Università degli Studi di Torino
Scuola di Scienze della Natura
Corso di Laurea Magistrale in Informatica

Tesi di Laurea Magistrale

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

RELATORE
Prof. Alessandro Mazzei

CO-RELATORI
Pier Felice Balestrucci
Michael Oliverio

CANDIDATO
Stefano Vittorio Porta
859133

Anno Accademico 2023/2024

Dichiarazione di Originalità

Dichiaro di essere responsabile del contenuto dell'elaborato che presento al fine del conseguimento del titolo, di non avere plagiato in tutto o in parte il lavoro prodotto da altri e di aver citato le fonti originali in modo congruente alle normative vigenti in materia di plagio e di diritto d'autore. Sono inoltre consapevole che nel caso la mia dichiarazione risultasse mendace, potrei incorrere nelle sanzioni previste dalla legge e la mia ammissione alla prova finale potrebbe essere negata.

Ringraziamenti

Todo

Abstract

Todo.

Parole chiave

NLU mediante classificazione, data annotation, data augmentation, data retrieval, NLG basata su LLM, software engineering

Indice

1 Introduzione

1.1 Contesto generale

dare un'idea di chatbot, cos'è NovaGraphS, perché è importante, come si inserisce nel contesto delle tecnologie assistive

1.2 Motivazioni e obiettivi della tesi

1.3 Struttura del documento

2 Natural Language Understanding

(introduzione all'argomento)

2.1 Come AIML gestisce la comprensione

Negli anni '90 iniziò a guadagnare popolarità il Loebner Prize [1], una competizione ispirata al Test di Turing [2].
Nella competizione, chatbot e sistemi conversazionali cercavano di "ingannare" giudici umani, facendo credere loro di essere persone reali. Molti sistemi presentati alla competizione erano basati su pattern matching e rule-based, a volte integrando euristiche per la gestione di sinonimi o correzione ortografica.

Tra questi, uno dei più celebri è *ALICE* (Artificial Linguistic Internet Computer Entity), sviluppato da Richard Wallace utilizzando il linguaggio di markup AIML (Artificial Intelligence Markup Language) da lui introdotto [3], [4].
ALICE vinse per la prima volta il Loebner Prize nel 2000, e in seguito vinse altre due edizioni, nel 2001 e 2004.

2.1.1 Struttura di un chatbot AIML

Basato sull'XML [3], di base l'AIML fornisce una struttura formale per definire regole di conversazione attraverso **categorie** di *pattern* e *template*:

- `<pattern>` : la frase (o le frasi) attese in input a cui il chatbot deve reagire;
- `<template>` : la risposta (testuale o con elementi dinamici) che il chatbot fornisce quando si verifica il match del pattern.

La forma più semplice di categoria è:

```
<category>
  <pattern>CIAO</pattern>
  <template>Ciao! Come posso aiutarti oggi?</template>
</category>
```

In questo caso, se l'utente scrive "Ciao"[1], il sistema restituisce la risposta associata nella sezione del `<template>`.

Naturalmente questa è una regola basilare: AIML permette di definire pattern molto più complessi.
Un primo passo verso la creazione di regole più flessibili è l'uso di wildcard: associando simboli quali $*$ e – a elementi di personalizzazione (`<star/>`), il motore che esegue la configurazione AIML può gestire un certo grado di variabilità linguistica:

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

[1] Caratteri maiuscoli e minuscoli sono considerati uguali dal motore di riconoscimento.

In particolare, il simbolo `*` corrisponde a una wildcard che cattura qualsiasi sequenza di parole in input tra i due pattern specificati.
In questo caso, se l'utente digita "Mi chiamo Andrea", il sistema sostituisce `<star/>` con "Andrea", e risponde di conseguenza.

Spesso è necessario memorizzare informazioni fornite dall'utente per utilizzarle successivamente. A questo scopo, AIML offre i tag `<set>` e `<get>` che, rispettivamente, memorizzano e recuperano valori da variabili di contesto:

```
<category>
  <pattern>IL MIO COLORE PREFERITO È *</pattern>
  <template>
    <think>
      <set name="colore"><star/></set>
    </think>
    Ok, ricorderò che il tuo colore preferito è <star/>.
  </template>
</category>

<category>
  <pattern>QUAL È IL MIO COLORE PREFERITO</pattern>
  <template>
    Il tuo colore preferito è <get name="colore"/>.
  </template>
</category>
```

Nella prima `<category>`, il tag `<think>` fa sì che l'operazione di memorizzazione non produca output testuale per l'utente, ma aggiorni internamente la variabile `colore`.
Nel secondo blocco, si utilizza `<get name="colore"/>` per restituire all'utente il valore memorizzato.

Il tag `<condition>` permette di definire regole condizionali in base a variabili di contesto.

Se la variabile stagione (presumibilmente impostata altrove con un `<set>`) ha valore `inverno`, verrà restituito "Fa piuttosto freddo…". Un risultato simile si ottiene per `estate`, mentre per altri valori o mancanza di valore si restituisce l'ultimo *list item*.

```
<category>
  <pattern>CHE TEMPO FA</pattern>
  <template>
    <condition name="stagione">
      <li value="inverno">Fa piuttosto freddo, in questa stagione.</li>
      <li value="estate">Fa molto caldo, bevi tanta acqua!</li>
      <li>Non saprei, forse è primavera o autunno?</li>
    </condition>
  </template>
</category>
```

Il tag `<topic>` permette di raggruppare categorie che appartengono a un medesimo ambito di conversazione, per *facilitare la lettura* delle regole:

```
<topic name="saluti">
  <category>
    <pattern>CIAO</pattern>
    <template>
      Ciao, piacere di vederti!
    </template>
  </category>

  <category>
    <pattern>COME STAI</pattern>
    <template>
      Sto bene, grazie! E tu come stai?
    </template>
  </category>
</topic>
```

In questo modo le regole legate ai saluti sono tutte contenute all'interno di un blocco `<topic>` chiamato `saluti`.

Il tag `<srai>`[2] permette di reindirizzare l'input ad un'altra regola, che verrà processata come se fosse stata digitata dall'utente. È molto utile per riutilizzare risposte o logiche già definite:

```
<topic name="saluti">
  <category>
    <pattern>SALUTA *</pattern>
    <template>
      <srai>CIAO</srai>
    </template>
  </category>
</topic>
```

Se l'utente scrive "Saluta Andrea", la regola cattura "SALUTA *" e reindirizza il contenuto (in questo caso "CIAO") a un'altra categoria. Se esiste una categoria che gestisce il pattern "CIAO", verrà attivata la relativa risposta.

Esiste anche una versione contratta di `<srai>` chiamata `<sr>`, che è stata prevista come scorciatoia quando è necessario matchare un solo pattern. Secondo la documentazione, il tag corrisponde a `<srai><star/></srai>`.

Abbiamo già visto `<think>` in azione per evitare che il contenuto venga mostrato all'utente. In generale, `<think>` è utile quando vogliamo impostare o manipolare variabili senza generare output visibile, ad esempio:

```
<category>
  <pattern>ADESSO È *</pattern>
  <template>
    <think><set name="stagione"><star/></set></think>
    Grazie, ora so che la stagione attuale è <star/>!
  </template>
</category>
```

[2] Stimulus-Response Artificial Intelligence [3]

Il tag `<that>` permette di scrivere pattern che dipendono dalla risposta precedentemente fornita dal chatbot. È particolarmente utile per gestire contesti conversazionali più complessi:

```
<category>
  <pattern>SI</pattern>
  <that>VA TUTTO BENE</that>
  <template>Felice di averti aiutato!</template>
</category>
```

In questo caso la regola sarà attivata se la risposta precedente del bot era "VA TUTTO BENE" e l'utente risponde in modo affermativo.

Per rendere la conversazione più naturale, AIML 2.0 fornisce `<random>`, che permette di restituire una risposta fra più alternative:

```
<category>
  <pattern>COME VA</pattern>
  <template>
    <random>
      <li>Benissimo, grazie!</li>
      <li>Abbastanza bene, e tu?</li>
      <li>Non c'è male, e tu come stai?</li>
    </random>
  </template>
</category>
```

Ogni volta che l'utente scrive "Come va", il bot sceglierà casualmente una delle tre risposte elencate.

Alcune versioni di AIML supportano `<learn>`, che consente al bot di aggiungere nuove categorie "al volo" durante l'esecuzione:

```
<category>
  <pattern>TI INSEGNO *</pattern>
  <template>
    <think>
      <learn>
        <![CDATA[
          <category>
            <pattern><star/></pattern>
            <template>Ho imparato a rispondere a "<star/>"!</template>
          </category>
        ]]>
      </learn>
    </think>
    Ho imparato una nuova regola!
  </template>
</category>
```

2.1.2 Criticità e limiti di AIML

Grazie ai tag previsti dallo schema, AIML riesce a gestire conversazioni piuttosto complesse. Ciononostante, presenta comunque alcune limitazioni:

- Le strategie di wildcard e pattern matching restano prevalentemente letterali, con limitata capacità di interpretare varianti linguistiche non codificate nelle regole.
  Se una frase si discosta dal pattern previsto, il sistema fallisce il matching. Sono disponibili comunque alcune funzionalità per la gestione di sinonimi, semplificazione delle locuzioni e correzione ortografica (da comporre e aggiornare manualmente) che possono mitigare alcuni di questi problemi.
- La gestione del contesto (via `<that>`, `<topic>`, `<star>`, ecc.) è rudimentale, soprattutto se paragonata a sistemi moderni di NLU con modelli neurali che apprendono contesti ampi e riescono a tenere traccia di dettagli dal passato della conversazione.
- L'integrazione con basi di conoscenza esterne (KB, database, API) richiede estensioni o script sviluppati ad-hoc, poiché AIML di per sé non offre costrutti semantici o query integrate, e non permette di integrare script internamente alle regole [3].
- 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 (come invece avviene con LLM e modelli di generazione di linguaggio).

