ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

**DIPARTIMENTO DI SCIENZE**

**CORSO DI LAUREA IN**

**INFORMATICA**

# Apprendimento di lingue straniere attraverso IA


**Relatore**

**Prof. Andrea Asperti**

**Presentata da**

**Giuseppe Forciniti**


---


**Sessione IV**

**Anno Accademico 2023/2024**

# INDICE

**Metodi di apprendimento linguistico** ..... 8

1.1 Flipped Classroom ..... 8

1.2 Input a doppia modalità ..... 9

1.2.1 Effetti sull'apprendimento del vocabolario, sull'ascolto, sulla comprensione del testo e sulla grammatica ..... 11

1.3 Spaced Repetition Software ..... 12

**Natural Language Processing (NLP)** ..... 14

2.1 Definizione e concetti chiavi di NLP ..... 14

2.2 Recenti sviluppi in NLP ..... 15

**Confronto di diversi LLM per NLP** ..... 19

3.1 Benchmark di LLM e chatbot moderni ..... 19

3.2 Analisi e risultati di diversi modelli ..... 20

**ChatGPT per l'apprendimento di lingue straniere** ..... 26

4.1 Pragmatica e autenticità di ChatGPT ..... 26

4.2 Abilità grammaticali di ChatGPT ..... 27

4.3 Panoramica sull'utilità di ChatGPT per l'apprendimento di lingue straniere ..... 30

**ASR nell'apprendimento di lingue straniere** ..... 32

5.1 Confronto tra diversi ASR ..... 32

5.2 Whisper per la valutazione del livello dello studente ..... 35

**L'utilizzo di AI per migliorare i sistemi SR** ..... 37

6.1 Ottimizzazione degli algoritmi SRS ..... 37

6.2 Automatizzazione dei sistemi SR ..... 37

# Introduzione

Quanto sono precisi i vari strumenti di Intelligenza Artificiale (IA)? Possono sostituire completamente gli esseri umani nell'insegnamento delle lingue straniere o, più realisticamente, affiancarli come validi alleati nello studio? Rendono l'apprendimento più efficiente? E se sì, quali sono le migliori opzioni e come vanno utilizzate? Questa tesi proverà a rispondere a queste domande analizzando i dati fino ad ora ottenuti attraverso un'ampia varietà di studi coprendo una vasta gamma di argomenti.

Nel Capitolo 1 vedremo quali metodi di apprendimento di una lingua straniera vengono adottati oggi oltre a quelli tradizionali, come i metodi di Flipped Classroom, Input a doppia modalità, e sistemi di spaced repetition (SRS), dopodiché faremo una panoramica essenziale sulla storia del Natural Language Processing (NLP) e delle sue principali applicazioni nel Capitolo 2, e metteremo a confronto diverse AI per capire qual è la più adatta a determinati contesti nel Capitolo 3. Analizzeremo poi nel Capitolo 4 ChatGPT come strumento per lo studio. Nel Capitolo 5, ci soffermeremo su alcuni sistemi di Automatic Speech Recognition (ASR) per stabilire quello più efficace — in particolare Whisper — e infine nel Capitolo 6 capiremo come l'IA possa migliorare l'utilizzo degli Spaced Repetition Systems (SRS).

L'Intelligenza Artificiale (IA) è una delle rivoluzioni principali dell'ultimo decennio. Abbiamo assistito alla breccia dell'IA nella nostra società in svariati campi, dal Natural Language Processing (NLP), come Chatbots (ChatGPT, Bard, Claude), servizi di traduzione (Google Translate, DeepL), all'elaborazione di immagini e video, come strumenti di ricognizione facciale (eg. Apple FaceID), generazione di immagini (eg. DALL-E, MidJourney) e tanti altri.

Queste svariate applicazioni stanno vedendo un aumento esponenziale nel loro utilizzo, ma devono ancora trovare un assetto definitivo in vari settori, compreso quello

educativo, sia per la difficoltà nel superare metodi tradizionali di lavoro e/o studio, sia perché l'IA risulta ancora spesso inaffidabile, e in bisogno di supervisione e miglioramento. Tuttavia, anche prendendo in considerazione ciò, è innegabile che l'IA si sia rivelata un grande aiuto, e abbia già semplificato molte attività e reso più rapido il lavoro in vari campi. Eppure, in settori umanistici come lo studio delle lingue straniere, il potenziale dell'IA resta spesso sottovalutato.

A questo proposito, vale la pena ricordare che la proficienza in particolare nell'Inglese rimane un buon indicatore della capacità di una nazione nel produrre beni e generare una crescita economica, ed è correlata a buoni investimenti nazionali che permettono alle persone di raggiungere il loro massimo potenziale, offrendo educazione e un buon standard di vita.

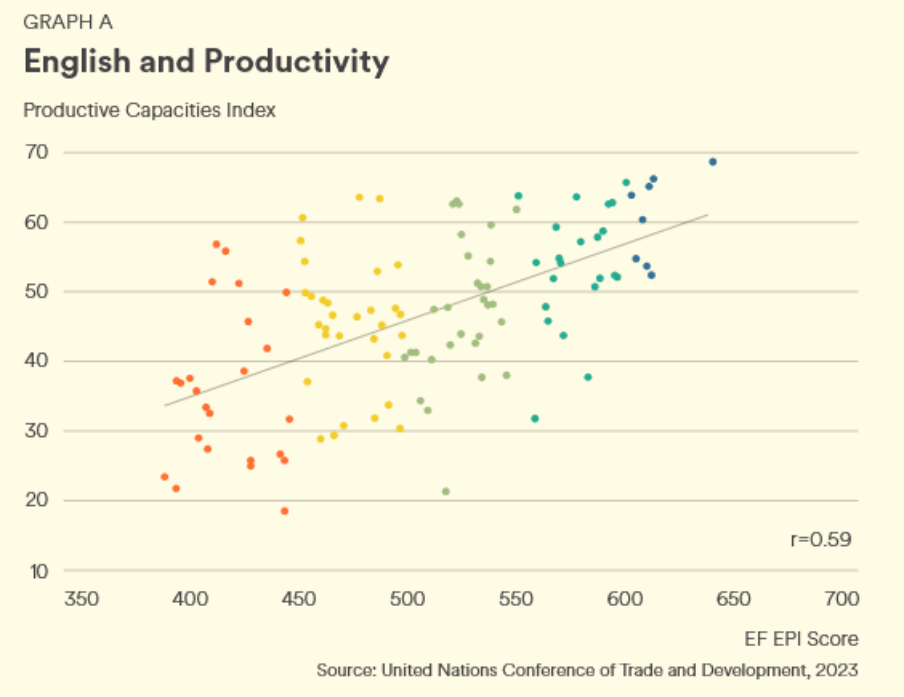


Figura 1: Correlazione positiva tra la conoscenza della lingua inglese (misurata dal punteggio EF EPI) e l'indice della capacità produttiva di una nazione. La correlazione è rappresentata da un coefficiente r=0,59, indicando una relazione moderatamente forte tra le due variabili. [1]

I sistemi tradizionali e scolastici risultano spesso inefficienti per grande parte della popolazione e in un mondo dove essere bilingue è ormai fondamentale, imparare una seconda lingua risulta ancora troppo problematico per molti.

Acquisire competenza in una lingua straniera richiede tempo, studio costante e un coinvolgimento attivo con il materiale di studio (Ortega, 2009 [125]; Seliger, 1977 [126]). L'apprendimento di una lingua è un processo molto diverso quando comparato ad altre materie accademiche, eppure molti studenti, e in particolare studenti universitari, tendono ad adottare le stesse strategie di studio utilizzate in altri corsi (Gardner, 2007 [127]; Oxford & Nyikos, 1989 [128]; Victori & Lockhart, 1995 [129]). Questo porta ad abitudini di studio poco efficaci ed efficienti, che potrebbero spiegare perché molti studenti di lingue si sentano insicuri rispetto alle proprie competenze linguistiche (Graham, 2007) [130]. Questa percezione può a sua volta ridurre la motivazione (Graham, 2007 [130]; Mercer & Ryan, 2010 [131]), la quale rappresenta il fattore più forte e costante nella predizione del successo nell'acquisizione di una lingua straniera (Lasagabaster, Doiz, & Sierra, 2014) [132].

L’AI offre un approccio più semplice ed efficace nello studio di lingue straniere consentendo una personalizzazione elevata per ogni singolo individuo; Tuttavia, rimangono molte domande a cui la tesi mira a rispondere.

# Capitolo 1

## METODI DI APPRENDIMENTO LINGUISTICO

Prima di esplorare l'utilità dell'IA nell'apprendimento delle lingue straniere, è utile capire in cosa consistono i metodi di apprendimento ad oggi utilizzati, come il modello Flipped Classroom, l'Input a doppia modalità e tutte le sue derivazioni e sistemi basati sulla spaced repetition (SRS), per capire come differiscono dall'approccio tradizionale e che risultati offrono.

### 1.1 FLIPPED CLASSROOM

Come evidenziato da Joseph P. Vitta (2023) [5] il modello della *flipped classroom* offre risultati migliori rispetto alla classe tradizionale, migliorando la performance degli studenti. Il metodo ribalta il processo educativo, gli studenti fanno i loro compiti e preparano il materiale prima di andare in classe, e il tempo in classe è invece dedicato alla pratica, discussione e altre attività di alto livello per consolidare l'apprendimento.

L'efficacia del metodo si basa sull'ottimizzazione del tempo in classe (Mehring & Leis, 2018 [64]; Voss & Kostka, 2019 [65]), ed è recentemente emerso in popolarità (van Alten, 2019) [66]. Rispetto alla lezione tradizionale, incoraggia gli studenti a sviluppare livelli cognitivi superiori (Låg & Sæle, 2019 [67];Mehring, 2016 [68], 2018 [69]), e specialmente "i livelli cognitivi superiori della tassonomia, dove avviene l'applicazione di conoscenze e lo sviluppo delle competenze" (Davis, 2016) [70]. Secondo la Tassonomia di Bloom, questi livelli superiori comprendono l'applicazione, l'analisi, la valutazione e la creazione, cioè l'utilizzo di conoscenze per risolvere problemi più complessi.

Anche per questo metodo ci sono criticità, per esempio il fatto che gli studenti principianti hanno difficoltà a comprendere il materiale della *flipped classroom* (Milman, 2012) [71], e non è ideale perché non hanno la chance di chiedere chiarimenti in tempo reale. Inoltre, non è molto pratica in quanto richiede una grande pianificazione da parte dell'insegnante (Mehring, 2016) [68].

| | Effect size (95% CI) | *k* | Domain |
|---|---|---|---|
| Strelan et al. (2020) | $g = 0.50$ (0.42, 0.57) | 198 | Cross-disciplinary |
| Cheng et al. (2019) | $g = 0.19$ (0.11, 0.27) | 55 | Cross-disciplinary |
| Låg and Sæle (2019) | $g = 0.35$ (0.31, 0.40) | 272 | Cross-disciplinary |
| Lo and Hew (2019) | $g = 0.29$ (0.17, 0.41) | 29 | Engineering education |
| Shi et al. (2020) | $g = 0.53$ (0.36, 0.70) | 60 | Cross-disciplinary |
| van Alten et al. (2019) | $g = 0.36$ (0.28, 0.44) | 114 | Cross-disciplinary |
| Xu et al. (2019) | $d = 1.79$ (1.32, 2.27) | 22 | Nursing education in China |

Tabella 1: Esempi di recenti metanalisi sul flipped learning. k: numero di ricerche, g: misura della dimensione dell'effetto, i numeri tra parentesi rappresentano la confidenza. [5]

Questa linea di insegnamento è basata in gran parte sulla tecnologia, anche se non si è specificato esattamente come il tempo fuori dalla classe sia stato usato. Alcune di queste applicazioni hanno adottato un framework Web 1.0 (Lomicka & Lord, 2016) [72]. Mori, Omori e Sato (2016) [73], ad esempio hanno utilizzato PowerPoint e altre tecnologie unidirezionali per il flipping dell'insegnamento dei caratteri di scrittura giapponesi, i kanji.

