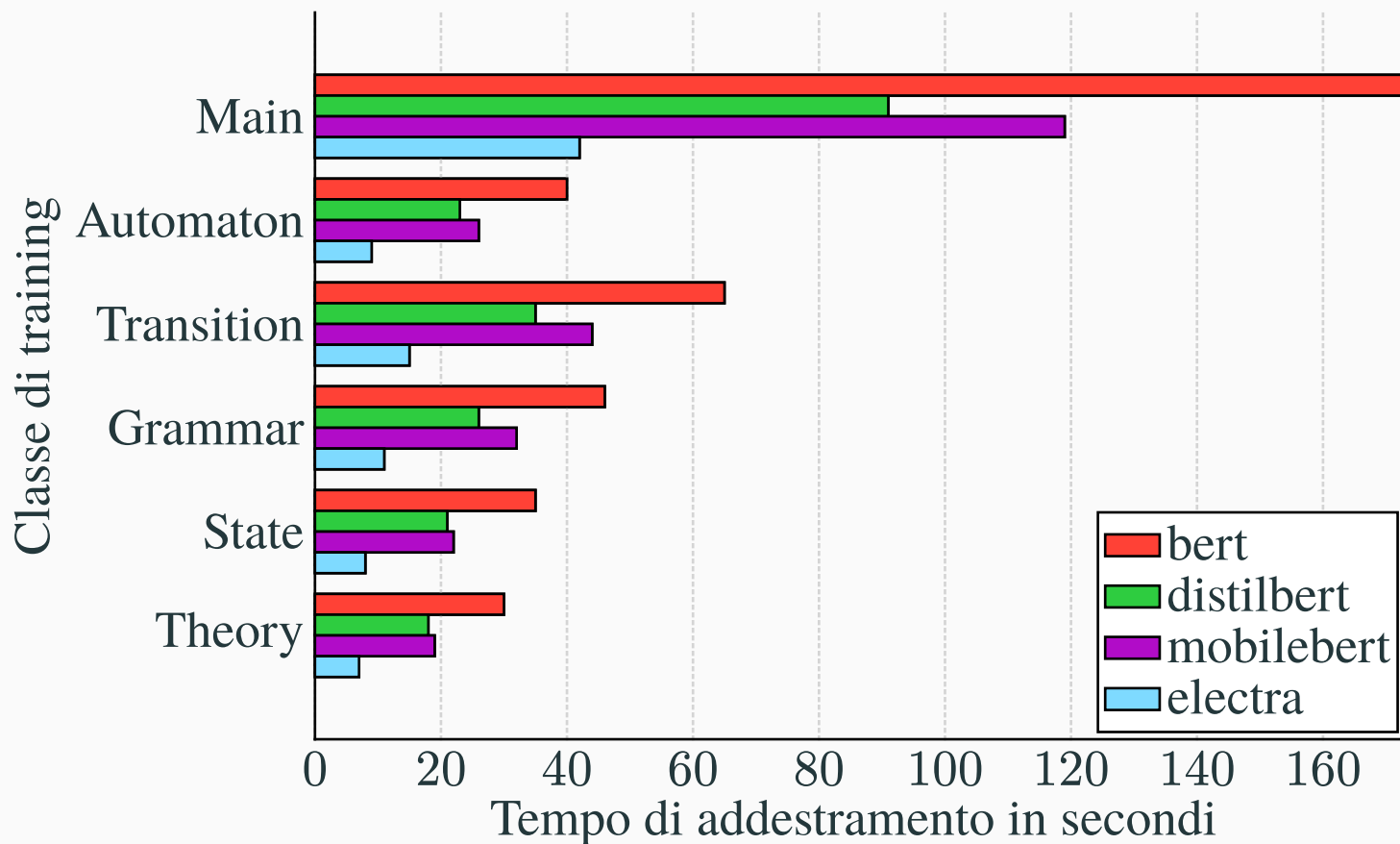


Figura 2: Tempo di addestramento per BERT, DistilBERT, MobileBERT ed ELECTRA

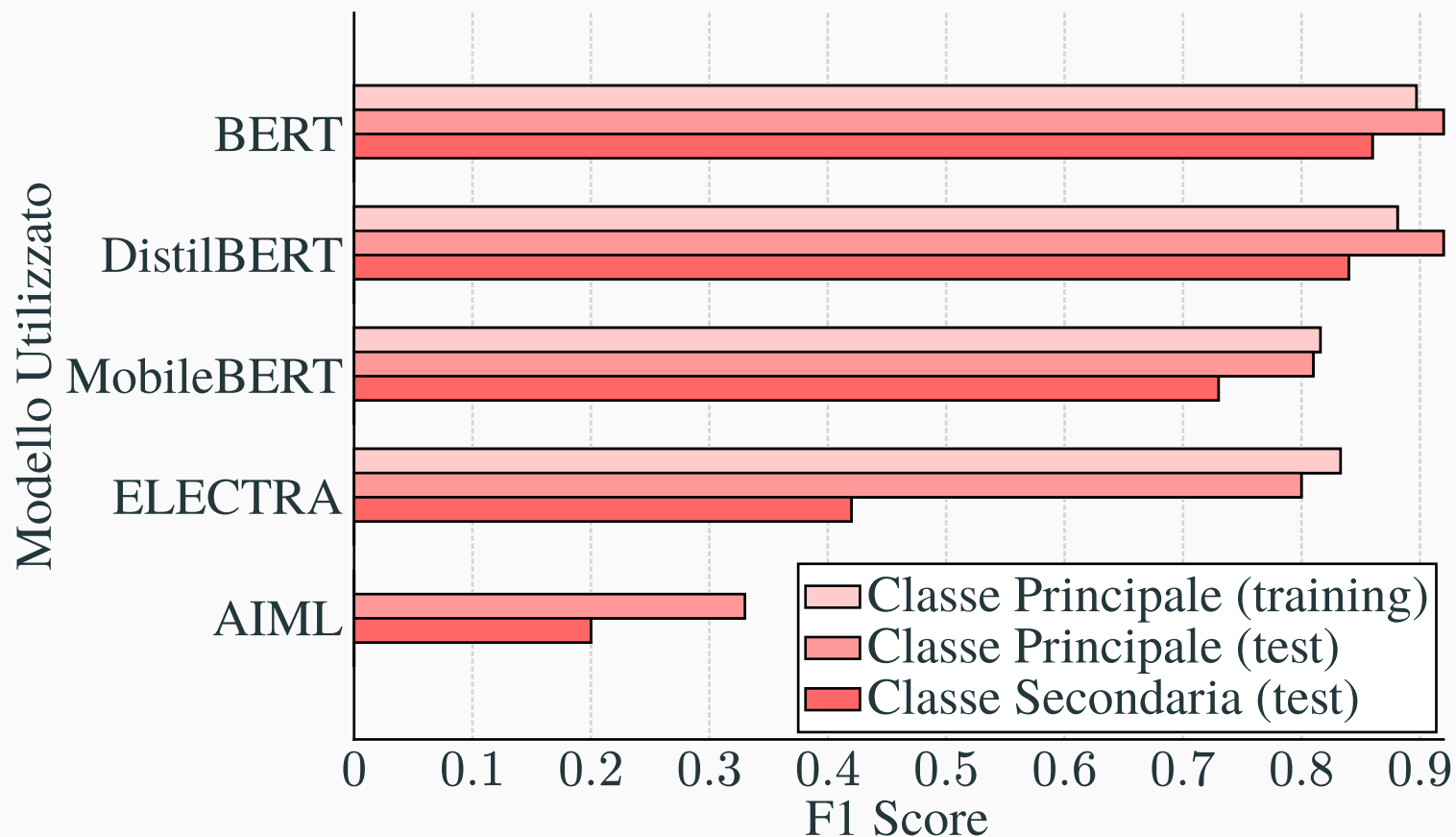


Figura 3: Performance su test set bilanciato da 468 ulteriori domande confrontato con AIML

2.4 Fine-tuning



	Performance AIML			Performance BERT			Esempi
	Precision	Recall	F1-score	Precision	Recall	F1-score	
Automaton	0.52	0.19	0.27	0.93	0.93	0.93	75
Grammar	0.90	0.13	0.23	0.78	0.83	0.81	70
Off-Topic	0.26	0.82	0.39	1.00	0.96	0.98	100
Start	1.00	0.53	0.69	1.00	0.90	0.95	40
State	0.17	0.19	0.18	0.96	1.00	0.98	43
Theory	0.00	0.00	0.00	0.57	0.57	0.57	30
Transition	0.67	0.28	0.40	0.97	0.99	0.98	110
Accuracy	0.35			0.92			468
Macro avg	0.44	0.27	0.27	0.89	0.88	0.88	468
Weighted avg	0.53	0.35	0.33	0.92	0.92	0.92	468