Nonostante questi limiti, AIML ha rappresentato un passo importante nell'evoluzione dei chatbot, offrendo un framework standardizzato e relativamente user-friendly per la creazione di agenti rule-based [4].
In alcuni ambiti ristretti (FAQ, conversazioni scriptate, assistenti vocali), costituisce ancora una soluzione valida e immediata. In domini più complessi, in cui la varietà del linguaggio e l'integrazione con dati dinamici sono essenziali, diventa indispensabile affiancare o sostituire AIML con tecniche di Natural Language Understanding basate su machine learning e deep learning.

Nelle sezioni successive sarà mostrato il percorso seguito per cercare di migliorare la comprensione degli input dell'utente, integrando tecniche di NLU basate su modelli di linguaggio neurali, e valutando le performance ottenute rispetto ad AIML.

2.2 Classificazione con LLM

Come detto poco sopra, uno dei limiti di AIML è la gestione limitata di varianti linguistiche e contesti conversazionali.

Per permettere all'AIML di generalizzare sulle richieste degli utenti, il botmaster deve dichiarare delle generalizzazioni esplicite, ad esempio utilizzando wildcard o pattern che catturano più varianti di una stessa richiesta. Questo processo richiede tempo e competenze linguistiche, oltre ad una grande attenzione per evitare ambiguità o sovrapposizioni tra regole.

Durante il mio percorso di ricerca ho deciso di seguire una strada simile a quella di AIML, ma facendo un passo indietro e ponendomi la domanda:

"Invece che cercare dei pattern nelle possibili richieste degli utenti, perchè non trovare un modello che possa generalizzare su queste richieste in modo automatico?"

Il percorso per arrivare al modello di classificazione di intenti ha richiesto i suoi tempi, ma alla fine ho ottenuto dei risultati che ritengo soddisfacenti.

I problemi principali da risolvere per poter classificare gli intenti sono due: la raccolta di dati etichettati e la scelta del modello di classificazione.

2.2.1 Dataset di training

Di base, nel mondo dell'apprendimento automatico supervisionato, per addestrare un modello di classificazione è necessario un dataset di esempi etichettati, cioè coppie di input e output che il modello deve apprendere a generalizzare.

Per la classificazione di intenti, i dataset più comuni sono quelli di chatbot e assistenti vocali, che contengono domande e richieste etichettate con l'intento che l'utente vuole esprimere.

Il dataset originario fornitomi è stato composto in seguito a una campagna di raccolta dati manuale, in cui diversi collaboratori hanno interagito con un prototipo di chatbot AIML, ponendo domande e richieste di vario tipo.

Il dataset è una collezione di circa 700 singole interazioni "botta e risposta" prodotte dagli utenti durante la prima fase di sperimentazione. Metà sono domande, l'altra metà coincide con ciò che il chatbot ha risposto.

Estrazione dei dati

Dovendo addestrare un modello di classificazione, ho proceduto innanzitutto con l'estrazione dei dati effettivamente a noi necessari. Un piccolo script python che adopera la libreria `pandas` [5] è stato sufficiente:

```python
import pandas as pd
from dotenv import load_dotenv

load_dotenv()

df_o = pd.read_excel('corpus/interaction-corpus.xlsx')

# Filter only the rows that have "Participant" as 'U'
df = df_o[df_o['Participant'] == 'U']
df = df[['Text']]
df = df.drop_duplicates()
df = df[df['Text'].apply(lambda x: isinstance(x, str))]
df['Text'] = df['Text'].str.strip()  # Remove trailing whitespace
texts = df['Text'].dropna()

df.to_csv("./filtered_data.csv")
```

Codice 1: Estrazione dei dati dal dataset di interazione.

Estratte le domande, ho potuto procedere con l'etichettatura.

In un primo step, ho considerato la possibilità di lasciare il compito di etichettatura delle domande ad un sistema che svolgesse il compito in automatico.
Questo permetterebbe di avere un dataset decorato, senza dover ricorrere a un'etichettatura manuale che sarebbe stata molto dispendiosa in termini di tempo e risorse, specialmente in ottica di un incremento dei dati del dataset in seguito a nuove interazioni con il chatbot.

Per fare ciò, ho rivolto la mia attenzione ai modelli di linguaggio neurali, in particolare ai Large Language Models (LLM), dal momento che sono in grado di generalizzare su una vasta gamma di task linguistici, inclusa la classificazione di intenti.

Con l'enorme disponibilità attuale di modelli pre-addestrati e API che permettono di interagire con essi, ho potuto sperimentare diverse soluzioni per l'etichettatura automatica delle

domande.
In particolare, ho deciso di sperimentare con modelli di LLM open-source, dal momento che sono eseguibili localmente e permettono di mantenere i dati sensibili all'interno dell'ambiente di lavoro, senza doverli condividere con servizi esterni.
Per utilizzarli, si sono rivelate fondamentali le API fornite da Ollama [6], un sistema per hostare localmente modelli di LLM open source (e in certi casi anche *open-weights*).

2.2.2 Etichettatura automatica del dataset

Per poter automatizzare l'etichettatura usando una LLM, prima di tutto ho identificato l'insieme delle possibili etichette:

```
LABELS: dict[str, str] = {
  "START": "Initial greetings or meta-questions, such as 'hi' or 'hello'.",
  "GEN_INFO": "General questions about the automaton that don't focus on specific components
or functionalities.",
  "STATE_COUNT": "Questions asking about the number of states in the automaton.",
  "FINAL_STATE": "Questions about final states of the automaton.",
  "STATE_ID": "Questions about the identity of a particular state.",
  "TRANS_DETAIL": "General questions about the transitions within the automaton.",
  "SPEC_TRANS": "Specific questions about particular transitions or arcs between states.",
  "TRANS_BETWEEN": "Specific question about a transition between two states",
  "LOOPS": "Questions about loops or self-referencing transitions within the automaton.",
  "GRAMMAR": "Questions about the language or grammar recognized by the automaton.",
  "INPUT_QUERY": "Questions about the input or simulation of the automaton.",
  "OUTPUT_QUERY": "Questions specifically asking about the output of the automaton.",
  "IO_EXAMPLES": "Questions asking for examples of inputs and outputs.",
  "SHAPE_AUT": "Questions about the spatial or graphical representation of the automaton.",
  "OTHER": "Questions not related to the automaton or off-topic questions.",
  "ERROR_STATE": "Questions related to error states or failure conditions within the
automaton.",
  "START_END_STATE": "Questions about the initial or final states of the automaton.",
  "PATTERN_RECOG": "Questions that aim to identify patterns in the automaton's structure or
behavior.",
  "REPETITIVE_PAT": "Questions focusing on repetitive patterns, especially in transitions.",
  "OPT_REP": "Questions about the optimal spatial or minimal representation of the
automaton.",
  "EFFICIENCY": "Questions about the efficiency or minimal representation of the automaton."
}
```

Codice 2: Etichette possibili per le domande del dataset.

In questa mappa, ad ogni etichetta è associata una descrizione che indica alla LLM un contesto in cui collocarla, con lo scopo di assistere la LLM ad etichettare correttamente le domande togliendo il più possibile le ambiguità.
Questo genere di task è del tipo **zero shot**, in cui il modello non ha mai visto i dati di training e deve etichettare le domande esclusivamente in base a un contesto fornito.

Con lo scopo di assicurare un'etichettatura corretta e affidabile, ho deciso di utilizzare due modelli di LLM differenti, in modo da poter fare un majority voting tra le etichette prodotte dai due modelli:

- *Gemma 2*, sviluppato da Google Deep Mind [7];
- *llama 3.1*, sviluppato da Meta AI [8].

I modelli sono stati utilizzati nelle loro varianti da 9 miliardi di parametri per Gemma 2 (dimensione intermedia) e 8 miliardi per llama 3.1 (il più piccolo dei modelli forniti), basandomi

sulle sperimentazioni che hanno mostrato un buon compromesso tra performance (intese come qualità dei risultati prodotti in seguito al prompting) e tempo di esecuzione [7], [8].

Un ulteriore modello, Qwen [9], prodotto da Alibaba, è stato utilizzato durante le sperimentazioni, ma i risultati non sono stati sufficientemente soddisfacenti da permettere un utilizzo all'interno del progetto.

Ho effettuato il prompting delle domande con i modelli di LLM utilizzando le risorse dell'hardware a mia disposizione, composto da:

- CPU AMD Ryzen 7 5800x (4.7GHz, 8 core, 16 thread, 32MB L3 cache)
- 64GB RAM DDR4 @3200MHz
- GPU Nvidia RTX 3070 Ti (8GB GDDR6, 6144 CUDA cores @1.77GHz)

Ad ogni modello è richiesto di etichettare ogni domanda. Il prompt utilizzato è stato progettato in modo da fornire un contesto chiaro e preciso, in modo da guidare la LLM verso l'etichetta corretta.
In particolare, ne sono stati utilizzati due per ogni modello, in modo da fornire un contesto più vario e permettere ai modelli di generalizzare meglio sulle domande. Ogni prompt risulta diverso dal punto di vista della composizione della richiesta, ma l'intento finale a livello semantico è lo stesso.