La tecnologia è stata anche impiegata per facilitare l'interazione tra studenti al di fuori della classe, con applicazioni Web 2.0, come forum e blog (Lin & Hwang, 2018 [74]; Lin, Hwang, Fu, & Chen, 2018 [75]).

## 1.2 INPUT A DOPPIA MODALITÀ

Concorrentemente con l'avanzamento della tecnologia, numerosi ricercatori hanno studiato l'effetto dell'input multimediale per migliorare l'apprendimento di lingue straniere. Nonostante ciò, ci sono stati risultati contrastanti riguardo l'esatta efficacia di diverse modalità di input multimediali. Per esempio, come Ruofei Zhang (2021) [97] fa notare, alcuni ricercatori hanno riportato che il testo puro era più efficace di testo + immagini nel migliorare l'apprendimento del vocabolario (e.g. Acha, 2009 [76]; Boers, Warren, He, & Deconinck, 2017 [78]), mentre altri hanno riportato tutt'altro (e.g. Bisson, van Heuven, Conklin, & Tunney, 2015 [77]; Warren, Boers, Grimshaw, & Siyanova-Chanturia, 2018 [79]). Similmente, alcuni ricercatori hanno affermato che guardare video senza sottotitoli ha migliorato la capacità di ascolto rispetto a guardarli con sottotitoli e.g. Lee & Mayer, 2015 [80]; Yang, 2014 [81]), mentre altri hanno affermato

l'opposto (e.g. Lin,Lee, Wang, & Lin, 2016 [82]; Montero-Perez, Peters, Clarebout, & Desmet, 2014 [83]).

Come evidenziato da Mark Feng Teng (2022) [6], gli input audio-visivi sono estremamente importanti ed efficaci per l'apprendimento e la ritenzione del vocabolario.

Gli studiosi hanno mostrato il potenziale dell'input multimediale nell'apprendimento del vocabolario (ad esempio, Chun e Plass, 1996 [84]; Ramezanali e Faez, 2019 [85]; Teng e Zhang, 2021 [86]; Yoshii, 2006 [87]; Yoshii e Flaitz, 2002 [88]). In linea con la teoria cognitiva dell'apprendimento multimediale (Mayer, 2001) [89], l'integrazione di input verbali e visivi risulta più efficace rispetto all'utilizzo del solo input verbale o visivo.

La teoria è basata su tre ipotesi fondamentali su come le persone elaborano le informazioni:

1. **Ipotesi del doppio canale**: Gli esseri umani possiedono canali separati per elaborare informazioni visive e uditive, come un canale per parole parlate e un canale per le immagini. Questo è in linea con la teoria della doppia codifica (Paivio, 1972 [90], 1986 [91], 1990 [92]).
2. **Ipotesi della capacità limitata**: Le persone hanno una capacità limitata di elaborare e memorizzare informazioni in un dato momento.
3. **Ipotesi dell'elaborazione attiva**: Per imparare, è necessario essere attivamente coinvolti nei processi cognitivi, non basta ricevere passivamente informazioni.

Tuttavia, in alcuni casi, l'input a singola modalità ha superato la doppia modalità nell'apprendimento e nella ritenzione delle parole. Come riportato in una metanalisi (Ramezanali, Uchihara e Faez, 2021) [93], diversi aspetti (es. competenza L2 dei discenti, lingua e tipo di glossario, design della ricerca) possono influenzare i risultati dell'apprendimento del vocabolario in ambienti multimediali.

Questo combacia con la metanalisi di Jhonrey C. Uy e colleghi (2023) [7], che hanno esplorato l'Anime-Inspired English Learning (AIEL), scovandone i vari benefici. La motivazione è un fattore critico nell'apprendimento linguistico (Dörnyei, 2001) [94], e l'AIEL sfrutta la intrinseca attrattiva degli anime per coinvolgere e motivare gli studenti. Diversi studi hanno evidenziato che gli studenti partecipanti a lezioni basate sugli anime

riportavano livelli di motivazione più alti e un maggiore entusiasmo per l'apprendimento (Alsubaie & Alabbad, 2020) [95]. Questo approccio offre un'esperienza multisensoriale che favorisce un'immersione più profonda nella lingua studiata. Questo risultato è in linea con il modello socioeducativo di Noels e colleghi (2020) [96], secondo cui la motivazione ad apprendere una lingua è strettamente legata al valore percepito e all'interesse per la cultura target.

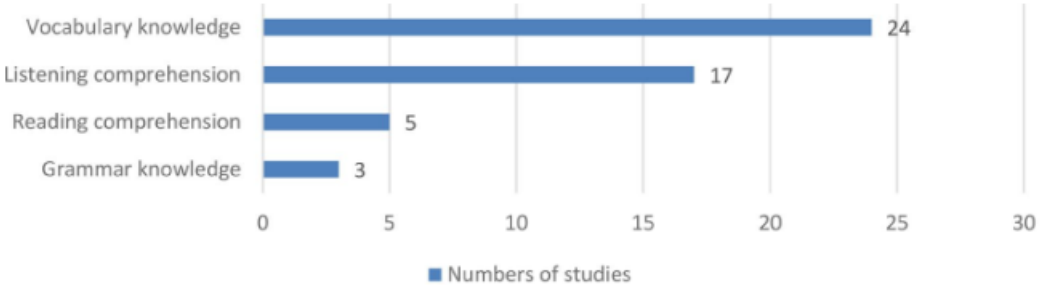


Figura 2: Aspetti dell'apprendimento di lingue straniere indagati da vari studi. [97]

### 1.2.1 Effetti sull'apprendimento del vocabolario, sull'ascolto, sulla comprensione del testo e sulla grammatica

Per migliorare l'apprendimento del vocabolario, i vari studi indicano molto frequentemente l'utilizzo di **audio + animazione + sottotitoli**, di solito correlato alla stimolazione dei canali visivi e uditivi. Questo permette di creare rappresentazioni multiple delle conoscenze e integrarle nella memoria a lungo termine (Aldera & Mohsen, 2013 [98]; Peters, 2019 [99]; Teng, 2019b [100]), inoltre aiuta a mantenere alta l'attenzione degli studenti (Lee & Révész, 2020 [101]; Montero-Perez, Peters, & Desmet, 2018 [103]), facilitando l'elaborazione profonda delle informazioni (Montero-Perez, 2014 [102], 2018 [103]; Winke, 2010 [104]).

Anche la modalità **testo + immagine** risulta utile, e riduce il rischio di sovraccarico cognitivo rispetto alla modalità con audio. Così come **audio + animazione**, che consente agli studenti di costruire rappresentazioni multimediali del vocabolario e di integrarle nella memoria (Peters, 2019) [105]. e animazioni, essendo vivide e coinvolgenti, aumentano l'interesse per il materiale di studio (Lee & Mayer, 2015) [106]. Infine, **audio + sottotitoli** facilita la comprensione e l'estrazione delle informazioni chiave, grazie al supporto visivo fornito dai sottotitoli (Peters, 2019) [105].

## 1.3 Spaced Repetition Software

La Spaced Repetition è una tecnica di apprendimento in cui il materiale che deve essere ricordato è presentato a intervalli che aumentano nel tempo. Questa ripetizione spaziata permette alla memoria dello studente di peggiorare fino ad un certo punto nel quale l'informazione sarà ripresentata per essere revisionata. L'algoritmo che determina la lunghezza di ogni intervallo è basato sul principio che aumentare il tempo tra ripetizioni di oggetti appresi, aumenta sia la durata sia la robustezza della memoria secondo la legge di Jost:

1. La proposizione secondo cui, se due associazioni apprese hanno la stessa forza ma durate diverse, la ripetizione aumenterà la forza di quella più vecchia più di quella più recente.
2. La proposizione secondo cui, se due associazioni apprese hanno la stessa forza ma durate diverse, quella più vecchia decadrà più lentamente di quella più recente. (Colman, 2015, p. 398) [133]

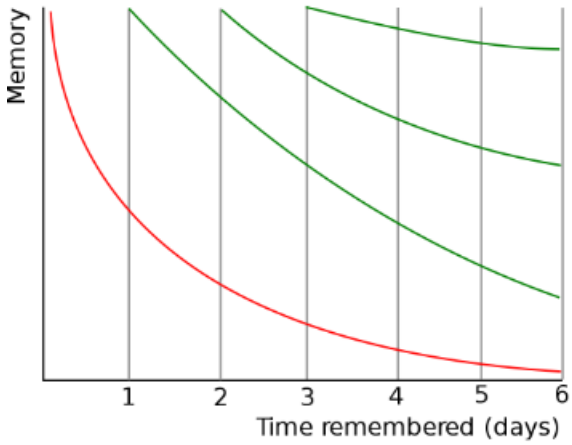


Figura 3: La Forgetting Curve. Dopo solo cinque revisioni, il materiale dovrebbe essere ritenuto per più mesi. [146]

Hermann Ebbinghaus (1885) [134], verificò il principio SR cercando il divario di revisione minimo necessario per memorizzare una parola, studiando migliaia di parole inventate. L'efficacia del metodo è stata testata e verificata diverse volte (Dempster, 1988 [135]; Hintzman, 1974 [136]; Melton, 1970 [137]; Underwood, 1970 [138]; Von Wright, 1971 [139]). Studi di Spitzer (1939) [140] e Cain e Willey (1939) [141] confermarono che si

iniziano a dimenticare rapidamente le informazioni dopo una sola revisione, e questo intervallo aumenta esponenzialmente dopo ogni revisione.

La maggior parte dei sistemi basati sull'apprendimento del vocabolario tendono a usare Spaced Repetition Software (SRS), che involve la creazione di flashcard digitali inserendo manualmente coppie di domanda/risposta. Esempi popolari includono: Anki, Brainscape, Memrise, Quizlet, Supermemo. [145]


Figura 4: Fronte e Retro di una flashcard di Anki. I bottoni sottostanti, "Again" nel caso non si sia ricordata la parola, "Easy", "Good", "Hard" per la difficoltà con cui si è ricordata, i 3 bottoni influenzano l'intervallo di tempo della flashcard.

Secondo lo studio di Nakata (2011) [142], la maggior parte dei sistemi SRS sono progettati in modo da massimizzare l'apprendimento del vocabolario. Hirschel e Fritz (2013) [143], hanno evidenziato come l'apprendimento di vocabolario a lungo termine fosse più elevato tra gli studenti di un'università Giapponese che avevano utilizzato SRS rispetto a quelli che non lo avevano utilizzato. Allo stesso modo, in uno studio più piccolo di Bower e Ruston-Griffiths (2016) [144] ha trovato un miglioramento nella conoscenza del vocabolario attraverso l'uso di SRS, che risultò in punteggi migliori per studenti Giapponesi nel Test Of English for International Communication (TOEIC).

# Capitolo 2

## NATURAL LANGUAGE PROCESSING (NLP)

In questo capitolo vengono fornite informazioni basilari quali la definizione di NLP e i suoi principali ambiti di studio. Successivamente si analizzerà la storia e i più recenti sviluppi nel campo del NLP: dalla nascita dei modelli di linguaggio neurale ai concetti di feed-forward neural networks, lookup tables e multitask learning, fino all'introduzione dei word embeddings, alle architetture delle principali reti neurali come CNN, RNN, LSTM e GRU, e il passaggio cruciale all'utilizzo dei Transformers alla base dei Large Language Models e i moderni sistemi di Automatic Speech Recognition.