2.4 Fine-tuning

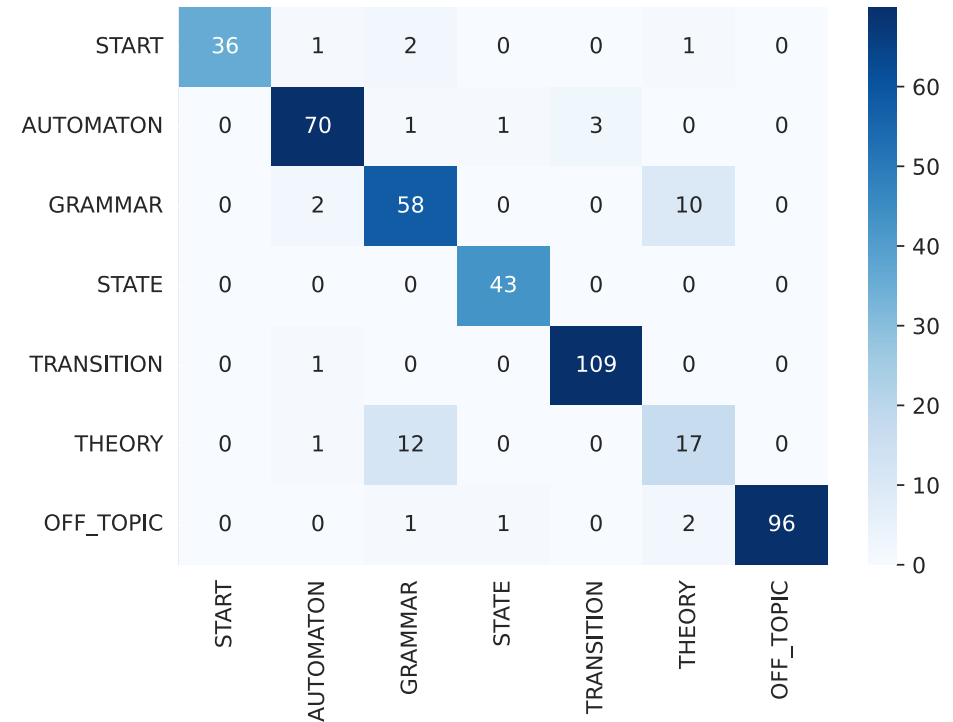
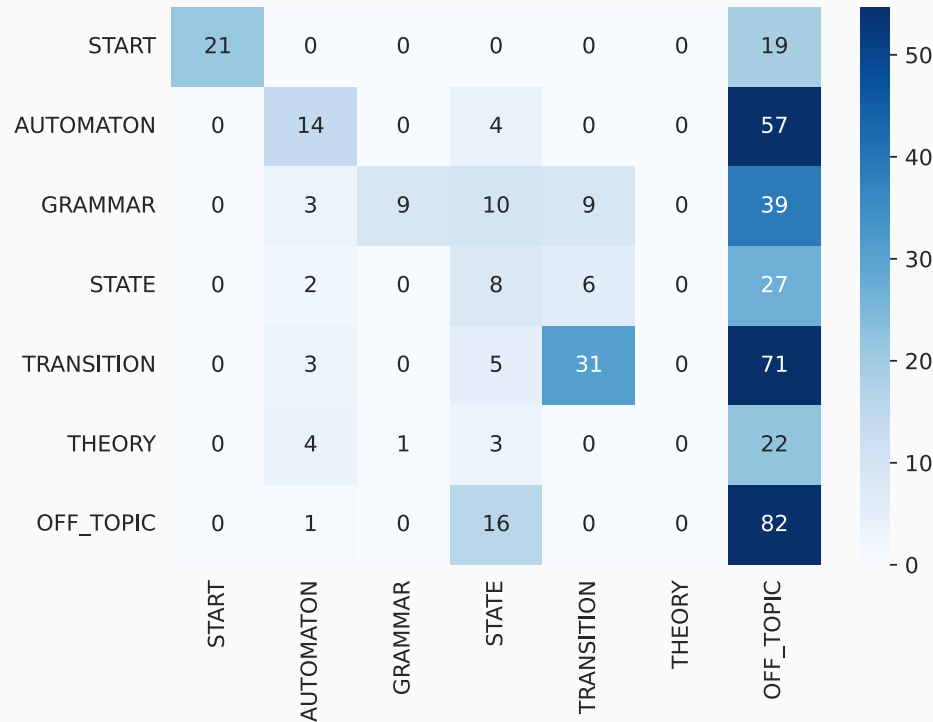





Figura 4: Matrici di confusione per le classi principali con AIML e BERT




Dobbiamo poter estrarre le parti variabili delle domande, in modo da capire quali informazioni l'utente sta cercando. È stato usato un tool per l'annotazione delle frasi, *Doccano* ^[1].

^[1]H. Nakayama, T. Kubo, J. Kamura, Y. Taniguchi, e X. Liang, «doccano: Text Annotation Tool for Human». [Online]. Disponibile su: <https://github.com/doccano/doccano>

2.5 Named entity recognition



  PROJECTS 

1 of 3 < >

The Hitchhiker's Guide to the Galaxy (sometimes referred to as HG2G , HHGTTGor H2G2) is a comedy science fiction series created by Douglas Adams . Originally a radio comedy broadcast on BBC Radio 4 in 1978 , it was later adapted to other formats, including stage shows , novels , comic books , a 1981 TV series, a 1984 video game, and 2005 feature film.

Annotations: Title (The Hitchhiker's Guide to the Galaxy), Title (HG2G), Title (HHGTTGor H2G2), Genre (comedy science fiction), Person (Douglas Adams), Others (BBC Radio 4), Date (1978), Genre (stage shows), Genre (novels), Genre (comic books), Date (1981), Date (1984), Date (2005).

Key	Value
No data available	

Figura 5: Interfaccia di Doccano per l'annotazione dei dati di NER.

In seguito all'etichettatura sono risultate tre classi di entità:

- input: per input o sequenze di simboli.

«Does it only accept 1s and 0s?» \Rightarrow [20, 21, "input"], [27, 28, "input"]

- node: per i frammenti di testo che contengono nodi o stati dell'automa.

«Is there a transition between q2 and q0?» \Rightarrow [30, 32, "node"], [37, 39, "node"]

- language: informazioni sulla grammatica accettata dall'automa.