I prompt sono stati scelti in modo da fornire informazioni utili ai modelli per etichettare le domande, insieme ad un contesto che effettivamente faccia comprendere alla LLM quale sia l'argomento della domanda:

```
prompts = [
  # First prompt
  """You are going to be provided a series of interactions from a user regarding questions
about finite state automatons.
  Each message has to be labelled, according to the following labels:

  {labels}

  You only need to answer with the corresponding label you've identified.
  Do not explain the reasoning, do not use different terms from the labels you've received
now.
  Interaction:
  {text}
  Label:
  """,

  # Second prompt
  """You are an AI assistant trained to classify questions into the following categories:

  {labels}

  Please classify the following question:
  {text}
  Category:
  """
]
```

Codice 3: Prompt utilizzati per l'etichettatura delle domande.

Si notino le differenze tra i due prompt: il primo è più dettagliato e fornisce una spiegazione più approfondita delle etichette, mentre il secondo è più conciso e diretto.

I tag tra parentesi graffe vengono sostituiti con i valori attualmente in uso, in modo da rendere il prompt generico e riutilizzabile.

Segue un estratto di codice python che mostra come è stato effettuato il prompting. Viene importata una classe `Chat`, da me sviluppata, che permette di interagire con i modelli di LLM in modo più semplice, astraendo le API di ollama.

```python
from tqdm import tqdm
from chat_helper import Chat
import pandas as pd

# ollama_models = ["llama3.1:8b", "gemma:7b", "qwen:7b"]
ollama_models = ["gemma2:9b", "llama3.1:8b"]

# We are initializing a new dataframe with the same index as the original one
res_df = pd.DataFrame(index=df.index)

for model in ollama_models:
    chat = Chat(model=model)

    dataset_size = len(df)

    for p_i, prompt_version in enumerate(prompts):
        progress_bar = tqdm(
          total=dataset_size,
          desc=f"Asking {model} with prompt {p_i}", unit="rows"
        )

        for r_i, row in df.iterrows():
            text = row["Text"]

            prompt = prompts[0].replace("{text}", text)

            inferred_label = chat.interact(
              prompt,
              stream=True,
              print_output=False,
              use_in_context=False
            )
            inferred_label = inferred_label.strip().replace("'", "")

            res_df.at[r_i, f"{model} {p_i}"] = inferred_label
            progress_bar.update()

        print(progress_bar.format_dict["elapsed"])
        progress_bar.close()
```

Codice 4: Prompting delle domande con i modelli di LLM.

Ecco un esempio dei risultati dell'etichettatura del bronze dataset, in seguito al prompting con i modelli di LLM:

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	SPEC_TRANS	SPEC_TRANS	TRANS_BETWEEN	TRANS_BETWEEN
4	Please provide the interaction. : START	START	START	START
…	…	…	…	…
285	OPT_REP	OPT_REP	OPT_REP	OPT_REP
286	GRAMMAR	GRAMMAR	GRAMMAR	GRAMMAR
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

Tabella 1: Esempio di etichettatura delle domande del bronze dataset.

Come è possibile notare, i modelli hanno etichettato le domande in modo coerente tra di loro, ma non sempre con le etichette corrette.
In certi casi, le etichette sono state completamente sbagliate, e in altre occorrenze sono state prodotte risposte che o non sono presenti nel set di etichette fornito, o hanno ignorato il prompt fornito, fornendo risposte completamente estranee.

Come accennato, è stato adoperato un sistema di majority voting per combinare i risultati delle due LLM, in modo da ottenere un'etichettatura più affidabile:

```
from collections import Counter

def majority_vote(row: pd.Series):
    label_counts = Counter(row)
    majority_label = label_counts.most_common(1)[0][0]
    return majority_label
```

Codice 5: Funzione di majority voting per combinare le etichette.

Tuttavia, in seguito ad una prima fase di fine tuning, ho verificato che nonostante un'etichettatura valida, le classi identificate erano troppo sbilanciate, con alcune classi che contenevano un numero troppo esiguo di esempi. In più, ho realizzato che le classi scelte erano troppo generiche: questo problema non avrebbe permesso di identificare con precisione l'argomento della domanda.

Per questo motivo ho proceduto con una revisione delle etichette, e una successiva etichettatura manuale delle domande.

2.2.3 Nuove classi e etichettatura manuale

Prima di proseguire con l'etichettatura, ho provveduto a ripulire il dataset da domande non pertinenti o duplicate. Una volta fatto, ho deciso di ridurre il numero di classi, in modo da poter avere un dataset più bilanciato e con classi più specifiche.
Avendone ridotto il numero, per ottenere un livello di granularità maggiore, ho deciso di

utilizzare un sistema di etichettatura gerarchico, in modo da poter identificare con maggiore precisione l'argomento della domanda.
Ne sono risultati sono due livelli di classi:

- Le *classi principali* (o *question intent*, si veda la Tabella 2), che rappresentano l'argomento generale della domanda, per un totale di 7 classi;
- Le *classi secondarie*, che rappresentano l'argomento specifico della domanda, dipendono dalla classe principale e sono 33 in totale. A seconda della classe principale, il numero di classi secondarie varia.

Il numero ristretto di classi di domande ha permesso di creare una suddivisione più bilanciata tra le classi, e di ottenere un dataset generalmente più equilibrato.

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 2: Classi principali del dataset.

Come è possibile notare dalle tabelle che seguono, alcune classi secondarie contengono un numero esiguo di esempi, non sufficiente per una classificazione affidabile.

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 3: Le 6 classi secondarie del dataset per la classe primaria dell'**automa**.

Sottoclassi	Scopo	Numero di Esempi
count	Numero di transizioni	10
cycles	Domande riguardo anelli tra nodi	4
description	Descrizioni generali sugli archi	2
existence_between	Esistenza di un arco tra due nodi	12
existence_directed	Esistenza di un arco da un nodo a un altro	9
existence_from	Esistenza di un arco uscente da un nodo	18
existence_into	Esistenza di un arco entrante in un nodo	1
input	Ricezione di un input da parte di un nodo	1
label	Indicazione di quali archi hanno una certa etichetta	4
list	Elenco generico degli archi	15
self_loop	Esistenza di self-cycles	1

Tabella 4: Le 11 classi secondarie del dataset per la classe primaria delle **transizioni**.

Sottoclassi	Scopo	Numero di Esempi
count	Numero di stati	19
details	Dettagli specifici su uno stato	1
list	Elenco generale degli stati	1
start	Qual è lo stato iniziale	8
final	Esistenza di uno stato finale	7
final_count	Numero di stati finali	2
final_list	Elenco degli stati finali	3
transitions	Connessioni tra gli stati	8

Tabella 5: Le 8 classi secondarie del dataset per la classe primaria degli **stati**.

Sottoclassi	Scopo	Numero di Esempi
accepted	Grammatica accettata dall'automa	14
example_input	Input di esempio accettato dall'automa	4
regex	Regular expression corrispondente all'automa	2
simulation	Simulazione dell'automa con input dell'utente	8
symbols	Simboli accettati dalla grammatica	7
validity	Validità di un input fornito	2
variation	Richiesta di simulazione su un automa modificato	2

Tabella 6: Le 7 classi secondarie del dataset per la classe primaria della **grammatica**.

2.2.4 Data Augmentation

Come evidenziato nella sezione precedente, diverse classi secondarie contengono un numero esiguo di esempi, non sufficiente per una buona classificazione in seguito al fine-tuning.

Avendo solo 229 esempi, ho arricchito i dati con ulteriori domande scritte manualmente e anche generate artificialmente.

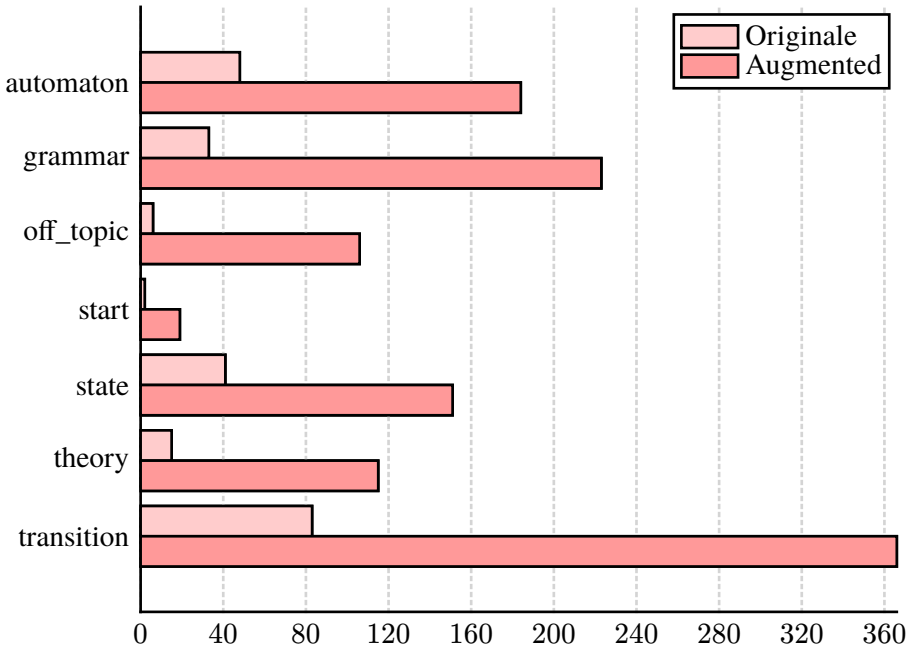
Le domande artificiali sono state prodotte in grandi quantità adoperando diversi modelli disponibili online e locali, tra cui:

- ChatGPT `4o`, `o1` e `o3-mini` [10];
- Llama3.1 [8];
- DeepSeek R1 [11], [12].