## 2.1 DEFINIZIONE E CONCETTI CHIAVI DI NLP

Un linguaggio può essere definito come un set di regole o un set di simboli dove i simboli sono combinati e usati per comunicare informazioni. NLP è un ramo dell'Intelligenza Artificiale e Linguistica, che si concentra nel far capire ai computer frasi o parole scritte in lingue umane. Può essere classificato in due parti:

- Natural Language Linguistics (NLL)
- Natural Language Generation (NLG)

NLL è la scienza del linguaggio che include Fonologia, cioè il suono, Morfologia, cioè la formazione di parole, Sintassi, cioè la struttura delle frasi, Sintassi semantica e Pragmatica, cioè il senso delle frasi. (Lass R, 1998 [13], Umber A, Bajwa I, 2011 [14], Liddy ED, 2001 [15], Feldman S 1999 [16], Walton D 1996 [17])

NLG è il processo di produrre frasi e paragrafi sensati partendo da una rappresentazione delle informazioni interna.

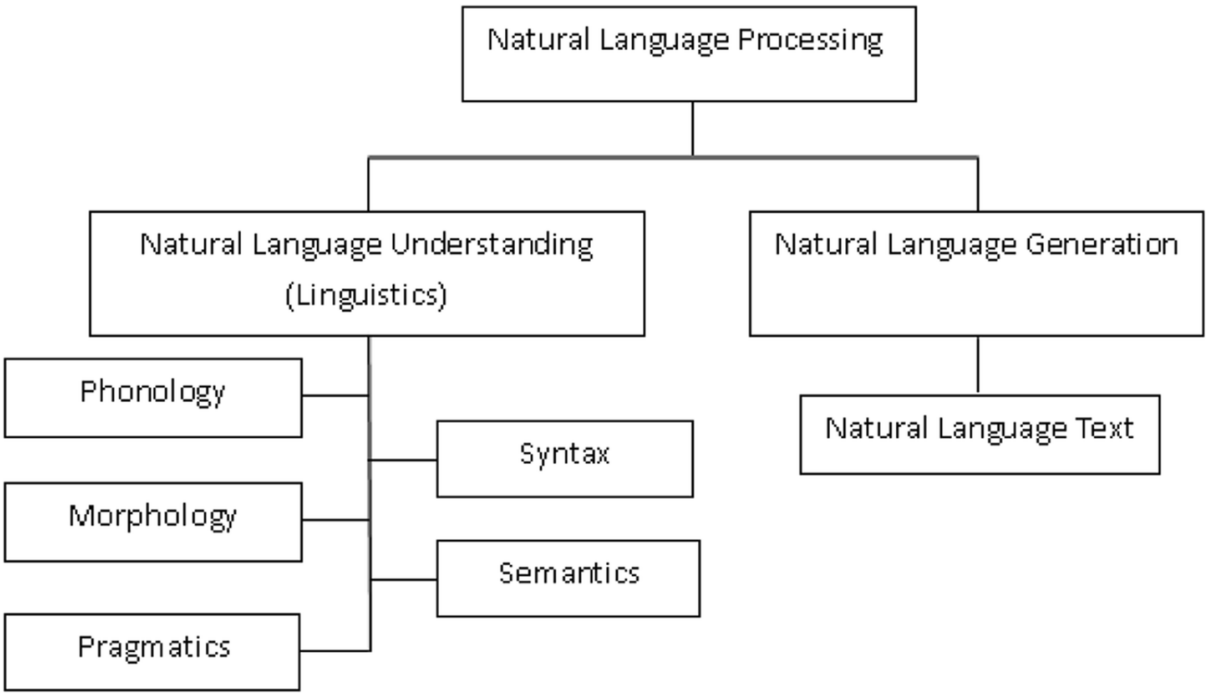


Figura 5: In cosa si distingue il Natural Language Processing. [2]

## 2.2 RECENTI SVILUPPI IN NLP

Nei primi anni 2000, si è sviluppato il *neural language modeling* dove la probabilità dell'occorrenza della parola successiva (token) è determinata data le precedenti n parole. In questo contesto sono stati introdotti concetti come le *feed-forward neural networks,* cioè reti neurali in cui i dati fluiscono in un'unica direzione, da input ad output, senza una memoria intrinseca, *e le lookup tables* che mappano gli elementi in input a rappresentazioni numeriche, che negli NLP generalmente risultano essere vettori. (Bengio Y, 2001) [18].

Successivamente sono state esplorate applicazioni di *multitask learning,* ovvero l'allenamento di un singolo modello per svolgere più compiti contemporaneamente, tra questi, il *part-of-speech tagging,* ossia il riconoscimento della funzione grammaticale delle parole in una frase*,* e il *Named Entity Recognition* (NER), che riconosce entità come nomi di persone e luoghi all'interno di un testo. (Collobert R, 2008) [19]

Un passo cruciale è stato l'introduzione del *word embeddings,* cioè un metodo per rappresentare parole come vettori densi, superando i limiti delle rappresentazioni sparse come i modelli *bag of words***,** che mappavano ogni parola a un dizionario

producendo vettori di alta dimensionalità ed eccessiva lunghezza. I word embeddings hanno quindi aperto la strada per reti neurali più avanzate. (Mikolov T, 2013) [20]

I *convolutional neural networks* (CNN) che erano inizialmente usati per la classificazione e analisi di immagini, hanno trovato applicazione anche nel contesto NLP grazie ai word embeddings (Socher R, 2013 [21], Tan KL, 2022 [22], Santoro A, 2018 [23], Yu S, 2018 [24], Luong MT, 2014 [25], Wiese G, 2017 [26], Newatia R, 2019 [27], Wang W, 2019 [28]).

Parallelamente altri Neural networks come i *Recurrent neural networks* (RNN), hanno guadagnato popolarità per la loro capacità di ricordare informazioni precedenti all'interno di una sequenza (Thomas C, 2019) [29]. Le RNN si sono poi evolute in architetture più sofisticate come le *Long Short-Term Memory* (LSTM), capaci di filtrare le informazioni meno rilevanti (Greff K, 2016 [30], Hochreiter S, 1997 [31]), e le *Gated Recurrent Unit* (GRU), una variante semplificata e ottimizzata delle LSTM (Cho K, 2014 [32], Chung J, 2014 [33]).

Vengono poi introdotti i *Transformers*, che eliminano il bisogno della "sequenza" degli RNN. Hanno un meccanismo di attenzione, che permette al modello di concentrarsi su parti rilevanti del testo (Bahdanau D, 2015 [34], Vaswani A, 2017 [35]). Questo ha permesso il successo di azioni più complesse come la traduzione o il riassunto di documenti.

Modelli basati sui *Transformers*, come *Bidirectional Encoder Representations from Transformers* (BERT), hanno migliorato le prestazioni ulteriormente, analizzando il contesto di una parola sia da destra che da sinistra. (Devlin J, 2018) [36]

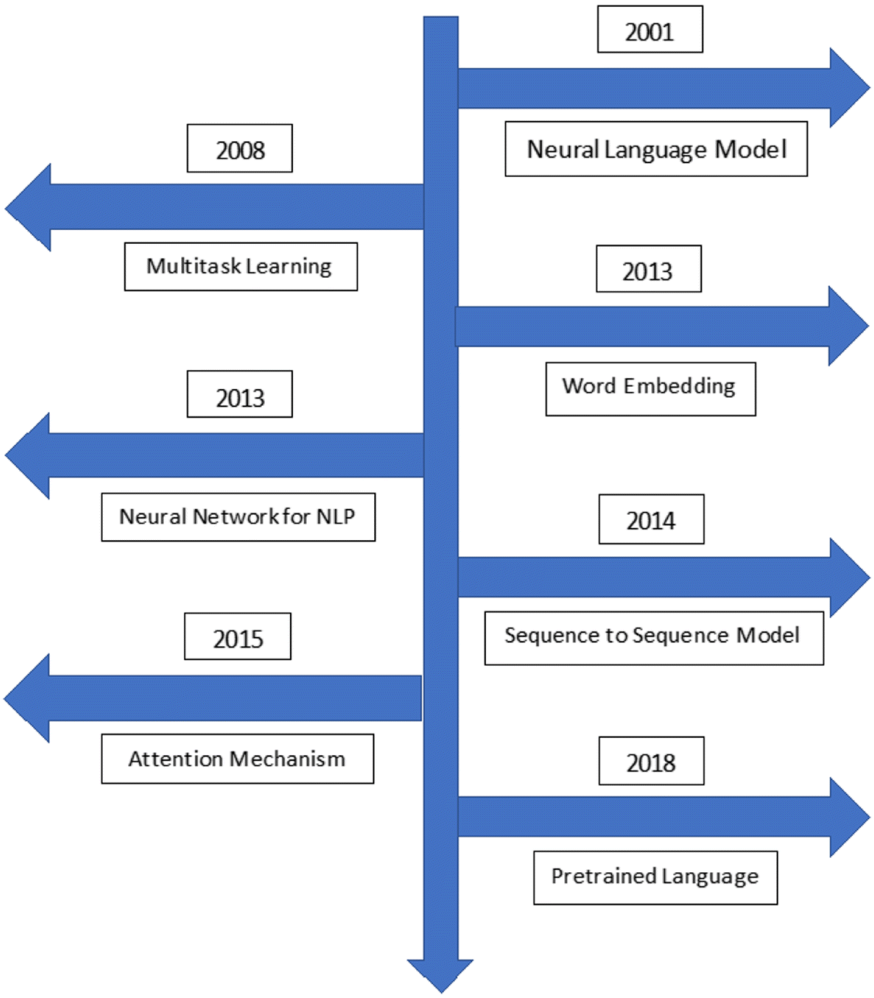


Figura 6: Evoluzione dei modelli NLP. [2]

Questi progressi hanno portato allo sviluppo di *Large Language Models* (LLM) e sistemi di *Automatic Speech Recognition* (ASR). I LLM, basati su Transformers, utilizzano grandi quantità di dati per generare, comprendere e analizzare testi, mentre i sistemi ASR, combinano tecniche NLP con reti neurali per convertire il linguaggio parlato in testo.

Le reti neurali sono oggi applicate in vari compiti di NLP:

- Classificazione del testo (Zeroual, 2017 [37], Tapaswi & Jain, 2012 [38], Ranjan & Basu, 2003 [39])
- Analisi del sentimento (Nasukawa, 2003) [40]
- Riconoscimento delle entità (Ritter, 2011) [41]
- Rilevamento delle emozioni (Sharma, 2016 [42], Seal D, 2020 [43])
- Etichettatura dei ruoli semantici (Palmer, 2005) [44]

Tuttavia, alcuni compiti, come l’analisi di testi informali o il trattamento di lingue meno diffuse, continuano a rappresentare una sfida significativa.

# Capitolo 3

## CONFRONTO DI DIVERSI LLM PER NLP

Per l'obiettivo di questa tesi, è fondamentale analizzare quali LLM siano attualmente più diffusi, e valutare le loro performance nel campo NLP, poiché è un risultato importante per valutare la loro utilità come strumento di supporto nell'insegnamento o apprendimento di lingue straniere. Capire i punti di forza di questi modelli può aiutare a capire come sfruttarli al meglio, migliorando l'esperienza di studio.