«Does the automaton accept strings over the alphabet {0,1}?» \Rightarrow [53, 58, "language"]

2.5 Named entity recognition

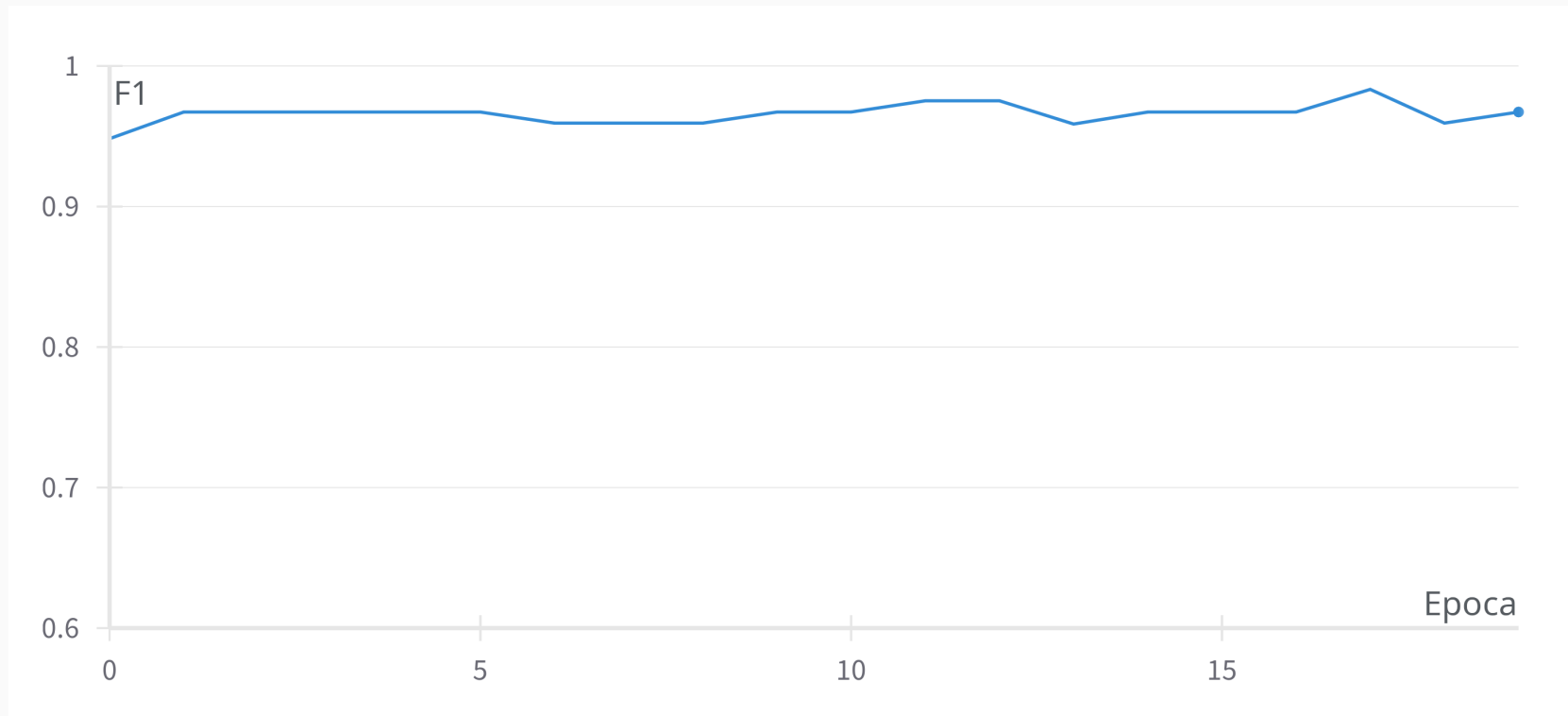


Figura 6: Performance di F1 del modello di NER durante l'addestramento tramite SPACY ^[1].

^[1]M. Honnibal e I. Montani, «spaCy: Industrial-strength Natural Language Processing in Python». [Online]. Disponibile su: <https://spacy.io/>

3. Retrieval Augmented Generation

- Per fornire risposte complete e pertinenti, il sistema deve essere in grado di interrogare una base di conoscenza esterna.

- Per fornire risposte complete e pertinenti, il sistema deve essere in grado di interrogare una base di conoscenza esterna.
- Esistono diverse tecniche, che dipendono dal tipo di dati e dalla loro rappresentazione:
 - Basi di conoscenza strutturate: SQL, SPARQL, ecc.

- Per fornire risposte complete e pertinenti, il sistema deve essere in grado di interrogare una base di conoscenza esterna.
- Esistono diverse tecniche, che dipendono dal tipo di dati e dalla loro rappresentazione:
 - Basi di conoscenza strutturate: SQL, SPARQL, ecc.
 - Corpus: ricerca full-text, TF-IDF, embeddings...

- Per fornire risposte complete e pertinenti, il sistema deve essere in grado di interrogare una base di conoscenza esterna.
- Esistono diverse tecniche, che dipendono dal tipo di dati e dalla loro rappresentazione:
 - Basi di conoscenza strutturate: SQL, SPARQL, ecc.
 - Corpus: ricerca full-text, TF-IDF, embeddings...
 - API e servizi esterni: REST, JMESPath

```
SELECT customer_name, order_date, total_amount
FROM orders
WHERE order_id = :orderId
      AND customer_id = :customerId;
```

Query 1: Esempio di query SQL per il recupero di dettagli di un ordine in un sistema e-commerce.

```
PREFIX dbo: <http://dbpedia.org/ontology/>
PREFIX dbr: <http://dbpedia.org/resource/>

SELECT ?director
WHERE {
    dbr:The_Incredibles dbo:director ?director .
}
```

Query 2: Query SPARQL per il recupero del regista del film *Gli Incredibili* (Brad Bird).