Ad ogni modello è stato presentato un insieme di domande con lo stesso topic principale o secondario, assieme al contesto in cui vengono poste e ad una richiesta di produzione di ulteriori domande simili semanticamente. Per maggiore convenienza, è stato richiesto ai modelli di rispondere fornendo le nuove domande formattate in markdown [13].

Dato il grosso volume di risposte, per verificare l'adesione dei modelli alle richieste è stato effettuato un controllo a campione, che non ha evidenziato particolari problematiche nella precisione di nessuno dei modelli.

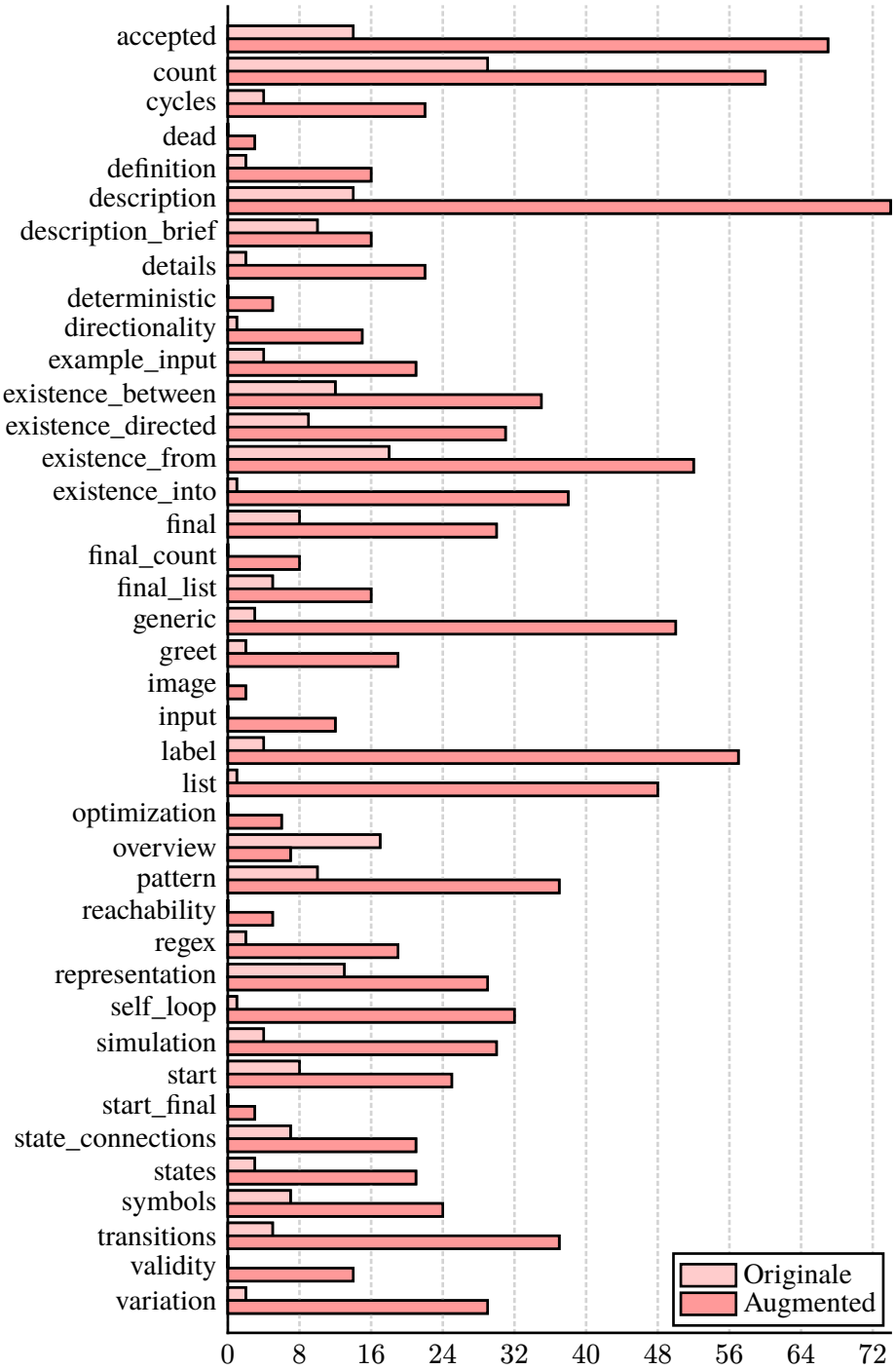
In totale sono stati aggiunti 851 nuovi quesiti, con la seguente distribuzione:



Le domande off-topic aggiuntive sono state estratte dal dataset SQUAD[3] v2 [14], [15], per avere una sufficiente varietà di domande non pertinenti.

Anche le classi secondarie hanno ricevuto alcune migliorie alla distribuzione, che rimane comunque ancora sbilanciata, com'è possibile vedere nella Figura 2:

[3]Stanford Question Answering Dataset



Nonostante lo sbilanciamento, è stato possibile ottenere dei buoni risultati in seguito al fine-tuning.

L'utilizzo del dataset SQUAD ha anche introdotto un'ulteriore incremento delle performance, portando a una diminuzione dell'erronea classificazione di esempi off-topic come domande lecite. In particolare, le metriche di entropia e confidenza durante il fine tuning sono migliorate rispettivamente del 17 e del 7%.

2.2.5 Fine-tuning

Per poter utilizzare le Large Language Models (LLM) per la classificazione di intenti, ho dovuto seguire un processo di fine-tuning.

Il fine-tuning avviene verso la fine della preparazione di un modello di machine learning. In particolare, è la fase in cui si prende un modello pre-addestrato su un compito generale (o su una grande quantità di dati non etichettati) e lo si "specializza" su un compito specifico, come la classificazione di intenti, l'analisi del sentiment o il riconoscimento di entità nominate.
Si parte quindi da un modello che possiede già una buona conoscenza linguistica di base (perché allenato, ad esempio, su quantità imponenti di testo come Wikipedia, libri o pubblicazioni) e lo si ri-addestra su un dataset mirato, così da fargli apprendere le particolarità e le sfumature del nuovo scenario applicativo.

Sul piano tecnico, il processo di fine-tuning si fonda sugli stessi principi del *learning by example*: si forniscono al modello coppie di input e output (nel caso di una classificazione, l'output è la classe corretta), e si calcola la loss (ad esempio la cross-entropy tra le probabilità previste dal modello e quelle desiderate).
Tramite la *backpropagation* dell'errore, i pesi del modello vengono aggiornati iterativamente, così da allineare le predizioni alle etichette reali. Il risultato è che, dopo un numero sufficiente di iterazioni (o epoche), il modello impara a predire con buona approssimazione la classe corretta anche per esempi non ancora visti.

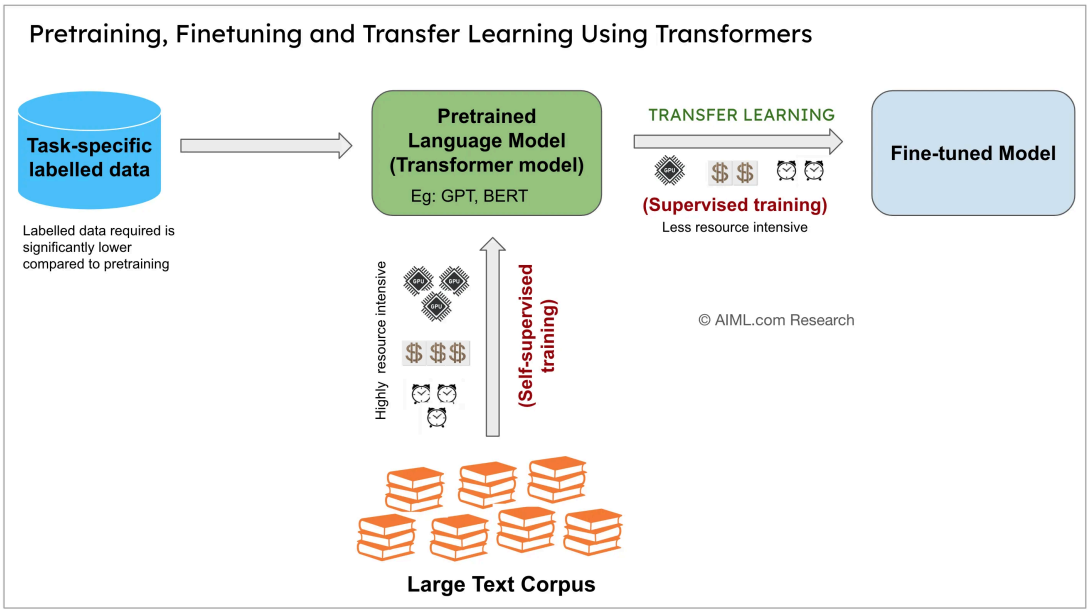


Figura 3: Processo di fine-tuning di un modello di LLM. **IMMAGINE DA SOSTITUIRE**

L'elemento distintivo del fine-tuning rispetto a un addestramento "da zero" (o from scratch) sta nel fatto che la maggior parte dei pesi del modello non parte da valori iniziali casuali, bensì da un punto in cui il modello ha già "appreso" molte regole e pattern del linguaggio. Se nel pre-addestramento ha appreso, ad esempio, la nozione di contesto, la correlazione fra parole vicine e la loro valenza semantica, durante il fine-tuning deve semplicemente specializzarsi nel riconoscere come queste informazioni si combinano per risolvere il compito target. Questo

riduce drasticamente la quantità di dati e di risorse computazionali necessarie a raggiungere buone prestazioni.