### 3.1 BENCHMARK DI LLM E CHATBOT MODERNI

L'introduzione dei *Large Language Models* (LLM) ha portato una rivoluzione nel campo della generazione del testo. Come esempi noti abbiamo GPT-4 di OpenAI, Bard di Google, Claude di Anthropic, e molti altri.

Come Ali Borji (2023) [3] ci fa notare, negli ultimi anni sono stati introdotti numerosi benchmark per gli LLM. Oltre a test specifici come RACE per la comprensione del testo (Lai, 2017) [45] e FEVER per fact-checking (Thorne, 2018) [46], matematica (Frieder, 2023 [47], Azaria, 2022 [48]), programmazione (Chen, 2021) [49], traduzione (Hendy, 2023 [50], Jiao, 2023 [51]), logica (Valmeekam, 2022) [52] e bias (Nadeem, 2020 [53], Liang, 2021 [54], Vig, 2020 [55]), ci sono benchmark compositi come BIG-bench, che coprono più attività (Srivastava, 2022 [56], Qin, 2023 [57]).

Alcuni studi recenti hanno valutato approfonditamente modelli come ChatGPT su svariati dataset, analizzando le risposte in ambiti finanziari, medici, legali e psicologici. Ad esempio, il dataset HC3 (Guo, 2023) [58] confronta risposte generate sia da ChatGPT sia da esperti umani.

Oltre all'accuratezza delle risposte, sono stati valutati altri aspetti come la tossicità e l'etica dei modelli (Welbl, 2021 [59], Zhuo, 2023 [60]), anche se molte di queste analisi sono soggettive e non completamente scientifiche.

Approcci più recenti valutano le risposte generate da modelli con quelle umane, quantitativamente e sistematicamente (Borji, 2023 [61], Bubeck, 2023 [62], Davis, 2023 [63]), permettendo di analizzare la performance più in dettaglio, rivelando quando il sistema sceglie la risposta giusta per motivi sbagliati o viceversa.

## 3.2 Analisi e risultati di diversi modelli

Dai risultati di Bubeck e colleghi (2023) [3], che hanno valutato i modelli basandosi sull'accuratezza delle risposte a domande del dataset Wordsmith, che comprende 1000 domande in diverse categorie. GPT-4 è emerso come il modello primo classificato, seguito da ChatGPT, Claude, e Bard. Si nota che le risposte sono perlopiù non ambigue, questo vuol dire che i modelli tendono a dare risposte completamente corrette o incorrette, ma raramente parzialmente corrette. I risultati corrispondono con i risultati presentati da Khurana e colleghi (2022), che lo hanno confrontato con Gemini, Claude e LlaMa, attraverso l'uso dei benchmark: MMLU, che misura la capacità del modello nel rispondere a domande di cultura generale e specialistiche. GPQA, che misura la capacità del modello nel rispondere a domande difficili. HumanEval, che misura la capacità del modello nella programmazione. GSM-8K, che misura la capacità del modello nel risolvere problemi matematici di livello elementare e medio formulati in linguaggio naturale. MATH, che misura la capacità del modello nel risolvere problemi più avanzati rispetto a GSM-8K.

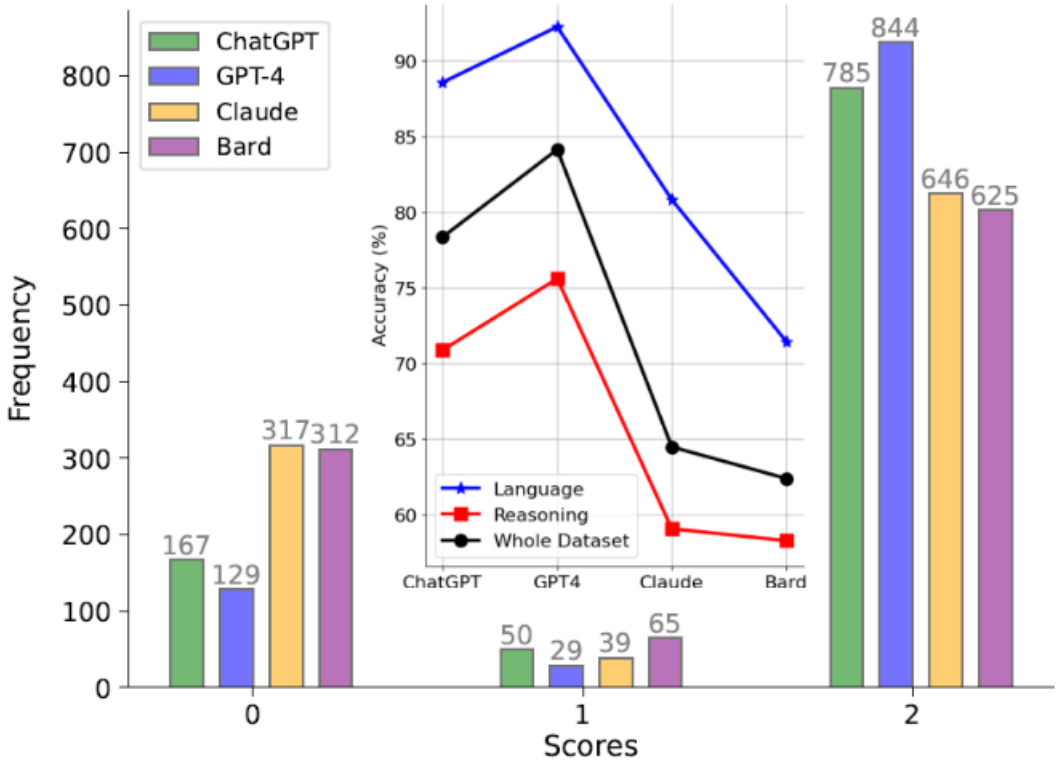


Figura 7: Valutazione dei 4 modelli. Il grafico esterno mostra la correttezza delle risposte e quale frequenza hanno, dove 0: Risposta incorretta, 1: Parzialmente corretta, 2: Totalmente corretta. Il grafico interno rappresenta l'accuratezza delle risposte. [3]

In particolare, con l'eccezione di Bard, i modelli hanno esibito comprensione del linguaggio avanzata e ottime performance nella comprensione del testo, vocabolario, riassunti, grammatica e composizione.

| Category | 0s | | | | 1s | | | | 2s | | | | Acc. (%) | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Ch | Gp | Cl | Ba | Ch | Gp | Cl | Ba | Ch | Gp | Cl | Ba | (Ch, Gp, Cl, Ba) | Acc. |
| Reasoning | 26 | 26 | 48 | 47 | 11 | 5 | 4 | 6 | 90 | **96** | 75 | 74 | (70.87, **75.59**, 59.06, 58.27) | 65.95 |
| Logic | 37 | 27 | 48 | 56 | 3 | 3 | 2 | 4 | 64 | **74** | 54 | 44 | (61.54, **71.15**, 51.92, 42.31) | 56.73 |
| Math and Arithmetic | 26 | 22 | 44 | 51 | 6 | 2 | 4 | 6 | 100 | **108** | 84 | 75 | (75.76, **81.82**, 63.64, 56.82) | 69.51 |
| Facts | 5 | 3 | 4 | 9 | 1 | 2 | 0 | 0 | 29 | 30 | **31** | 26 | (82.86, 85.71, **88.57**, 74.29) | 82.86 |
| Bias and Discrimination | 1 | 0 | 0 | 2 | 1 | 0 | 0 | 5 | 21 | **23** | **23** | 16 | (91.30, **100.0**, **100.0**, 69.57) | **90.22** |
| Wit and Humor | 2 | 2 | 7 | 2 | 4 | 2 | 1 | 3 | 15 | **17** | 13 | 16 | (71.43, **80.95**, 61.90, 76.19) | 72.62 |
| Coding | 6 | 8 | 15 | 19 | 8 | 7 | 5 | 14 | **54** | 53 | 48 | 35 | (**79.41**, 77.94, 70.59, 51.47) | 69.85 |
| Language Understanding | 19 | 16 | 32 | 53 | 9 | 3 | 15 | 17 | 217 | **226** | 198 | 175 | (88.57, **92.24**, 80.82, 71.43) | 83.26 |
| Riddles | 38 | 20 | 102 | 52 | 3 | 2 | 3 | 3 | 130 | **149** | 66 | 116 | (76.02, **87.13**, 38.60, 67.84) | 67.40 |
| Self-Awareness | 0 | 0 | 2 | 3 | 2 | 2 | 2 | 2 | **18** | **18** | 16 | 15 | (**90.00**, **90.00**, 80.00, 75.00) | 83.75 |
| Ethics and Morality | 2 | 2 | 1 | 8 | 1 | 1 | 2 | 5 | **29** | **29** | **29** | 19 | (**90.62**, **90.62**, **90.62**, 59.38) | 82.81 |
| IQ | 5 | 3 | 14 | 10 | 1 | 0 | 1 | 0 | 18 | **21** | 9 | 14 | (75.00, **87.50**, 37.50, 58.33) | 64.58 |

Tabella 2: Valutazioni dei modelli per categoria. Ch: ChatGPT, Gp: GPT-4, Cl: Claude, Ba: Bard [3]

Dal lavoro di Walid Hariri [4], GPT-4 emerge ancora come modello migliore rispetto ai diversi modelli LlaMA, e Gemini.

| Benchmark (shots) | GPT-3.5 | GPT-4 | PaLM | PaLM-2-L | LLaMA 2 |
|---|---|---|---|---|---|
| MMLU (5-shot) | 70.0 | 86.4 | 69.3 | 78.3 | 68.9 |
| TriviaQA (1-shot) | - | - | 81.4 | 86.1 | 85.0 |
| Natural Questions (1-shot) | - | - | 29.3 | 37.5 | 33.0 |
| GSM8K (8-shot) | 57.1 | 92.0 | 56.5 | 80.7 | 56.8 |
| HumanEval (0-shot) | 48.1 | 67.0 | 26.2 | - | 29.9 |
| BIG-Bench Hard (3-shot) | - | - | 52.3 | 65.7 | 51.2 |

Tabella 3: Valutazione di GPT, PaLM e LLaMA attraverso vari benchmarks. [4]

Gemini pone un forte accento sull'adozione di pratiche etiche e riduzione di bias nel campo dell'IA. Tuttavia, come annotato da Khurana e colleghi, risultano pratiche intrinsecamente soggettive e quindi, paradossalmente soggette a bias. Nonostante questa criticità, Gemini dimostra comunque un'elevata performance in numerose casistiche.

| | Meta LLama 3 70B | Gemini Pro 1.5 | Claude 3 Sonnet |
|---|---|---|---|
| MMLU (5-shot) | 82.0 | 81.9 | 79.0 |
| GPQA (0-shot) | 39.5 | 41.5 | 38.5 |
| HumanEval (0-shot) | 81.7 | 71.9 | 73.0 |
| GSM-8K (8-shot, CoT) | 93.0 | 91.7 | 92.3 |
| MATH (4-shot, CoT) | 50.4 | 58.5 (Minerva prompt) | 40.5 |

Tabella 4: Valutazione di LlaMA, Gemini e Claude attraverso vari benchmarks. [4]

Quindi GPT-4 risulta essere il migliore quando confrontato a Gemini, Llama, Claude e Bard. Vediamo invece che risultati ottiene a confronto con i nuovissimi modelli Deepseek R-1 e GPT o4, e o1.

Dai risultati pubblicati da OpenAI (2024), 4o ha risultati migliori rispetto a GPT-4 (e altri modelli).