3.1 Retrieval

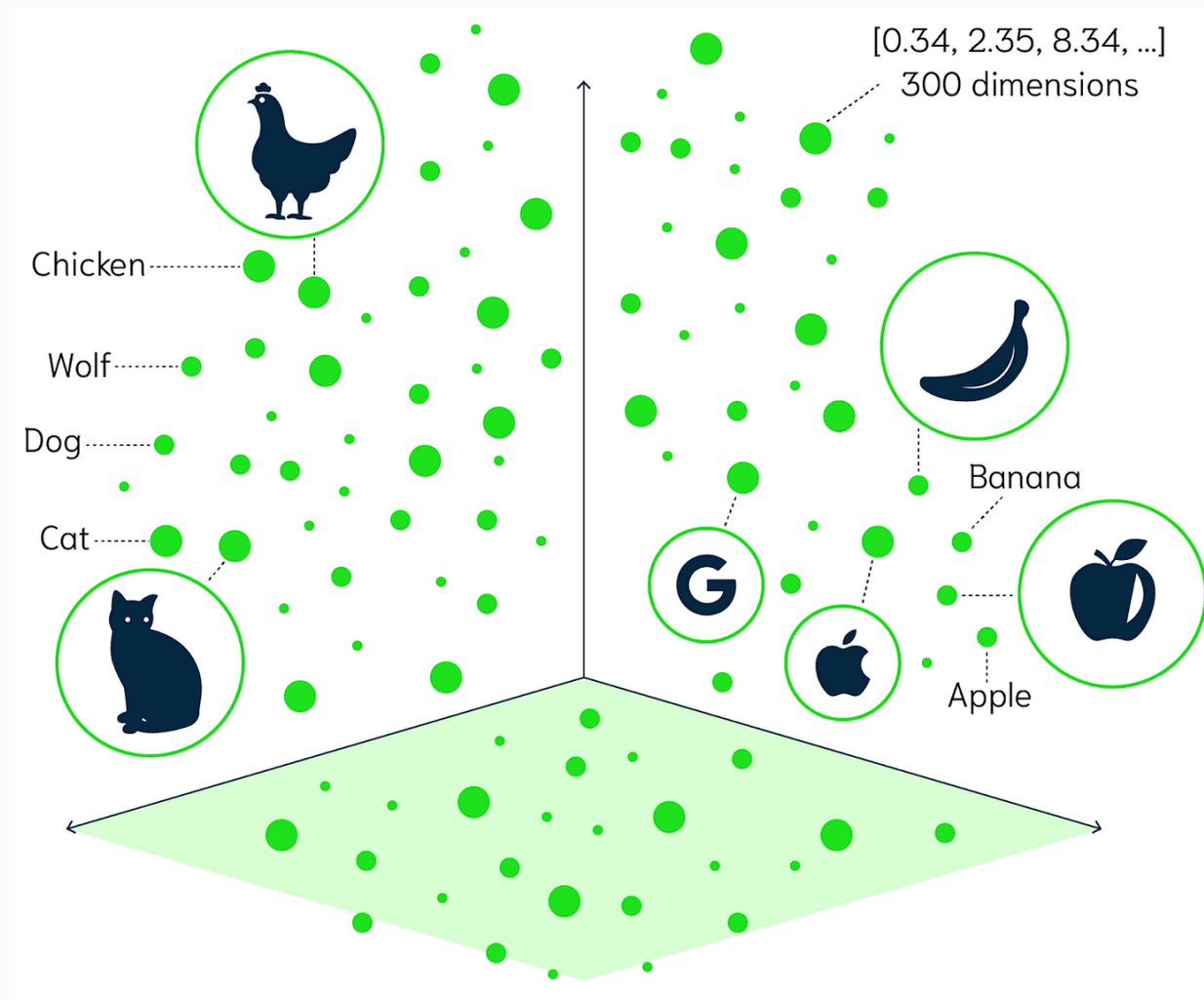


Figura 7: Illustrazione della cosine similarity.

Prompt

Ecco i dati estratti dal database di e-commerce:

- Prezzo dell'articolo: 49,99 euro
- Tempi di spedizione: 2 giorni
- Colori disponibili: rosso, blu, verde

Genera una breve risposta da mostrare al cliente, evitando informazioni non pertinenti e senza inventare nulla.

Risposta da GPT-4o

L'articolo è disponibile nei colori rosso, blu e verde al prezzo di 49,99 euro. I tempi di spedizione sono di 2 giorni.

Prompt

Sei un assistente che fornisce informazioni su ordini online. Di seguito trovi una selezione parziale delle interazioni con l'utente:

Utente: “Qual è lo stato di avanzamento del mio ordine?”

Il sistema ha reperito i seguenti dati:

- Ordine n. 1357
- Stato: in spedizione
- Previsione di consegna: 10/04/2025

Ora, rispondi alla domanda dell'utente in modo chiaro e conciso, mantenendo la coerenza con le interazioni precedenti.