Nel caso di una classificazione testuale multi-classe, si aggiunge in genere un piccolo strato di output (o head) in cima al modello pre-addestrato. La testa è una semplice rete feed-forward, spesso costituita da uno o due livelli di neuroni, che produce un vettore di dimensione pari al numero di possibili etichette. Il resto del modello rimane pressoché invariato: l'architettura interna, come i vari encoder o layer del Transformer, resta la stessa, ma i loro pesi continuano ad aggiornarsi durante il training, almeno in un contesto standard (è anche possibile, in alcuni scenari, "congelare" i primi strati e addestrare solo quelli finali, in base a considerazioni di efficienza e dimensione del dataset).

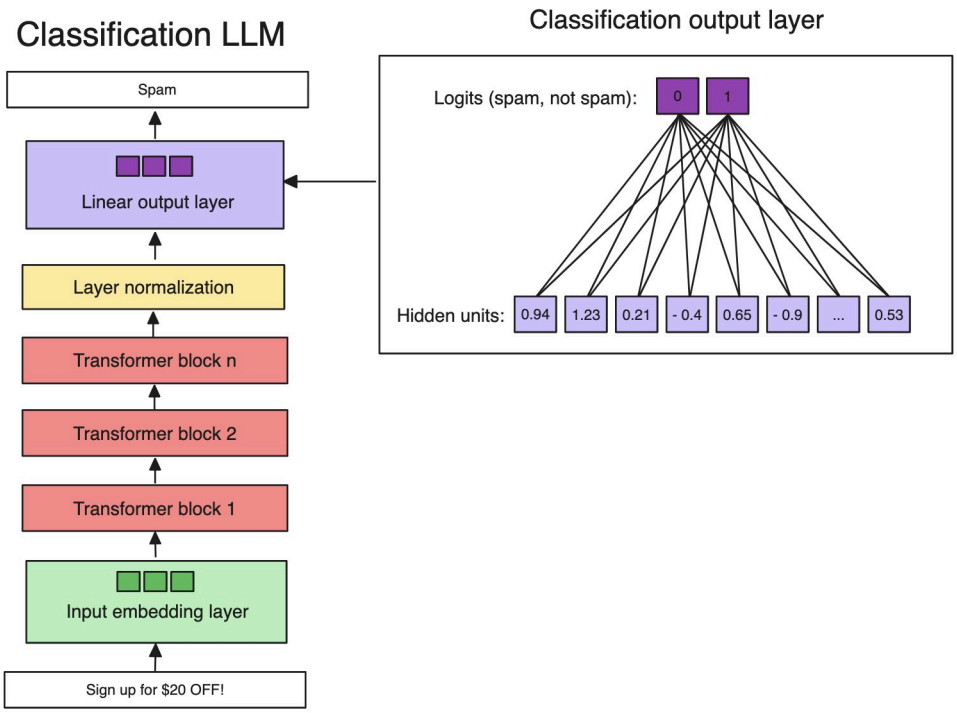


Figura 4: Struttura di un modello di classificazione basato su LLM. **IMMAGINE DA SOSTITUIRE**

2.2.6 BERT

Prima di illustrare più nel dettaglio il fine-tuning, è utile introdurre BERT[4] [16], progenitore della famiglia di modelli da me utilizzati per la sperimentazione, nonchè uno dei modelli più noti e influenti degli ultimi anni in ambito Natural Language Processing.

BERT è stato proposto nel 2018 da J. Devlin, M.-W. Chang, K. Lee, e K. Toutanova [16] come un sistema capace di apprendere rappresentazioni contestuali del testo in modo bidirezionale, basandosi sull'architettura Transformer introdotta in precedenza da A. Vaswani *et al.* [17].

L'idea portante di BERT è quella di addestrare un modello neurale a predire, data una sequenza testuale, le parole mascherate (ovvero rimosse o sostituite) e la relazione tra

[4]Bidirectional Encoder Representations from Transformers

frasi adiacenti. Queste due tecniche di pre-addestramento vengono rispettivamente chiamate Masked Language Modeling e Next Sentence Prediction.

Nel Masked Language Modeling, BERT maschera casualmente alcune parole del testo in input e chiede al modello di indovinare quali fossero, costringendolo così a sviluppare una comprensione profonda del contesto circostante.

Nel Next Sentence Prediction, invece, il modello riceve in ingresso due frasi (A e B) e impara a classificare se B segue effettivamente A o se le due frasi appartengono a contesti disgiunti. Addestrando in parallelo su questi due compiti, BERT acquisisce rappresentazioni interne che colgono sfumature sintattiche, semantiche e relazionali del linguaggio [16].

Una volta pre-addestrato su grandi corpora di testo (come Wikipedia ed estrazioni di libri), BERT può essere facilmente "specializzato" per vari task supervisionati, tra cui la classificazione di testi, l'analisi del sentiment, il question answering e, in generale, tutto ciò che riguarda la comprensione del linguaggio naturale, essendo un modello encoder. La peculiarità di BERT è che, essendo già addestrato a livello linguistico di base, necessita di meno esempi per ottenere risultati spesso notevoli su compiti altamente specializzati.

Esistono diverse varianti del modello, in termini di dimensioni e capacità. Le versioni più comuni sono `BERT-base` e `BERT-large`, differenziate per numero di livelli (encoder) e di parametri totali.
In generale, la versione `base` è più rapida e ha requisiti meno elevati in termini di memoria, mentre la versione large offre performance maggiori a fronte di tempi di calcolo e requisiti hardware superiori.

Nella libreria di Huggingface `transformers` [18], BERT è messo a disposizione come un modello pretrained, pronto per essere caricato e ulteriormente addestrato. In un contesto di classificazione di intenti, ad esempio, si può utilizzare `AutoModelForSequenceClassification` specificando il checkpoint "bert-base-uncased" (o simili).

Un esempio di codice di inizializzazione è il seguente:

```
from transformers import AutoModelForSequenceClassification, AutoTokenizer

model_name = "bert-base-uncased"
tokenizer = AutoTokenizer.from_pretrained(model_name)
model = AutoModelForSequenceClassification.from_pretrained(model_name,
num_labels=num_classes)
```

`model` è in grado di elaborare sequenze di token generate dal tokenizer e, una volta fine-tuned, produce come output le probabilità di appartenere alle varie classi (o intenti) da classificare. Questa è la base su cui mi sono appoggiato per la classificazione delle domande del dataset.

2.2.7 Implementazione

In questa sezione sarà presentata la procedura di fine-tuning che ho implementato per addestrare un modello di classificazione di intenti basato su architetture Transformer.
L'intero processo sfrutta principalmente la libreria `transformers` di Huggingface [18],

in combinazione con altri strumenti sempre dell'ecosistema FOSS[5] di Huggingface, come `datasets`.

L'utilizzo di queste librerie permette di semplificare notevolmente il processo di fine-tuning, fornendo API intuitive e funzionalità di alto livello per la gestione dei dati, la creazione dei modelli e la valutazione delle performance. In questo modo è possibile addestrare un modello di classificazione di intenti in poche righe di codice, senza dover scrivere manualmente i loop di training e validation, o implementare da zero la logica di salvataggio e caricamento dei modelli, nonostante questa via sia sempre possibile.

L'obiettivo è utilizzare un modello pre-addestrato (ad esempio BERT, DistilBERT o qualsiasi altro compatibile con `AutoModelForSequenceClassification`) con lo scopo di specializzarlo nel riconoscimento di specifiche categorie di intenti, e successivamente salvarlo per l'uso nel chatbot.

Preparazione dei dati

Un primo punto cruciale è la preparazione del dataset, gestita dalla funzione `prepare_dataset`. Qui effettuo la suddivisione stratificata tra train e validation, tokenizzo i testi tramite un `AutoTokenizer` e converto le etichette da stringhe a interi, in accordo con la mappatura definita nella classe `LabelInfo`[6].

```
def prepare_dataset(df: DataFrame,
                    tokenizer: PreTrainedTokenizer,
                    label_info: LabelInfo,
                    examples_column: str,
                    labels_column: str) -> tuple[Dataset, Dataset]:
  """
  Prepares the dataset for training and evaluation by tokenizing the text
  and encoding the labels.
  """
  def tokenize_and_label(example: dict) -> BatchEncoding:
      question = example[examples_column]
      encodings = tokenizer(question, padding="max_length", truncation=True, max_length=128)
      label = label_info.get_id(example[labels_column])
      encodings.update({'labels': label})
      return encodings

  split = StratifiedShuffleSplit(n_splits=1, test_size=0.2, random_state=42)
  train_index, val_index = next(split.split(df, df[labels_column]))
  strat_train_set = df.iloc[train_index].reset_index(drop=True)
  strat_val_set = df.iloc[val_index].reset_index(drop=True)

  train_dataset = Dataset.from_pandas(strat_train_set)
  eval_dataset = Dataset.from_pandas(strat_val_set)

  train_dataset = train_dataset.map(tokenize_and_label,
remove_columns=train_dataset.column_names)
  eval_dataset = eval_dataset.map(tokenize_and_label,
remove_columns=eval_dataset.column_names)

  return train_dataset, eval_dataset
```

Codice 6: Funzione per la preparazione del dataset.

[5]Free and Open Source Software, cioè Software **Libero** e Open Source

[6]Si veda l'appendice per la completa definizione.

In questo modo, ottengo due oggetti di tipo `Dataset` che rappresentano il training set e il validation set. Ciascun esempio è stato trasformato in una struttura pronta per essere gestita dal `Trainer` di Huggingface, con un campo `labels` che indica la classe corretta da apprendere.

Una volta create e preparate queste componenti (funzione di metriche, funzioni di training, dataset tokenizzato), eseguo il fine-tuning chiamando `run_fine_tuning` (presentata poco più avanti).