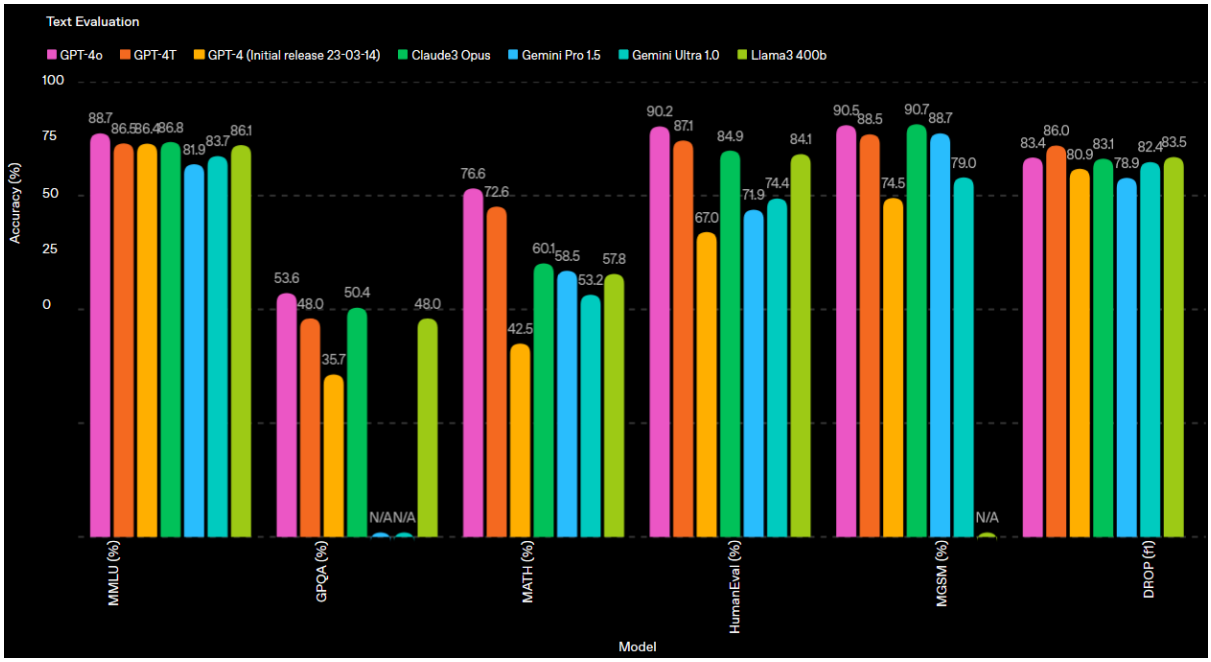


Figura 8: Risultato di valutazione del testo da GPT4, 4o, Claude, Gemini e Llama su vari benchmark. [148]

Sempre dallo studio di OpenAI, o1 risulta nettamente migliore rispetto a 4o, evidenziando come l'adozione della tecnica "chain of thought" – ossia far "pensare" il modello prima di rispondere – consenta a sistemi come o1 di affrontare compiti complessi (in fisica, chimica, biologia, matematica e programmazione) con prestazioni paragonabili a quelle di studenti di dottorato. Infatti, o1, dedicato a compiti di ragionamento, supera nettamente 4o nei benchmark che richiedono una elaborazione multi-step e una riflessione approfondita.

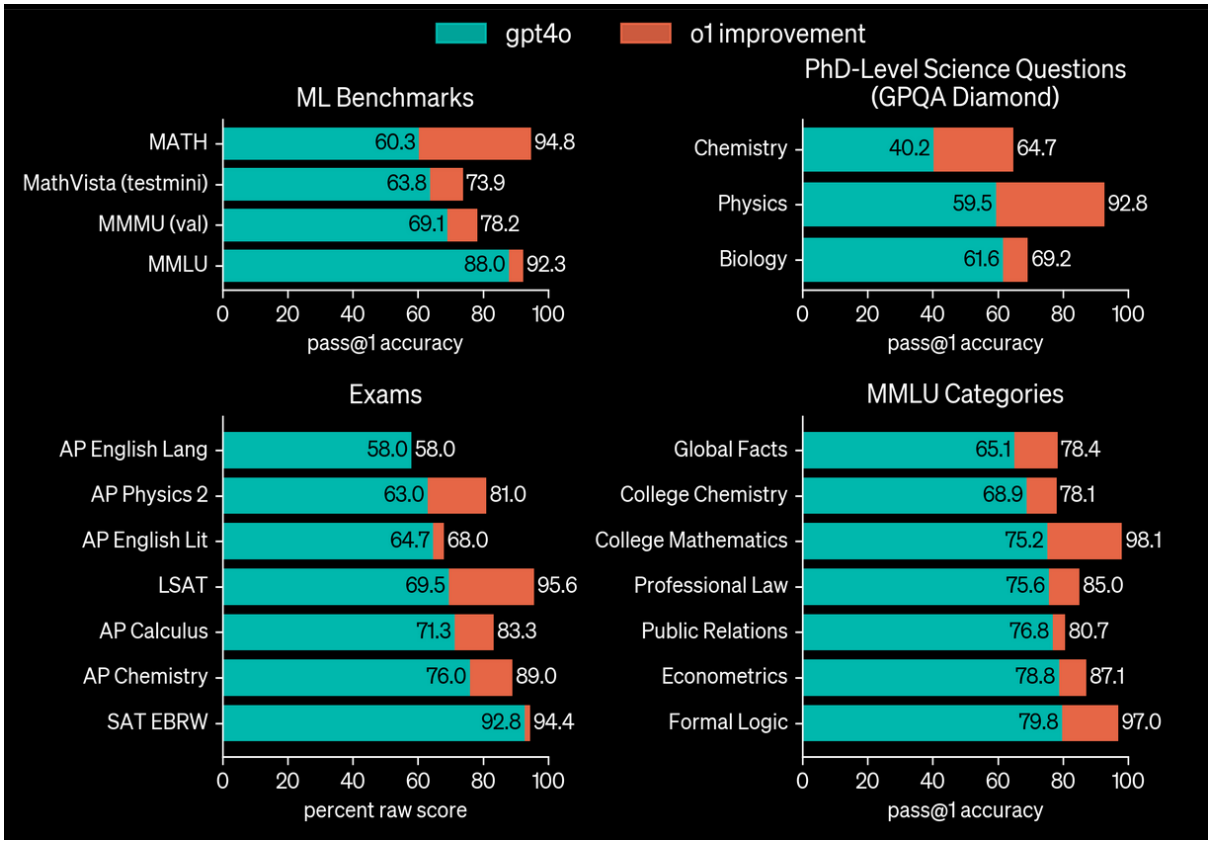


Figura 9: Confronto su diversi benchmark tra o1 e 4o. [149]

Parallelamente, il lavoro condotto da DeepSeek (Guo e colleghi, 2025) conferma l'importanza di questo approccio: DeepSeek-R1, sviluppato sfruttando tecniche di reinforcement learning integrate con il chain-of-thought, dimostra performance superiori rispetto a GPT-o1 su vari benchmark di ragionamento. In particolare, DeepSeek-R1 eccelle in compiti di ragionamento matematico e di codifica, risultando il modello più performante nel campo del NLP per applicazioni di problem solving complesso.

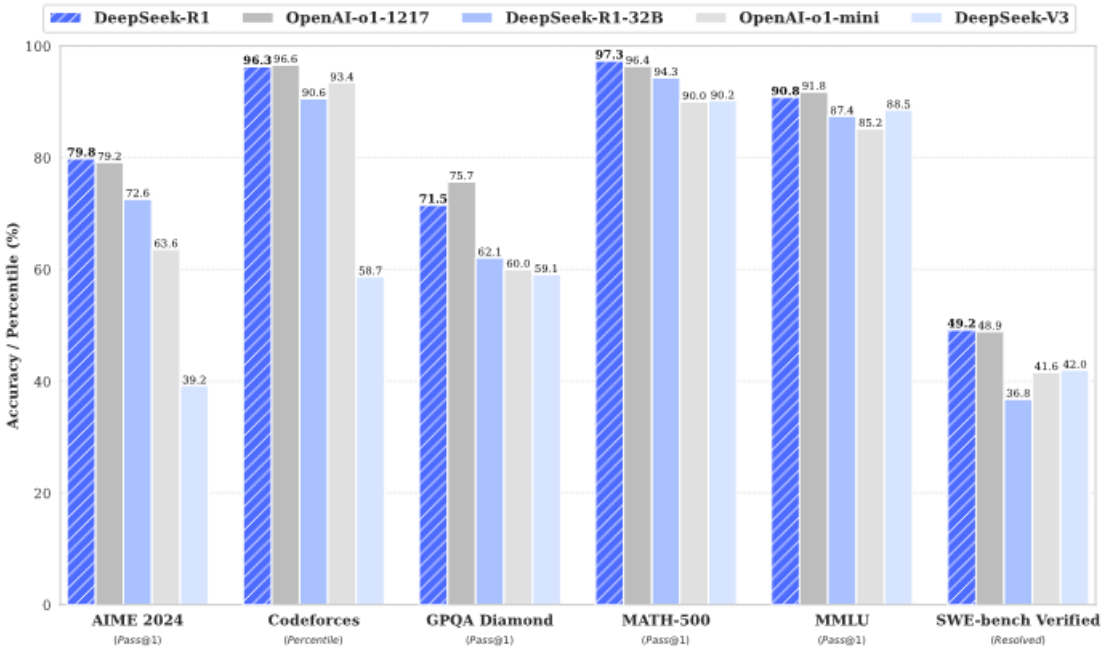


Figura 10: Benchmark performance di Deepseek-R1. [147]

In conclusione, mentre GPT-4o si conferma la scelta principale per applicazioni che richiedono rapidità ed efficienza, i modelli che "pensano" – come o1 e, in modo ancora più notevole, DeepSeek-R1 – offrono un vantaggio decisivo in compiti complessi. Inoltre, GPT-4 è attualmente il più usato e studiato nel campo dell'apprendimento linguistico, quindi per gli scopi della mia tesi, sarà comunque utile andare a studiarlo più in dettaglio, mettendo da parte Deepseek e o1.

# Capitolo 4

## CHATGPT PER L'APPRENDIMENTO DI LINGUE STRANIERE

Come evidenziato dal precedente capitolo, e in particolare dalla flipped classroom, ChatGPT e in generale i vari LLM possono risultare un'ottima scelta per l'apprendimento delle lingue straniere, in quanto gli studenti possono utilizzare i modelli per pratica in autonomia, ad esempio attraverso esercizi di traduzione, dialoghi o spiegazioni grammaticali, per ottimizzare il tempo di studio a casa ed in aula. Come chiarito nel Capitolo 2, ChatGPT, o GPT-4, risulta essere uno dei modelli migliori per NLP ed è quindi adatto per questo scopo. Per questo motivo è utile analizzare quanto sia effettivamente performante andando ad analizzare più nello specifico le sue capacità grammaticali, di pragmatica, e come insegnante basandosi sul metodo del Task Based Language Teaching (TBLT), dove gli studenti ricevono attività interattive da completare, per poi discutere con l'insegnante sul linguaggio usato.