- Dobbiamo assicurarci che il sistema sia in grado di rispondere in modo non solo coerente e pertinente, ma anche efficace alle domande degli utenti.
- Per effettuare le valutazioni sono stati utilizzati diversi LLM
- Una volta generate tutte le risposte con i vari modelli, sono state tutte annotate
- I risultati delle annotazioni hanno fornito dettagli essenziali per la scelta del modello finale

Sono state scelte delle LLM open-weights dati i multipli vantaggi che offrono:

- Trasparenza e accountability
 - Accesso ai pesi e alla struttura dei modelli per identificare bias e anomalie.
 - Maggiore controllo sulla sicurezza e affidabilità dell'AI.

Sono state scelte delle LLM open-weights dati i multipli vantaggi che offrono:

- Trasparenza e accountability
 - Accesso ai pesi e alla struttura dei modelli per identificare bias e anomalie.
 - Maggiore controllo sulla sicurezza e affidabilità dell'AI.
- Verifiche indipendenti e replicabilità per ricercatori e sviluppatori. Maggiore fiducia.

Sono state scelte delle LLM open-weights dati i multipli vantaggi che offrono:

- Trasparenza e accountability
 - Accesso ai pesi e alla struttura dei modelli per identificare bias e anomalie.
 - Maggiore controllo sulla sicurezza e affidabilità dell'AI.
- Verifiche indipendenti e replicabilità per ricercatori e sviluppatori. Maggiore fiducia.
- Innovazione e competizione
 - Riduzione delle barriere d'ingresso per startup e centri di ricerca.
 - Non sono necessarie per forza grandi risorse hardware
 - Protezione della privacy e riduzione dei costi operativi.

Sono state scelte delle LLM open-weights dati i multipli vantaggi che offrono:

- Trasparenza e accountability
 - Accesso ai pesi e alla struttura dei modelli per identificare bias e anomalie.
 - Maggiore controllo sulla sicurezza e affidabilità dell'AI.
- Verifiche indipendenti e replicabilità per ricercatori e sviluppatori. Maggiore fiducia.
- Innovazione e competizione
 - Riduzione delle barriere d'ingresso per startup e centri di ricerca.
 - Non sono necessarie per forza grandi risorse hardware
 - Protezione della privacy e riduzione dei costi operativi.
- Supporto a uno sviluppo sostenibile e responsabile dell'AI.

3.3 Generazione delle risposte



You are a helpful assistant expert in finite state automata.

Answer the question given by the user using the retrieved data, using plain text only.

Avoid referring to the data directly; there is no need to provide any additional information.

Keep the answer concise and short, and avoid using any additional information not provided.

The system has retrieved the following data:

\ \ \

{data}

\ \ \

The user has asked the following question:

\ \ \

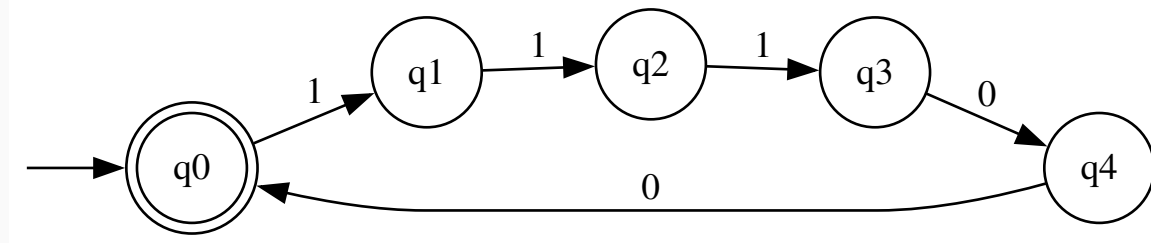
{question}

\ \ \

3.3 Generazione delle risposte



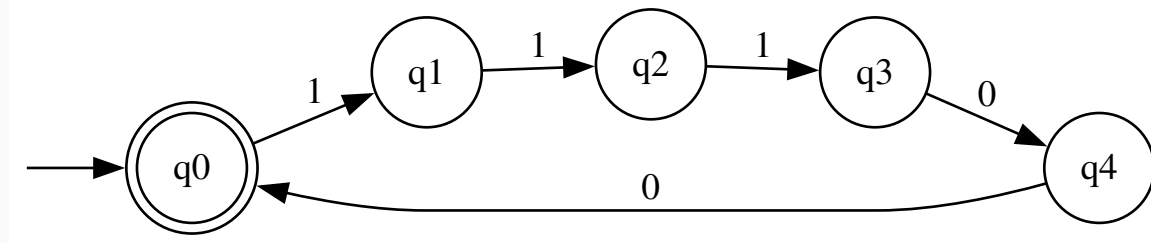
Per la generazione delle risposte è stato selezionato un sottinsieme di domande che richiedono informazioni riguardo questo specifico automa a stati finiti:



3.3 Generazione delle risposte



Per la generazione delle risposte è stato selezionato un sottinsieme di domande che richiedono informazioni riguardo questo specifico automa a stati finiti:



Se il tema della domanda è riguardante le *transizioni uscenti da un nodo*, prima di generare la risposta, il sistema recupera le i dettagli utili e li presenta al modello.

The transitions exiting from the node are the following:

- From q0 to q1, with label '1'

3.3 Generazione delle risposte



```
digraph FSA {  
    rankdir=LR;  
    node [shape = circle];  
    q0 [shape = doublecircle];  
    q1; q2; q3; q4;  
  
    start [shape=none, label=""];  
    start -> q0;  
  
    q0 -> q1 [label = "1"];  
    q1 -> q2 [label = "1"];  
    q2 -> q3 [label = "1"];  
    q3 -> q4 [label = "0"];  
    q4 -> q0 [label = "0"];  
}
```

Snippet 1: Rappresentazione in formato Graphviz dell'automa a stati finiti utilizzato come input per le domande.