Metriche di valutazione

Per prima cosa, ho definito una funzione in grado di calcolare le metriche di valutazione, che permetteranno di valutare le performance del modello in fase di fine-tuning in modo automatico.

Ho scelto di considerare **accuratezza**, **precision**, **recall** e **F1** come indicatori classici di performance; in aggiunta, calcolo anche l'**entropia media** e la **confidenza media**, allo scopo di misurare rispettivamente il grado di incertezza delle previsioni e la probabilità media associata alla classe predetta. Lo snippet seguente mostra la funzione `compute_metrics`:

```
def compute_metrics(eval_pred):
  """
  Compute evaluation metrics for the model predictions.
  """
  predictions, labels = eval_pred
  probabilities = np.exp(predictions) / np.sum(np.exp(predictions), axis=1, keepdims=True)
  preds = np.argmax(probabilities, axis=1)

  acc = accuracy_score(labels, preds)
  precision, recall, f1, _ = precision_recall_fscore_support(labels, preds,
average='weighted', zero_division=0)

  entropies = entropy(probabilities.T)
  avg_entropy = np.mean(entropies)
  avg_confidence = np.mean(np.max(probabilities, axis=1))

  metrics = {
      'accuracy': acc,
      'precision': precision,
      'recall': recall,
      'f1': f1,
      'avg_entropy': avg_entropy,
      'avg_confidence': avg_confidence,
  }
  return metrics
```

Codice 7: Funzione per il calcolo delle metriche di valutazione.

Può essere utile soffermarci un momento a spiegare le metriche scelte:

L'**accuratezza** (o tasso di classificazione corretta) misura la proporzione di esempi classificati correttamente, senza distinzione tra le varie classi. Formalmente:

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N}\{\hat{y}_i = y_i\} \tag{1}$$

dove $\{\hat{y}_i = y_i\}$ vale 1 se la previsione è corretta, 0 altrimenti. Più il valore è vicino a 1, migliore è la performance complessiva del modello.

Quando si lavora con problemi di classificazione con etichette binarie, o si valuta ciascuna classe indipendentemente, esistono alcuni conteggi che possono essere utili per valutare la qualità delle previsioni:

- i **true positives** (TP) indicano i casi in cui il modello ha predetto correttamente la classe positiva;
- i **false positives** (FP) indicano i casi previsti come positivi dal modello, ma che in realtà sono negativi;
- i **false negatives** (FN) i casi previsti negativi ma in realtà positivi.

Sulla base di queste definizioni, si introducono due metriche fondamentali:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2}$$

che indica la percentuale di esempi classificati come positivi che erano effettivamente positivi.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

stima la quota di esempi positivi che sono stati effettivamente riconosciuti come tali dal modello.

L'**F1-score** fornisce una media armonica fra Precision e Recall, combinando entrambe le metriche in un singolo indice:

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

Un F1 score alto richiede che entrambe le metriche siano elevate; se una delle due è bassa, il valore di F1 tende drasticamente a ridursi. Questo lo rende particolarmente utile in casi di class imbalance o quando è importante non trascurare né la precisione né la capacità di recuperare tutti i positivi.

L'**entropia** è una misura della disordine o incertezza di un sistema, in questo caso delle previsioni del modello.

Per un singolo esempio, se il modello produce una distribuzione di probabilità $\boldsymbol{p}_i = \big(p_{i,1}, ..., p_{i,k}\big)$ sulle k classi, è possibile calcolare l'entropia dell'esempio come

$$\text{H}(\boldsymbol{p}_i) = -\sum_{j=1}^{k} p_{i,j} \log\big(p_{i,j}\big) \tag{5}$$

Tale quantità esprime quanto "incerte" sono le previsioni del modello: se il modello assegna un'alta probabilità a una sola classe e bassa probabilità alle altre, l'entropia tende a essere prossima a zero (predizione più "sicura"); se distribuisce le probabilità in modo pressoché uniforme, l'entropia aumenta (maggiore incertezza).

L'entropia media su tutto il set di validazione di dimensione N è:

$$\text{Average Entropy} = \frac{1}{N} \sum_{i=1}^{N} H(p_i) \tag{6}$$

Un valore basso di entropia media indica che, in media, le previsioni del modello sono piuttosto concentrate su una specifica classe; un valore più alto suggerisce che il modello sia spesso incerto.

Sempre definita a partire dalla distribuzione $\boldsymbol{p}_i$, la confidenza per il singolo esempio i può essere definita come la probabilità associata alla classe di output che ha la confidenza massima:

$$C(\boldsymbol{p}_i) = \max_j p_{i,j} \tag{7}$$

Maggiore è il valore di $C(\boldsymbol{p}_i)$, più il modello risulta "sicuro" di quella predizione. Analogamente, la confidenza media sul dataset si calcola come:

$$\text{Average Confidence} = \frac{1}{N} \sum_{i=1}^{N} \max_j p_{i,j} \tag{8}$$

Un valore prossimo a 1 indica che, spesso, il modello prende decisioni molto nette; un valore più basso può rivelare maggiore cautela o incertezza.

Usata congiuntamente all'entropia media, la confidenza media può fornire indicazioni interessanti su come il modello pesa le varie classi e quanto tende a "sbilanciarsi" sulle previsioni.

Addestramento

Per effettuare l'addestramento vero e proprio, ho definito anche la funzione `run_fine_tuning`, che si fa carico di gestire i parametri di training (come numero di epoche, learning rate, batch size), di configurare gli strumenti di logging e salvataggio, e di lanciare effettivamente il training tramite la classe `Trainer` della libreria `transformers`.

La classe `Trainer` semplifica notevolmente la gestione di molteplici aspetti, come la schedulazione del learning rate o la stratificazione della validazione.

Il metodo espone diversi parametri significativi:

- `load_best_model_at_end=True` consente di caricare automaticamente al termine dell'addestramento i pesi del modello con il miglior valore di F1 (impostato in metric_for_best_model='f1');
- `warmup_ratio=0.1` configura un periodo iniziale di warm-up, durante il quale il learning rate cresce gradualmente prima di stabilizzarsi nella fase successiva. Questo contribuisce a rendere l'ottimizzazione più stabile ed evitare picchi di aggiornamento eccessivi nelle primissime iterazioni.
  La configurazione del warmup, assieme alla learning rate sono state scelte basandomi sull'utilissimo paper di M. Mosbach, M. Andriushchenko, e D. Klakow [19] che fornisce una guida pratica per il fine-tuning di BERT.
- `metric_for_best_model='f1'` indica che il modello migliore sarà scelto in base al valore di F1, calcolato dalla funzione `compute_metrics`. F1 torna utile in quanto è in grado di bilanciare le due metriche di precision e recall, fornendo un'indicazione complessiva delle performance del modello.

Un'ultima considerazione molto importante riguarda il parametro `report_to`, che consente di specificare a quali servizi di logging inviare i risultati del training.

Nel mio caso, ho scelto di fare affidamento a **Weights and Biases**[7] in modalità online, in modo da poter monitorare in tempo reale le performance del modello durante il fine-tuning.

```python
def run_fine_tuning(model: AutoModelForSequenceClassification,
                    tokenizer: AutoTokenizer,
                    train_dataset: Dataset,
                    eval_dataset: Dataset,
                    wandb_mode: str,
                    num_train_epochs=20) -> Trainer:
  """
  Fine-tunes a pre-trained model on the provided training dataset and evaluates it on the
evaluation dataset.
  """
  report_to = ["wandb"] if wandb_mode == "online" else None

  training_args = TrainingArguments(
      output_dir='./temp',  # Directory to save the model and other outputs
      num_train_epochs=num_train_epochs,  # Number of training epochs
      learning_rate=2e-5,  # Learning rate for the optimizer
      warmup_ratio=0.1,  # Warmup for the first 10% of steps
      lr_scheduler_type='linear',  # Linear scheduler
      per_device_train_batch_size=16,  # Batch size for training
      per_device_eval_batch_size=16,  # Batch size for evaluation
      save_strategy='epoch',  # Save the model at the end of each epoch
      logging_strategy='epoch',  # Log metrics at the end of each epoch
      eval_strategy='epoch',  # Evaluate the model at the end of each epoch
      logging_dir='./temp/logs',  # Directory to save the logs
      load_best_model_at_end=True,  # Load the best model at the end by evaluation metric
      metric_for_best_model='f1',  # Use subtopic F1-score to determine the best model
      greater_is_better=True,  # Higher metric indicates a better model
      save_total_limit=1,  # Limit the total number of saved models
      save_only_model=True,  # Save only the model weights
      report_to=report_to,  # Report logs to Wandb if mode is "online"
  )

  trainer = Trainer(
      model=model,  # The model to be trained
      args=training_args,  # Training arguments
      train_dataset=train_dataset,  # Training dataset
      eval_dataset=eval_dataset,  # Evaluation dataset
      processing_class=tokenizer,  # Tokenizer for processing the data
      compute_metrics=compute_metrics  # Function to compute evaluation metrics
  )

  print(f"Trainer is using device: {trainer.args.device}")

  trainer.train()  # Start the training process

  return trainer
```

La quasi totalità dei dati mostrati in questo documento sono stati raccolti tramite Wandb, riducendo enormemente il tempo necessario per l'analisi e la visualizzazione dei risultati: il salvataggio automatico ad ogni run e la possibilità di confrontare run diversi in un'unica dashboard sono state funzionalità fondamentali per la mia sperimentazione.

[7] Weights and Biases, abbreviato `Wandb`, è un servizio di monitoraggio e logging per l'addestramento di modelli di machine learning

Modelli utilizzati

Tutti i modelli che ho utilizzato per la sperimentazione sono basati su BERT, o ELECTRA [20], entrambi fondati sull'architettura encoder [16].