### 4.1 PRAGMATICA E AUTENTICITÀ DI CHATGPT

Come fatto notare da Robert Godwin-Jones (2024) [107], la maggior parte degli studi sulla pragmatica delle IA usano il principio di cooperazione di Grice (Grice, 1989) [108], cioè principi formulati dal filosofo Paul Grice che descrivono come gli interlocutori possono rendere la comunicazione efficace e cooperativa. Altri invece si sono concentrati sulle prestazioni relative agli atti linguistici (Gubelman, 2024 [109]; Tao et al., 2024 [110]) e ai significati impliciti (Ruis et al., 2024) [111]. Gli studi mostrano come le AI abbiano problemi nel campo della pragmatica. C'è una mancanza di "terreno comune" sia linguistico che culturale nel rapporto tra esseri umani ed IA, molto simile a quello presente tra parlanti di lingue madri diverse.
Il concetto di autenticità è complesso e controverso. Gilmore (2019) [112] identifica vari tipi di autenticità, ma non tutti i ricercatori concordano che l'autenticità sia desiderabile, e il modo in cui gli studenti percepiscono testi o compiti autentici è imprevedibile. Un aspetto fondamentale dell'autenticità è la parte sociale e culturale della comunicazione, questo è problematico per l'IA, che manca di un'esperienza sensoriale e socioculturale

reale. Studi hanno infatti dimostrato che ha difficoltà a generare discorsi contestualmente appropriati, rispettando le intenzioni e le emozioni dell'interlocutore.

I chatbot non possono quindi sostituire completamente un insegnante reale nel campo della pragmatica; tuttavia, rimangono utili per ricevere feedback sulla grammatica, scelta di parole e altre sfide nell'apprendimento di lingue straniere. Questo non implica che la presenza di un insegnante o una precisione assoluta nella pragmatica siano indispensabili per imparare una lingua. Infatti, le limitazioni di un approccio tradizionale in classe, come la lentezza, la scarsa personalizzazione e la generale inefficienza, potrebbero rendere gli insegnanti superflui. Un metodo alternativo, basato sull'uso dell'IA combinato con l'immersione nella lingua target e altre risorse, potrebbe dimostrarsi più efficace.

## 4.2 Abilità grammaticali di ChatGPT

Grammatical Error Correction (GEC) è un'attività che si occupa di correggere diversi tipi di errori in un testo, come errori di ortografia, punteggiatura, grammatica e scelta delle parole (Ruder, 2022) [115]. Gli errori grammaticali possono essere classificati in tre categorie principali:

- Errori di omissione
- Errori di sostituzione
- Errori di inserzione

Per valutare la performance dei sistemi GEC sono stati creati diversi dataset di riferimento.

| System | Short | | | Medium | | | Long | | |
|---|---|---|---|---|---|---|---|---|---|
| | Precision | Recall | $F_{0.5}$ | Precision | Recall | $F_{0.5}$ | Precision | Recall | $F_{0.5}$ |
| GECToR | 76.9 | 38.5 | 64.1 | 68.8 | 37.5 | 58.9 | 71.8 | 38.9 | 61.5 |
| Grammarly | 62.5 | 60.6 | 62.1 | 68.9 | 56.0 | 65.9 | 67.3 | 45.3 | 61.4 |
| ChatGPT | 58.5 | 66.7 | 60.0 | 48.7 | 60.7 | 50.7 | 51.0 | 62.8 | 53.0 |

Tabella 5: Performance GC di GECToR, Grammarly e ChatGPT, dove Precision=$\frac{TP}{TP+FP}$ quindi la precisione nel correggere solamente gli errori, Recall = $\frac{TP}{TP+FN}$ quindi quanti errori ha trovato di tutti quelli presenti, F0.5 = $\frac{1.25\times Precision\times Recall}{1.25\times Precision\times Recall}$ quindi una media pesata che da più peso alla Precision. [8]

Grazie allo studio di Haoran Wu e colleghi (2023) [8], è evidente che ChatGPT tende a fare meno errori rispetto a GECToR, ma a correggere più errori possibili, che porta spesso a sovra correzioni, che sebbene porti a ridurre il punteggio di Precision e F0.5, consente anche espressioni linguistiche più flessibili nella correzione grammaticale. GECToR ottiene il valore più alto nella precisione, e Grammarly risulta il più bilanciato tra i 3.

| System | #Under | #Mis | #Over |
|---|---|---|---|
| GECToR | 13 | 4 | 0 |
| Grammarly | 14 | 0 | 1 |
| ChatGPT | 3 | 3 | 30 |

Tabella 6: Numero di sotto correzioni, correzioni errate, e sovra correzioni prodotte dai diversi sistemi GEC. [8]

Anche gli insegnanti beneficiano dal tempo risparmiato nel non dover correggere e dare feedback a errori grammaticali semplici quando un GEC potrebbe essere utilizzato al loro posto. (Toncic, 2020 [113]; Al-Ahdal, 2020 [114])

Secondo Nguyen e colleghi (2022) [116] la quantità di tempo a disposizione ha spinto i docenti della Van Lang University a concentrarsi principalmente sulla correzione degli errori più comuni come metodo principale di correzione. Utilizzare GEC consentirebbe agli insegnanti di dedicare più tempo a una valutazione globale dei testi concentrandosi su aspetti più importanti come l'espressione di idee e significati.

Dallo studio di Ronald Schmidt-Fajlik (2023) [9], che ha confrontato Grammarly, ProWritingAid e ChatGPT nella correzione di errori grammaticali, è risultato che gli errori grammaticali individuati da ChatGPT sono stati descritti in modo molto più dettagliato,

le spiegazioni fornite sugli errori sono più chiare e dirette, rendendo molto più comprensibili le specifiche ragioni per cui tali errori vengono considerati tali. Inoltre, è possibile tradurre le spiegazioni, consentendo così agli studenti di comprenderle più facilmente.

Per esempio:

1. "I experienced a big earthquake" because "I have experienced big earthquake" should be corrected. The reason is "a" should be used before "big" to indicate that the earthquake was a single, specific event.

Col seguente prompt: *Translate the following explanations into Japanese, but only translate the explanations so that a Japanese person can understand their mistakes*: Ha prodotto il seguente risultato con ChatGPT:

1.「I have experienced big earthquake」は、「a」を「big」の前に使用することで、地震が単一の特定のイベントであることを示し、「I experienced a big earthquake」であるべきです.

Gli studenti dell'università Giapponese hanno valutato l'utilità di ChatGPT dopo un semestre, rispondendo a un breve questionario anonimo.

| Question | Yes | No | Other |
|---|---|---|---|
| 1. Was ChatGPT easy to use? | 62 (89.86%) | 7 (10.14%) | 0 (0%) |
| 2. Were the prompts easy to use? | 59 (85.51%) | 8 (11.59%) | 2 (2.9%) |
| 3. Did ChatGPT help you correct your paragraph? | 66 (95.65%) | 3 (4.35%) | 0 (0%) |
| 4. Did ChatGPT find many mistakes? | 65 (94.20%) | 3 (4.35%) | 1 (1.45%) |
| 5. Did you translate the English explanation to Japanese? | 62 (89.86) | 7 (10.14%) | 0 |

Tabella 7: L'uso di ChatGPT come GEC da studenti universitari Giapponesi (N=68). [9]

## 4.3 Panoramica sull'utilità di ChatGPT per l'apprendimento di lingue straniere

Per valutare l'idoneità di ChatGPT come strumento per l'apprendimento di lingue straniere, dallo studio di Sunyoung Kim e colleghi (2023) [10], è stato testato per completare 2 obiettivi:

- Progettare contenuti del corso
- Insegnare agli studenti utilizzando il metodo Task-Based Language Teaching (TBLT)

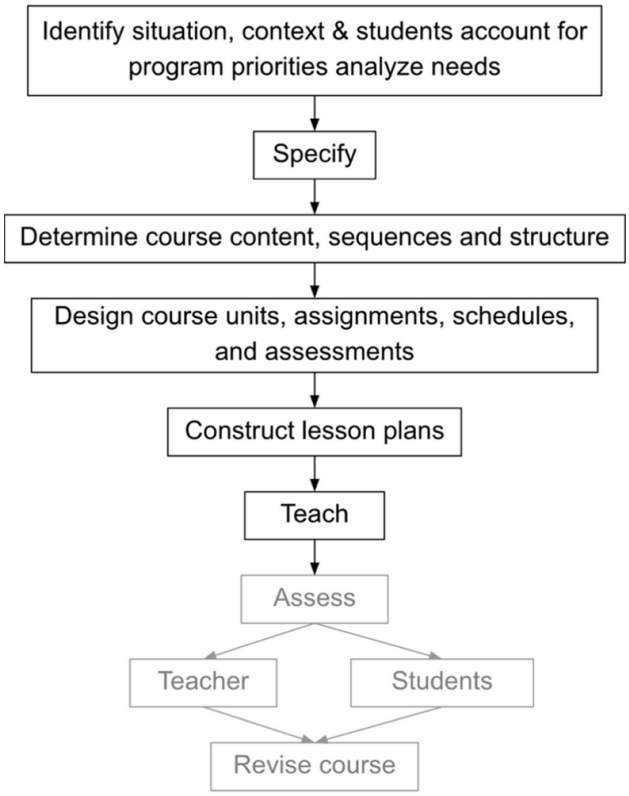


Figura 11: Processo di progettazione di un corso per l'apprendimento di una seconda lingua da H. Douglas Brown (2015). [10]

I risultati indicano che ChatGPT può insegnare una lingua a un certo livello utilizzando il metodo TBLT, ma ha varie limitazioni. Innanzitutto, la mancanza di interazione attiva è uno dei problemi principali. Inoltre, ChatGPT si limita a dare modelli di risposta senza stimolare il pensiero indipendente degli studenti, propone contenuti troppo semplici ed ha una capacità pragmatica limitata. Tuttavia, risulta efficace in quanto permette ampia personalizzazione per diverse lingue, competenze, argomenti e livelli. In secondo luogo, è flessibile e adattabile ad ogni esigenza. Infine, è altamente reattivo, e risponde quasi immediatamente.

# Capitolo 5

## ASR NELL'APPRENDIMENTO DI LINGUE STRANIERE

Come evidenziato in precedenza nel capitolo su input a doppia modalità, i sistemi ASR possono risultare utili nell'apprendimento di lingue straniere, grazie alla creazione di trascrizioni in ogni formato, tra cui sottotitoli, che possono risultare estremamente utili come supporto di apprendimento di una lingua, in quanto possono essere creati nel caso non fossero disponibili e quindi rendere lo studente indipendente sul materiale di studio, per esempio creando sottotitoli per Anime, film e serie televisive, o creando trascrizioni di podcast, video o documentari. Inoltre, i sistemi ASR possono essere usati come strumento per migliorare la pronuncia, o altre casistiche simili.
Per questo è fondamentale esplorare i principali modelli usati al momento, Whisper di OpenAI e Wav2Vec di Meta, e confrontarli.

### 5.1 CONFRONTO TRA DIVERSI ASR

I progressi nell'ASR sono stati stimolati dallo sviluppo di tecniche di pre-addrestamento non supervisionato, come per esempio Wav2Vec 2.0 (Baevski, 2020) [117].

Poiché questi metodi apprendono direttamente dall'audio grezzo senza bisogno di etichette umane, possono utilizzare grandi insiemi di dati parlati e sono stati rapidamente scalati fino a 1.000.000 di ore di dati di addestramento (Zhang, 2021) [118], molte di più delle tipiche ore di un set di dati supervisionato.

La ricerca sul riconoscimento vocale utilizza generalmente la Word Error Rate (WER) come metrica per valutare e confrontare i sistemi. Tuttavia, WER penalizza le differenze tra la trascrizione prodotta dal modello e quella di riferimento, comprese differenze stilistiche o formattazioni. Il problema è evidente nei modelli zero-shot come Whisper. Rimane comunque il metodo migliore in quanto non esistono ancora metodi ampiamente accettati.