- È necessario annotare le risposte per valutare le performance degli LLM candidati
- Basato sulle ricerche di Z. Kasner e O. Dušek ^[1]
- Verrà utilizzato il software Factgenie ^[2] da loro sviluppato per le annotazioni

^[1]Z. Kasner e O. Dušek, «Beyond Traditional Benchmarks: Analyzing Behaviors of Open LLMs on Data-to-Text Generation», in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, L.-W. Ku, A. Martins, e V. Srikumar, A c. di, Association for Computational Linguistics, 2024, pp. 12045–12072. [Online]. Disponibile su: <https://aclanthology.org/2024.acl-long.651>

^[2]*Factgenie*. [Online]. Disponibile su: <https://github.com/ufal/factgenie>

Gli annotatori sono liberi di evidenziare nelle risposte frammenti problematici semplicemente selezionandoli. Sono stati definiti quattro generi di errori:

- **INCORRECT^I**: la risposta contiene informazioni che contraddicono i dati forniti o che sono chiaramente sbagliate.
- **NOT_CHECKABLE^{NC}**: la risposta contiene informazioni che non possono essere verificate con i dati forniti.
- **MISLEADING^M**: la risposta contiene informazioni fuorvianti o che possono essere interpretate in modo errato.
- **OTHER^O**: la risposta contiene errori grammaticali, stilistici o di altro tipo.

12345678910111213141516171819202122232425

What are the incoming and outgoing transition paths for q1?

Instructions

Incorrect

Not checkable

Misleading

Other

Drag your mouse over the text to highlight the span:

Incoming transition for q1 comes from q0 with label "1". Outgoing transition goes to q2 with label "1".
Incoming: q0 → q1 (label "1") Outgoing: q1 → q2 (label "1")

Please check if you agree with any of the following statements:

☒ The text 100% accurate and does not contain any errors.

☐ The text is missing or incomplete.

☐ The text is severely off-topic (seems completely unrelated to the data).

Answer Clearness (if it is understandable)

Completely clear

Answer length (if it was too short to be 100% clear or

Figura 9: Interfaccia di Factgenie

44 / 52

Oltre ad evidenziare parti problematiche delle risposte, è stato richiesto agli annotatori anche di fornire delle valutazioni qualitative su alcune metriche:

- **Accuratezza della risposta:** da selezionare quando la risposta è corretta al 100% e non contiene errori sui dati;
- **Assenza o incompletezza di informazioni:** da selezionare quando la risposta non contiene tutte le informazioni rilevanti;
- **Totale incongruenza della risposta:** da selezionare quando la risposta appare completamente scorrelata o non pertinente alla domanda;
- **Chiarezza della risposta:** se è comprensibile e ben strutturata;
- **Lunghezza della risposta:** se la comprensione della risposta è facilitata dalla sua lunghezza (o brevità);
- **Utilità percepita della risposta:** se la risposta è utile e fornisce informazioni rilevanti;
- **Apprezzamento generale:** se la risposta è apprezzata o gradita.

In totale, 12 annotatori hanno partecipato alla valutazione delle risposte generate dai modelli.

Riguardo gli annotatori:

- 8 su 12 sono studenti del Dipartimento di Informatica;
- 2 hanno un background ingegneristico;
- I rimanenti 2 provengono dal Dipartimento di Storia e Biologia.

L'età media è di 28 anni, con un range tra i 21 e i 68 anni.

Ogni volontario ha valutato un sottoinsieme delle risposte, garantendo comunque un overlap sulle 75 risposte totali.

Oltre agli annotatori umani, sono stati utilizzati anche due modelli di LLM (GPT-o3-mini e GPT-4.5) per svolgere automaticamente una valutazione simile a quella dei volontari (ϵ_{hum}).

Modello	Incorrect		Not Checkable		Misleading		Other		Globale	
	ϵ_{hum}	$\epsilon_{4.5}$	ϵ_{hum}	$\epsilon_{4.5}$	ϵ_{hum}	$\epsilon_{4.5}$	ϵ_{hum}	$\epsilon_{4.5}$	ϵ_{hum}	$\epsilon_{4.5}$
Deepseek	20.0%	13.3%	6.67%	0.0%	26.7%	6.7%	66.7%	0.0%	83.0%	33.3%
Gemma2	26.7%	26.7%	0.0%	0.0%	33.3%	0.0%	20.0%	6.7%	30.6%	33.3%
Llama3.1	33.3%	33.3%	6.67%	0.0%	26.7%	6.7%	46.7%	0.0%	67.3%	53.3%
GPT-4o	20.0%	6.67%	13.3%	0.0%	20.0%	0.0%	40.0%	0.0%	36.0%	6.67%
GPT-o3-mini	0.0%	13.3%	0.0%	0.0%	13.3%	0.0%	20.0%	0.0%	18.6%	20.0%

Tabella 3: Percentuali di *risposte contenenti almeno un errore*, secondo le annotazioni umane (ϵ_{hum}) e le valutazioni automatiche ($\epsilon_{4.5}$). Più basso è il valore, migliore è la qualità delle risposte.