In particolare, dal repository di Huggingface dedicato ai modelli di classificazione ho deciso di utilizzare:

- `google-bert/bert-base-uncased`, versione da 110 milioni di parametri [21]. Si tratta del modello originale di BERT ideato da Google [16];
- `distilbert/distilbert-base-uncased` [22], versione distillata [23] di BERT, con circa il 40% in meno di parametri [24]. Il modello è il risultato di una operazione dove si addestra un modello più piccolo ad imitare al meglio l'originale;
- `google/mobilebert-uncased` [25], versione di BERT ingegnerizzata con lo scopo di essere eseguibile su dispositivi mobili. Ha un totale di 25 milioni di parametri [26].
- `google/electra-small-discriminator` [27], da 14 milioni di parametri. Questo modello è stato addestrato utilizzando tecniche simili a quelle utilizzate per addestrare le GAN[8] [20], [28]

Tutti questi modelli sono direttamente adoperabili per i nostri scopi essendo modelli encoder: dato un certo input produrranno una rappresentazione vettoriale o matriciale. Il risultato è successivamente classificabile da una rete feed-forward, restituendo così come risultato la classe più probabile (si veda la Sezione 2.2.5).

Sono state effettuate anche delle sperimentazioni con una variante della normale architettura, dove su un unico encoder vengono addestrati due modelli separati di classificazione, per riconoscere con un'unica esecuzione del modello entrambe le classi della domanda presentata.

L'idea, già utilizzata anche in altri ambiti per il Transfer Learning [29] o direttamente su BERT [30], [31] può permettere di ridurre notevolmente il costo e i tempi di addestramento, oltre ai requisiti di memoria. Infatti, avendo la quasi totalità dei pesi concentrati nei layer del transformer, lo strato finale di classificazione risulta molto "sottile", e richiede una percentuale minima rispetto al resto del modello.

Nel mio caso sfortunatamente l'architettura a doppia testa di classificazione non si è rivelato migliore, con performance in media inferiori del 20% rispetto al miglior modello addestrato finora. Nonostante le performance peggiori, l'utilizzo di un modello del genere può essere considerato in contesti soggetti da forti limiti hardware, come su dispositivi mobili, edge o low-end.

L'intera implementazione fa nuovamente fondamento sull'enorme flessibilità della libreria `transformers`. È stato sufficiente infatti soltanto aggiungere le due classification heads ed estendere il metodo `forward` che si occupa della predizione:

[8] Generative Adversarial Networks, modelli addestrati in coppia, dove uno impara a svolgere un certo compito generativo, e l'altro a riconoscere se un certo esempio presentato è generato o meno.

```python
from torch import nn as nn
from transformers import BertPreTrainedModel, BertModel


class BertForHierarchicalClassification(BertPreTrainedModel):
    def __init__(self, config, num_main_topics, num_subtopics):
        super().__init__(config)
        self.bert = BertModel(config)
        self.classifier_main = nn.Linear(config.hidden_size, num_main_topics)
        self.classifier_sub = nn.Linear(config.hidden_size, num_subtopics)
        self.init_weights()

    def forward(self, input_ids, attention_mask, labels_main=None, labels_sub=None):
        outputs = self.bert(input_ids=input_ids, attention_mask=attention_mask)
        pooled_output = outputs.pooler_output
        logits_main = self.classifier_main(pooled_output)
        logits_sub = self.classifier_sub(pooled_output)

        loss = None
        if labels_main is not None and labels_sub is not None:
            loss_fct = nn.CrossEntropyLoss()
            loss_main = loss_fct(logits_main, labels_main)
            loss_sub = loss_fct(logits_sub, labels_sub)
            loss = loss_main + loss_sub  # Adjust weighting if needed

        return {'loss': loss, 'logits_main': logits_main, 'logits_sub': logits_sub}
```

2.2.8 Valutazione e performance

Come spiegato nella Sezione 2.2.7.2, per compiere l'addestramento dei modelli è stato essenziale sfruttare metriche di valutazione adeguate, in grado di fornire un quadro completo delle performance del modello.

Iniziamo quindi ad osservare i risultati dell'addestramento sulla classe principale del dataset:

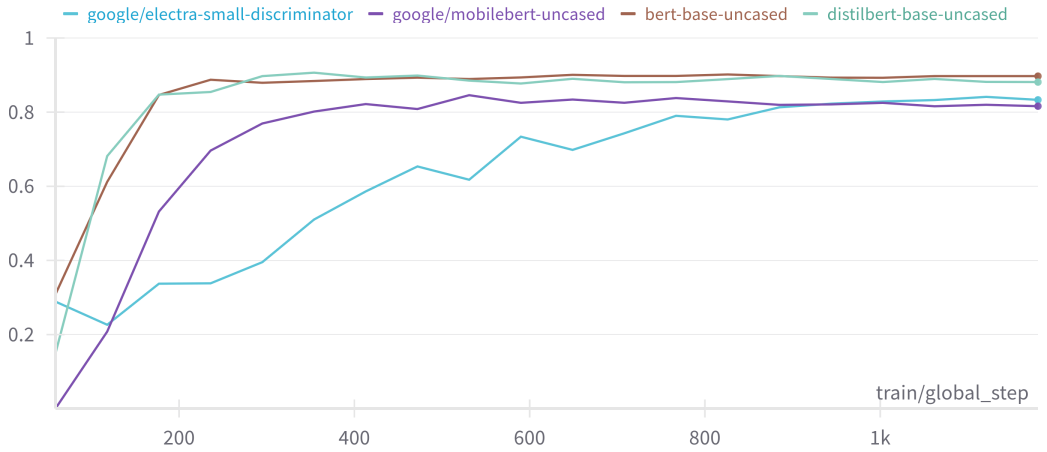


Figura 5: Confronto delle performance di F1 tra i modelli addestrati.

Come possiamo osservare, i modelli `bert` e `distilbert` hanno performance pressochè identiche (la differenza è dello 0.01%), mentre i modelli `mobilebert` e `electra` differiscono di circa l'8% rispetto a `bert`.

Le differenze di performance sono sempre da confrontare considerando anche il tempo di addestramento e la complessità del modello: `electra` ad esempio, pur avendo performance leggermente inferiori, è stato addestrato in meno della metà del tempo rispetto a `bert`.

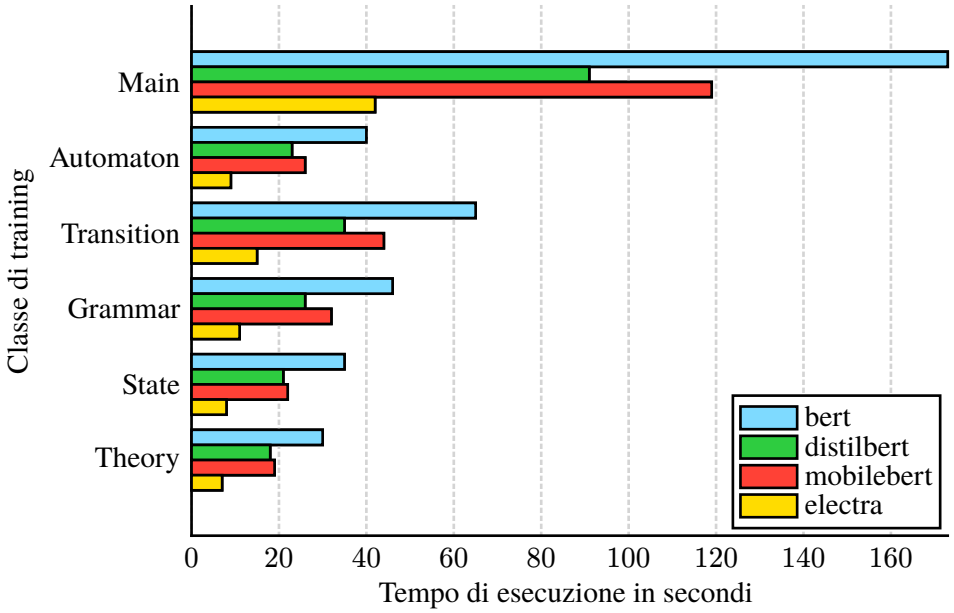


Figura 6: Confronto dei tempi di esecuzione per ciascuna classe di training.

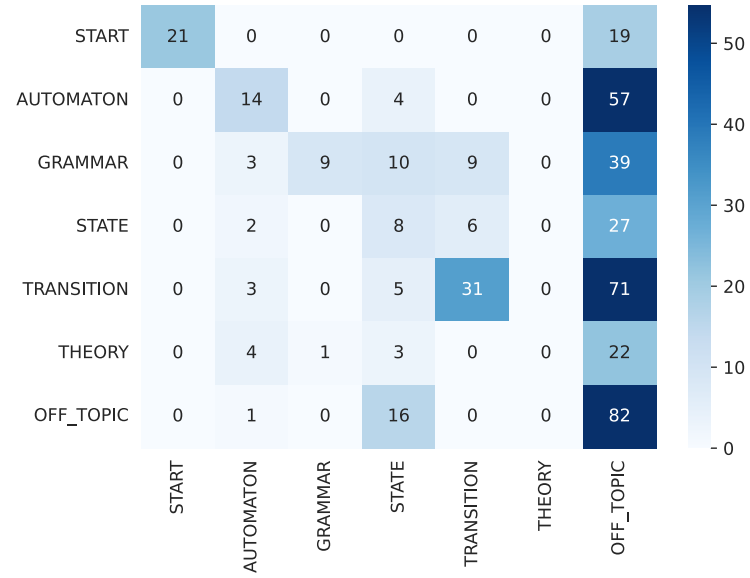
Questo salto nei tempi di addestramento così brusco in realtà si rivelerà essere un problema, come vedremo poco più avanti.