Confrontando Whisper e Wav2Vec, risulta che Whisper ottiene una riduzione degli errori del 55.2% di media.

| Dataset | wav2vec 2.0 Large (no LM) | Whisper Large V2 | RER (%) |
|---|---|---|---|
| LibriSpeech Clean | **2.7** | **2.7** | 0.0 |
| Artie | 24.5 | **6.2** | 74.7 |
| Common Voice | 29.9 | **9.0** | 69.9 |
| Fleurs En | 14.6 | **4.4** | 69.9 |
| Tedlium | 10.5 | **4.0** | 61.9 |
| CHiME6 | 65.8 | **25.5** | 61.2 |
| VoxPopuli En | 17.9 | **7.3** | 59.2 |
| CORAAL | 35.6 | **16.2** | 54.5 |
| AMI IHM | 37.0 | **16.9** | 54.3 |
| Switchboard | 28.3 | **13.8** | 51.2 |
| CallHome | 34.8 | **17.6** | 49.4 |
| WSJ | 7.7 | **3.9** | 49.4 |
| AMI SDM1 | 67.6 | **36.4** | 46.2 |
| LibriSpeech Other | 6.2 | **5.2** | 16.1 |
| Average | 29.3 | **12.8** | 55.2 |

Tabella 8: Confronto della robustezza su vari dataset. RER: WER con un normalizzatore di testo applicato, per avere un giudizio più giusto. [11]

Whisper è inoltre capace di lavorare su 75 lingue, quindi è importante valutare quanto sia performante rispetto ad altri modelli.

| Model | MLS | VoxPopuli |
|---|---|---|
| VP-10K + FT | - | 15.3 |
| XLS-R (1B) | 10.9 | 10.6 |
| mSLAM-CTC (2B) | 9.7 | 9.1 |
| Maestro | - | **8.1** |
| Zero-Shot Whisper | **7.3** | 13.6 |

Tabella 9: Multilingual speech recognition. Whisper Zero-Shot migliora la performance sul dataset MLS, ma è particolarmente indietro sul dataset VoxPopuli. [11]

I risultati probabilmente derivano dal fatto che altri modelli hanno incluso VoxPopuli come una fonte principale di dati per il loro pre-addestramento non supervisionato, o al fatto che VoxPopuli dispone di una quantità significativamente maggiore di dati supervisionati. Inoltre, entrambi i dataset sono piuttosto limitati includendo solo 15 lingue.

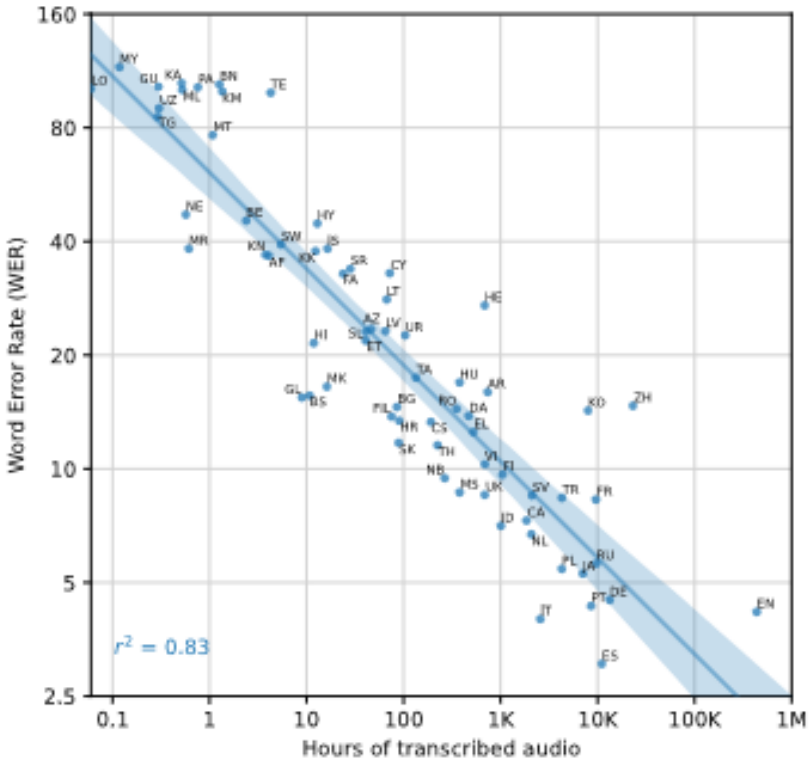


Figura 12: Correlazione tra la quantità di dati di riconoscimento vocale utilizzati nel pre-addestramento e il WER. [11]

La performance in diverse lingue è altamente dipendente dalla quantità di dati su cui il modello è stato addestrato. Come si può notare dalla Figura 12, testando le lingue presenti nel dataset Fleurs, c'è una forte correlazione tra quantità di dati e numero di errori.

| X → English | High | Mid | Low | All |
|---|---|---|---|---|
| XMEF-X | 34.2 | 20.2 | 5.9 | 14.7 |
| XLS-R (2B) | 36.1 | 27.7 | 15.1 | 22.1 |
| mSLAM-CTC (2B) | 37.8 | 29.6 | 18.5 | 24.8 |
| Maestro | **38.2** | 31.3 | 18.4 | 25.2 |
| Zero-Shot Whisper | 36.2 | **32.6** | **25.2** | **29.1** |

Tabella 10: Traduzione X->en, Whisper Zero-Shot ha performance migliori rispetto ad altri modelli su dataset CoVoST2. [11]

Come riportato nella Tabella 10, Whisper risulta il più performante nelle traduzioni da lingua X ad inglese.

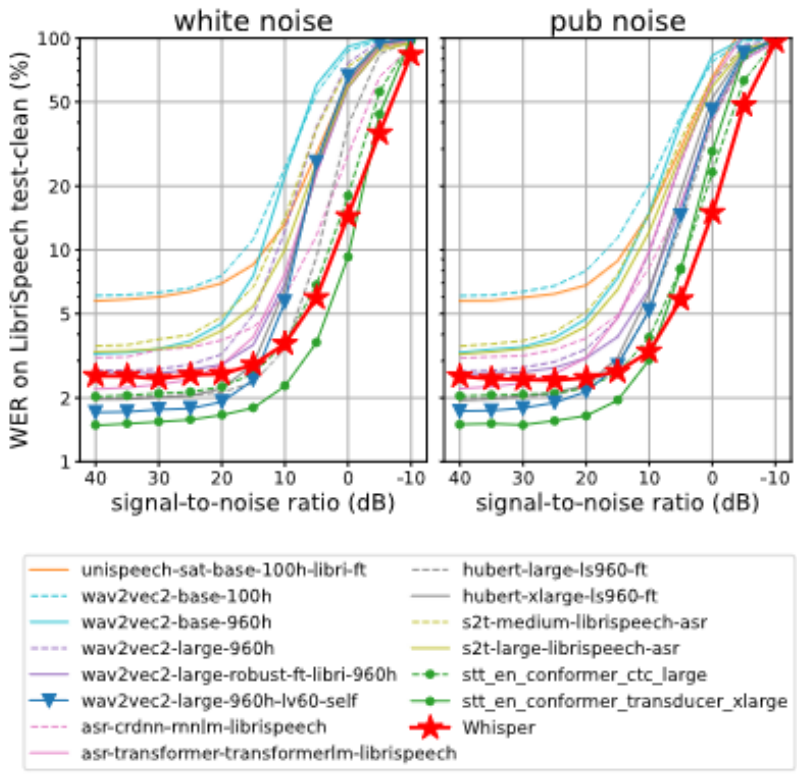


Figura 13: WER sul dataset LibriSpeech, con white noise e pub noise aggiunto. [11]

Whisper risulta il migliore quando viene aggiunto noise all'audio.

In conclusione, dai dati osservati Whisper risulta essere il modello più flessibile e con performance soddisfacenti nei campi della trascrizione e traduzione.

## 5.2 WHISPER PER LA VALUTAZIONE DEL LIVELLO DELLO STUDENTE

L'uso di ASR per l'allenamento della pronuncia risale agli anni 90. Studi mostrano come gli ASR migliorino l'intelligibilità negli studenti di lingue straniere e migliorino alcuni contrasti fonetici mirati (/iː/ vs. /ɪ/, /θ/ vs. /s/) non presenti nelle parole in cui si erano allenati. (Rogers et al.1994). Studi più recenti hanno utilizzato l'ASR di Google per misurare l'intelligibilità del parlato di studenti di lingue straniere con nativi della lingua. Altri studi si sono concentrati sulle discrepanze tra output dell'ASR e target atteso (Chanethom & Henderson, 2022 [123]; Inceoglu et al., 2020 [120]). Studi su Whisper (Radford et al., 2023) [121] hanno confrontato le prestazioni dell'ASR per il parlato indiano nativo e non nativo (Javed et al., 2023) [122]. L'uso dell'ASR per diagnosticare il

parlato di lingue straniere è stato esplorato in modo limitato, ma ricerche precedenti concordano sull'importanza delle sostituzioni fonetiche come principale fonte di errori (Chanethom & Henderson, 2022 [123]; Inceoglu et al., 2020 [120]).

# Capitolo 6

## L'UTILIZZO DI AI PER MIGLIORARE I SISTEMI SR

Come evidenziato in precedenza, i sistemi SRS sono uno strumento molto potente nello studio in generale, e particolarmente utile nello studio delle lingue straniere. Quindi è opportune per lo scopo di questa tesi andare ad analizzare come l'AI è utilizzata per migliorare gli SRS, sia algoritmicamente sia nella creazione delle flashcards stesse.

### 6.1 OTTIMIZZAZIONE DEGLI ALGORITMI SRS

Negli ultimi anni l'integrazione di modelli di intelligenza artificiale nei sistemi SR sta crescendo di popolarità, poiché permette una maggior efficienza nelle sessioni di revisione, rendendo i sistemi più personalizzabili e specifici per ogni studente. Un esempio è lo studio di Tabibian e colleghi (2018), dove sono state usate tecniche di machine learning, in particolare half-life regression, che sarebbe un modello ispirato dalla logistic regression, per stimare il tasso di dimenticanza di ogni elemento, consentendo di programmare le sessioni di revisione in modo dinamico e mirato, massimizzando la memorizzazione a lungo termine ed evitare di utilizzare algoritmi hard-coded. [150]

### 6.2 AUTOMATIZZAZIONE DEI SISTEMI SR

Un altro modo per sfruttare l'AI nell'uso degli SRS è la generazione automatica di domande o flashcards. Per esempio, nello studio di Robinson e Schneider (2023), si è ottimizzato l'SRS generando automaticamente delle domande estratte dai contenuti del corso, creando quiz che vengono poi somministrati agli studenti senza intervento manuale. È stata usato un modello di Logistic Regression per valutare l'efficacia dell'algoritmo, per determinare se l'uso dei quiz generati fosse associato a una maggiore probabilità di superare l'esame finale. I ricercatori hanno suddiviso gli studenti in base al numero di quiz completati, concentrandosi sul confronto tra il quartile superiore (top 25%) e il restante 75%. L'analisi ha rilevato che gli studenti nella top 25% avevano una probabilità aumentata del 152% di superare l'esame finale rispetto agli altri. [151]

Un altro contributo proviene dallo studio di Baillifard e colleghi (2023), dove è stata sperimentata l'implementazione di un tutor AI personalizzato che integrando i principi di sistemi SR, genera automaticamente domande basate sul materiale didattico. La rete neurale, utilizzata per modellare la comprensione dello studente, ha permesso di adattare gli intervalli di ripasso in base alle performance individuali.