Modello	Incorrect		Not Checkable		Misleading		Other		Globale	
	ε_{hum}	$\varepsilon_{4.5}$	ε_{hum}	$\varepsilon_{4.5}$	ε_{hum}	$\varepsilon_{4.5}$	ε_{hum}	$\varepsilon_{4.5}$	ε_{hum}	$\varepsilon_{4.5}$
Deepseek-r1:8b	0.13	0.26	0.07	0	0.1	0.06	0.53	0	0.83	0.33
Gemma2:9b	0.16	0.26	0	0	0.06	0	0.07	0.06	0.31	0.33
Llama3.1:8b	0.29	0.47	0	0	0.07	0.07	0.31	0	0.67	0.53
GPT-4o	0.15	0.07	0.1	0	0.02	0	0.09	0	0.36	0.07
GPT-o3-mini	0	0.2	0	0	0.01	0	0.17	0	0.19	0.2

Tabella 4: Numero medio di *errori per output*, per ogni categoria di errore e in totale, secondo le annotazioni umane (ε_{hum}) e le valutazioni automatiche ($\varepsilon_{4.5}$). Più basso è il valore, migliore è la qualità delle risposte.

Metrica	LLM Commerciale		LLM Open-Weights	
Chiarezza	GPT-o3-mini	93%	Deepseek-r1:8b	69%
Lunghezza	GPT-o3-mini	96%	Deepseek-r1:8b	78%
Utilità	GPT-o3-mini	98%	Deepseek-r1:8b	82%
Apprezzamento	GPT-o3-mini	95%	Deepseek-r1:8b	68%

Tabella 5: Percentuali di valutazione delle risposte generate dai modelli LLM commerciali e open-weights.

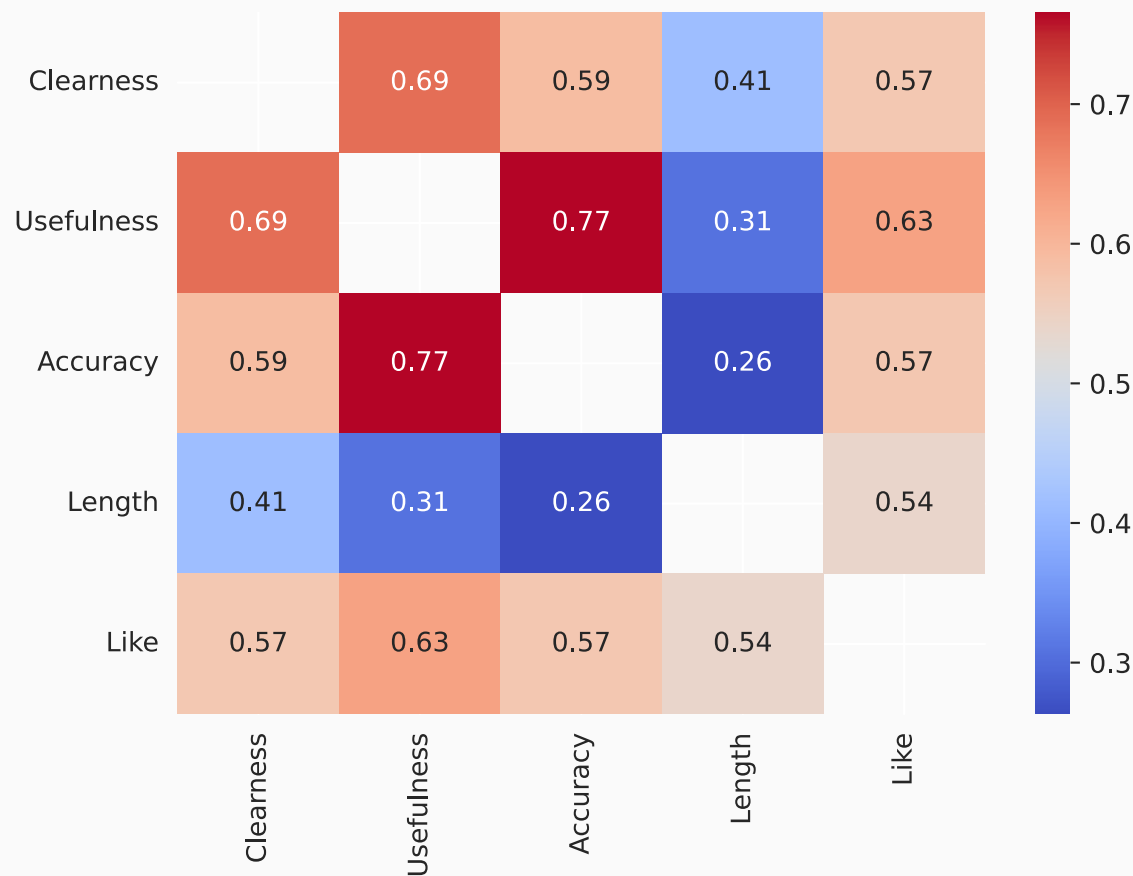


Diagramma 1: Correlazione tra le metriche di valutazione delle risposte per ogni modello.

4. Ingegnerizzazione del sistema

5. Conclusioni

Grazie per l'attenzione!
Domande?