Da un punto di vista qualitativo, iniziamo a osservare le performance di AIML, che useremo come baseline di riferimento per il confronto con gli altri modelli neurali.
Per poterlo fare, sfrutteremo le matrici di confusione per valutare le performance dei modelli, in particolare per osservare come si comportano in presenza di classi sbilanciate o di domande ambigue.[9]

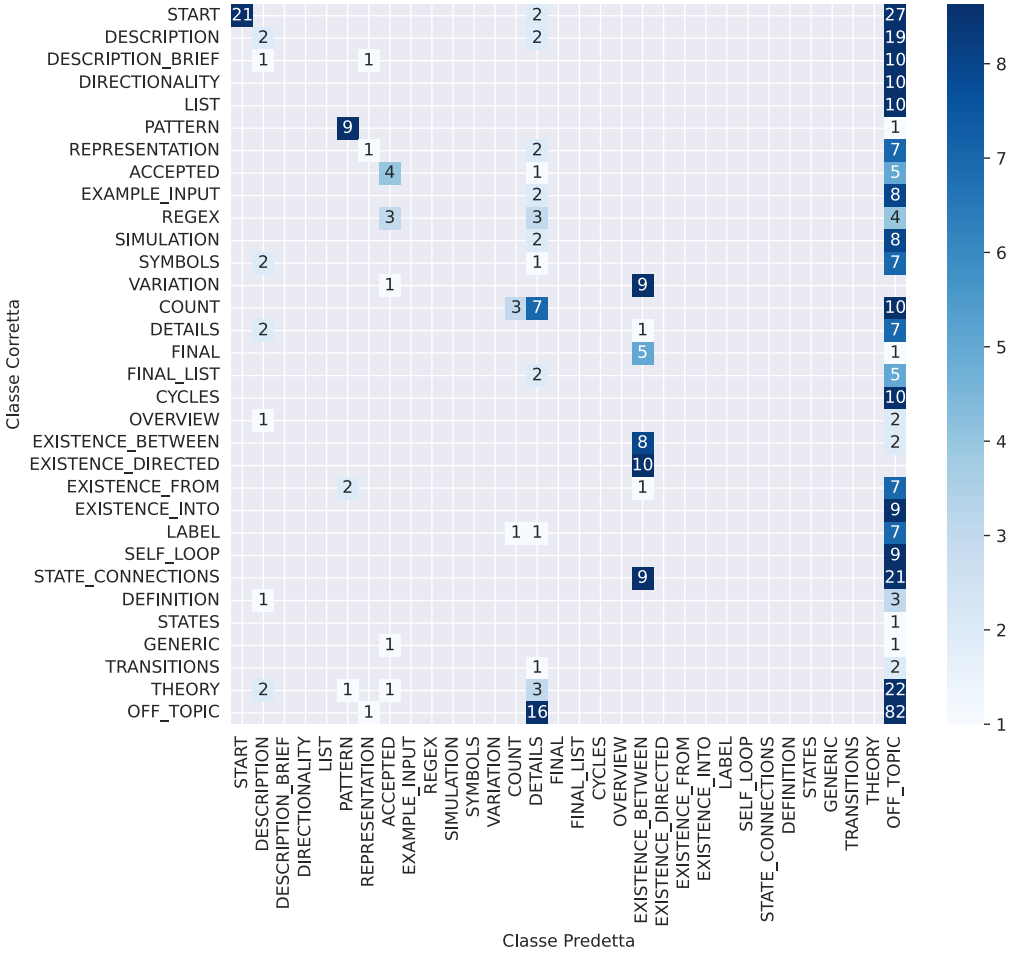
Tutte le valutazioni qualitative sono effettuate utilizzando un ulteriore dataset di test, separato dal dataset di training e di validazione, per evitare overfitting e garantire una valutazione imparziale. È composto da 468 ulteriori domande, distribuite in modo da assicurare una verifica sufficiente su tutte le classi di intenti secondarie, cruciali per la corretta classificazione e per fornire effettivamente risposte utili agli utenti.

Il modello, nonostante sia costituito da un numero non indifferente di regole e pattern (103), ha performance mediamente basse, con un F1 score medio del 33%. Possiamo anche vedere come, dove questo non è in grado di classificare una certa domanda, finisca col classificarla come off-topic, indicando una certa difficoltà nel riconoscere domande in realtà valide per il nostro dominio.

[9]Una matrice di confusione è una tabella che mostra il numero di predizioni corrette e incorrette fatte dal modello, confrontando le predizioni con le etichette reali.



Possiamo estendere queste affermazioni anche alle classi di intenti secondarie, dove il modello mostra un F1 score medio del 9%:



2.3 Riconoscimento delle entità

2.3.1 NER e Slot-filling

2.3.2 Spacy

2.3.3 Valutazione e performance

3 Natural Language Generation

3.1 Generazione di risposte tramite LLM

3.1.1 Parafrasi

3.1.2 Prompting

3.2 Data Retrieval

3.2.1 Retrieval tramite query

3.2.2 Retrieval basato su script

3.2.3 Retrieval automatico guidato dagli LLM

3.3 Qualità delle risposte

3.3.1 Valutazione automatica

3.3.2 Valutazione umana

4 Ingegnerizzazione

4.1 Composizione del sistema

4.2 Compilatore

4.2.1 Pipeline

4.3 Runner

Bibliografia

[1] «The Loebner Prize». [Online]. Disponibile su: https://www.ocf.berkeley.edu/~arihuang/academic/research/loebner.html

[2] A. M. TURING, «I.—COMPUTING MACHINERY AND INTELLIGENCE», *Mind*, vol. 59, fasc. 236, pp. 433–460, 1950, doi: 10.1093/mind/LIX.236.433.

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

[4] R. Wallace, «The anatomy of A.L.I.C.E», 2009, pp. 181–210. doi: 10.1007/978-1-4020-6710-5_13.

[5] *pandas - Python Data Analysis Library*. [Online]. Disponibile su: https://pandas.pydata.org/

[6] *Ollama - LLM local runner*. [Online]. Disponibile su: https://github.com/ollama/ollama

[7] G. Team *et al.*, «Gemma 2: Improving Open Language Models at a Practical Size», 2024. doi: 10.48550/arXiv.2408.00118.

[8] A. Grattafiori *et al.*, «The Llama 3 Herd of Models», 2024. doi: 10.48550/arXiv.2407.21783.

[9] J. Bai *et al.*, «Qwen Technical Report», 2023. doi: 10.48550/arXiv.2309.16609.

[10] T. B. Brown *et al.*, «Language Models are Few-Shot Learners». 2020. doi: 10.48550/arXiv.2005.14165.

[11] DeepSeek-AI *et al.*, «DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning». 2025. doi: 10.48550/arXiv.2501.12948.

[12] DeepSeek-AI *et al.*, «DeepSeek-V3 Technical Report». 2025. doi: 10.48550/arXiv.2412.19437.

[13] John Gruber, «Markdown». [Online]. Disponibile su: https://daringfireball.net/projects/markdown/

[14] 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.

[15] P. Rajpurkar, R. Jia, e P. Liang, «Know What You Don't Know: Unanswerable Questions for SQuAD», in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, I. Gurevych e Y. Miyao, A c. di, Association for Computational Linguistics, lug. 2018, pp. 784–789. doi: 10.18653/v1/P18-2124.

[16] J. Devlin, M.-W. Chang, K. Lee, e K. Toutanova, «BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding». 2019. doi: 10.48550/arXiv.1810.04805.

[17] A. Vaswani *et al.*, «Attention Is All You Need». 2023. doi: 10.48550/arXiv.1706.03762.

[18] T. Wolf *et al.*, «HuggingFace's Transformers: State-of-the-art Natural Language Processing», 2020. doi: 10.48550/arXiv.1910.03771.

[19] M. Mosbach, M. Andriushchenko, e D. Klakow, «On the Stability of Fine-tuning BERT: Misconceptions, Explanations, and Strong Baselines», 2021. doi: 10.48550/arXiv.2006.04884.

[20] K. Clark, M.-T. Luong, Q. V. Le, e C. D. Manning, «ELECTRA: Pre-training text encoders as discriminators rather than generators». [Online]. Disponibile su: https://openreview.net/pdf?id=r1xMH1BtvB

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

[22] *Distilbert base model*. [Online]. Disponibile su: https://huggingface.co/distilbert/distilbert-base-uncased

[23] G. Hinton, O. Vinyals, e J. Dean, «Distilling the Knowledge in a Neural Network», 2015. doi: 10.48550/arXiv.1503.02531.

[24] 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.

[25] *MobileBERT on Huggingface*. [Online]. Disponibile su: https://huggingface.co/google/mobilebert-uncased

[26] Z. Sun, H. Yu, X. Song, R. Liu, Y. Yang, e D. Zhou, «MobileBERT: a Compact Task-Agnostic BERT for Resource-Limited Devices», 2020. doi: 10.48550/arXiv.2004.02984.

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

[28] I. J. Goodfellow *et al.*, «Generative Adversarial Networks». 2014. doi: 10.48550/arXiv.1406.2661.

[29] R. Caruana, «Multitask Learning», *Machine Learning*, vol. 28, fasc. 1, pp. 41–75, 1997, doi: 10.1023/A:1007379606734.

[30] Sanjaye Elayattu, «Multi-task Fine-tuning with BERT». [Online]. Disponibile su: https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1234/final-reports/final-report-169425270.pdf

[31] Jiacheng Hu e Jack Hung, «Multitasking with BERT». [Online]. Disponibile su: https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1234/final-reports/final-report-169989122.pdf