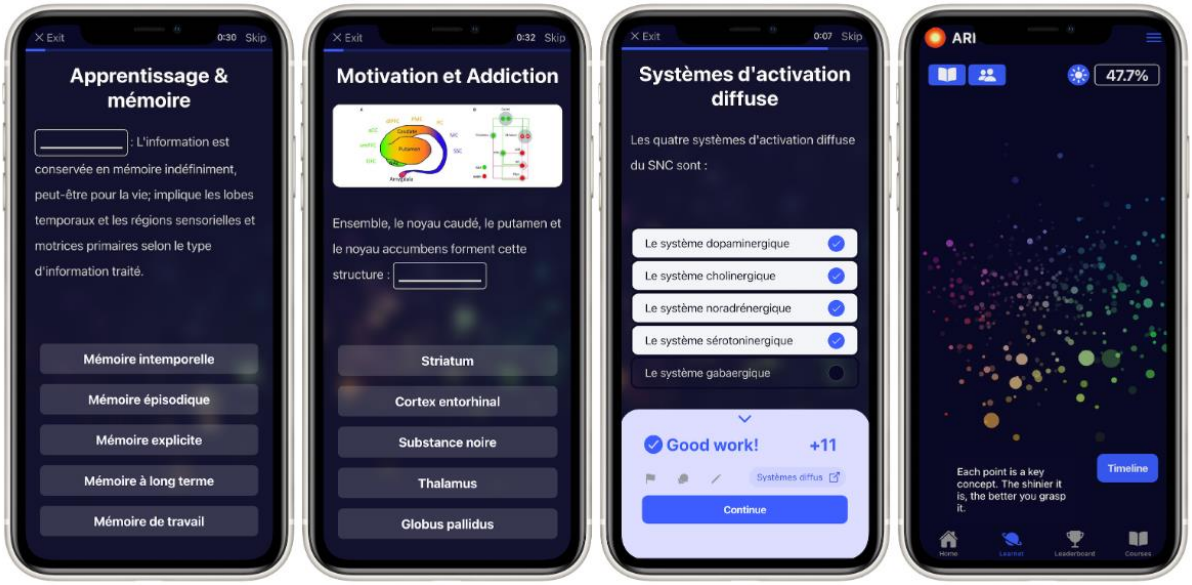


Figura 14: Esempi di domande generate dall'app dell'insegnante AI. Da sinistra verso destra, vediamo una definizione, una domanda basata su un'immagine, una domanda a scelta multipla con feedback, e sulla destra vediamo una "learnet", cioè un'organizzazione visuale dei concetti chiave e della comprensione degli studenti. [152]

Il tutor è stato sviluppato con GPT-3, e stimando in tempo reale il livello di ciascuno studente ha permesso di proporre esercizi con il giusto grado di difficoltà. I risultati hanno evidenziato che gli studenti che ne hanno usufruito hanno ottenuto performance significativamente migliori rispetto ai gruppi di controllo. [152]

# Conclusione

In conclusione, questa tesi ha evidenziato come l'integrazione dell'intelligenza artificiale nell'apprendimento delle lingue straniere rappresenti una svolta significativa nel panorama educativo attuale.

Analizzando i recenti studi nel campo dell'AI e l'adozione di metodologie didattiche innovative – quali la flipped classroom, l'input a doppia modalità e i sistemi di spaced repetition – è risultato che strumenti come i chatbot, per esempio ChatGPT, strumenti ASR, come Whisper, e modelli per ottimizzare sistemi SR, possono supportare e potenziare il percorso di apprendimento in modo personalizzato, interattivo, aumentandone l'efficacia ed efficienza.

L'uso di algoritmi di machine learning per l'ottimizzazione degli SRS e per la generazione automatica di esercizi consente di adattare il materiale didattico alle specifiche esigenze degli studenti, migliorando non solo la memorizzazione, ma anche l'engagement e la motivazione. Allo stesso tempo, l'analisi delle capacità di chatbot come ChatGPT nel correggere errori grammaticali e nel fornire feedback dettagliati ha messo in luce il potenziale di tali tecnologie nel supportare il lavoro degli insegnanti e il lavoro degli studenti sia in classe sia per studio autonomo permettendo una maggiore focalizzazione su aspetti qualitativi dell'apprendimento.

Nonostante l'utilità degli strumenti, questa tesi ha anche evidenziato alcune limitazioni, quali la necessità di affinare la capacità pragmatica degli strumenti AI e di garantire una maggiore autenticità nelle interazioni.

In sintesi, l'integrazione dell'IA nel campo dell'apprendimento delle lingue straniere non solo abilita nuove possibilità in termini di personalizzazione, flessibilità, efficienza ed efficacia, ma sfida anche l'approccio didattico tradizionale, ponendo le basi per un modello educativo più dinamico e centrato sullo studente. Questa tesi ha provato che la sinergia tra innovazione tecnologica e metodologie pedagogiche consolidate potrà contribuire a rendere l'apprendimento linguistico sempre più efficace e coinvolgente.

# Bibliografia


1. EF. (n.d.). *English Proficiency Index*. Retrieved from http://www.ef.com/epi
2. Khurana, D., Koli, A., Khatter, K., & Singh, S. (2022). *Natural language processing: State of the art, current trends and challenges*.
3. Borji, A., & Mohammadian, M. (2023). *Battle of the wordsmiths: Comparing ChatGPT, GPT-4, Claude, and Bard*.
4. Hariri, W. (2023). *Unlocking the potential of ChatGPT: A comprehensive exploration of its applications, advantages, limitations, and future directions in natural language processing*.
5. Vitta, J. P., & Al-Hoorie, A. H. (2023). *The flipped classroom in second language learning: A meta-analysis*.
6. Teng, M. F. (2022). *The effectiveness of multimedia input on vocabulary learning and retention*.
7. Uy, J. C. (2023). *Anime-inspired English learning: A unique approach*.
8. Wu, H., Wang, W., Wan, Y., Jiao, W., & Lyu, M. R. (2023). *ChatGPT or Grammarly? Evaluating ChatGPT on grammatical error correction benchmark*.
9. Schmidt-Fajlik, R. (2023). *ChatGPT as a grammar checker for Japanese English language learners: A comparison with Grammarly and ProWritingAid*.
10. Kim, S., Shim, J., & Shim, J. (2023). *A study on the utilization of OpenAI ChatGPT as a second language learning tool*.
11. Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). *Robust speech recognition via large-scale weak supervision*.
12. Zhang, R., & Zou, D. (2021). *A state-of-the-art review of the modes and effectiveness of multimedia input for second and foreign language learning*.
13. Lass, R. (1998). *Phonology: An introduction to basic concepts* (p. 1). Cambridge University Press.
14. Umber, A., & Bajwa, I. (2011). Minimizing ambiguity in natural language software requirements specification. In *Proceedings of the Sixth International Conference on Digital Information Management* (pp. 102–107).

15. Liddy, E. D. (2001). *Natural language processing*.
16. Feldman, S. (1999). *NLP meets the jabberwocky: Natural language processing in information retrieval*. Online – Weston Then Wilton, 23, 62–73.
17. Walton, D. (1996). A pragmatic synthesis. In *Fallacies arising from ambiguity* (Applied logic series, Vol. 1). Springer.
18. Bengio, Y., Ducharme, R., & Vincent, P. (2001). *A neural probabilistic language model*. In *Proceedings of Neural Information Processing Systems (NIPS)*.
19. Collobert, R., & Weston, J. (2008). *A unified architecture for natural language processing*. In *Proceedings of the 25th International Conference on Machine Learning* (pp. 160–167).
20. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Distributed representations of words and phrases and their compositionality*. In *Advances in Neural Information Processing Systems*.
21. Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A. Y., & Potts, C. (2013). *Recursive deep models for semantic compositionality over a sentiment treebank*. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing* (pp. 1631–1642).
22. Tan, K. L., Lee, C. P., Anbananthen, K. S. M., & Lim, K. M. (2022). *RoBERTa-LSTM: A hybrid model for sentiment analysis with transformers and recurrent neural network*. IEEE Access.
23. Santoro, A., Faulkner, R., Raposo, D., Rae, J., Chrzanowski, M., Weber, T., ... & Lillicrap, T. (2018). *Relational recurrent neural networks*. Advances in Neural Information Processing Systems, 31.
24. Yu, S., et al. (2018). *A multi-stage memory augmented neural network for machine reading comprehension*. In *Proceedings of the Workshop on Machine Reading for Question Answering*.
25. Luong, M. T., Sutskever, I., Le, Q. V., Vinyals, O., & Zaremba, W. (2014). *Addressing the rare word problem in neural machine translation*. arXiv preprint arXiv:1410.8206.
26. Wiese, G., Weissenborn, D., & Neves, M. (2017). *Neural domain adaptation for biomedical question answering*. arXiv preprint arXiv:1706.03610.

27. Newatia, R. (2019). *Sentence classification using convolutional neural networks*. Medium. Retrieved December 15, 2021, from https://medium.com/saarthi-ai/sentence-classification-using-convolutional-neural-networks-ddad72c7048c
28. Wang, W., & Gang, J. (2018). *Application of convolutional neural network in natural language processing*. In *Proceedings of the 2018 International Conference on Information Systems and Computer Aided Education (ICISCAE)* (pp. 64–70).
29. Thomas, C. (2019). *Recurrent neural networks and natural language processing*. Retrieved from https://towardsdatascience.com/recurrent-neural-networks-and-natural-language-processing-73af640c2aa1
30. Greff, K., Srivastava, R. K., Koutník, J., Steunebrink, B. R., & Schmidhuber, J. (2016). *LSTM: A search space odyssey*. IEEE Transactions on Neural Networks and Learning Systems, 28(10), 2222–2232.
31. Hochreiter, S., & Schmidhuber, J. (1997). *Long short-term memory*. Neural Computation, 9(8), 1735–1780.
32. Cho, K., van Merriënboer, B., Bahdanau, D., & Bengio, Y. (2014). *On the properties of neural machine translation: Encoder-decoder approaches*. arXiv preprint arXiv:1409.1259.
33. Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). *Empirical evaluation of gated recurrent neural networks on sequence modeling*. arXiv preprint arXiv:1412.3555.
34. Bahdanau, D., Cho, K., & Bengio, Y. (2015). *Neural machine translation by jointly learning to align and translate*. In *Proceedings of the International Conference on Learning Representations (ICLR)*.
35. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). *Attention is all you need*. In *Advances in Neural Information Processing Systems* (pp. 5998–6008).
36. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of deep bidirectional transformers for language understanding*. arXiv preprint arXiv:1810.04805.
37. Zeroual, I., Lakhouaja, A., & Belahbib, R. (2017). *Towards a standard part of speech tagset for the Arabic language*. Journal of King Saud University-Computer and Information Sciences, 29(2), 171–178.

38. Tapaswi, N., & Jain, S. (2012). *Treebank based deep grammar acquisition and part-of-speech tagging for Sanskrit sentences*. In *Proceedings of the 2012 CSI Sixth International Conference on Software Engineering (CONSEG)* (pp. 1–4). IEEE.
39. Ranjan, P., & Basu, H. V. S. S. A. (2003). *Part of speech tagging and local word grouping techniques for natural language parsing in Hindi*. In *Proceedings of the 1st International Conference on Natural Language Processing (ICON 2003)*.
40. Yi, J., Nasukawa, T., Bunescu, R., & Niblack, W. (2003). *Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques*. In *Proceedings of the Third IEEE International Conference on Data Mining (ICDM 2003)* (pp. 427–434). IEEE.
41. Ritter, A., Clark, S., & Etzioni, O. (2011). *Named entity recognition in tweets: An experimental study*. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 1524–1534). Association for Computational Linguistics.
42. Sharma, S., Srinivas, P. Y. K. L., & Balabantaray, R. C. (2016). *Emotion detection using online machine learning method and TLBO on mixed script*. In *Proceedings of the Language Resources and Evaluation Conference 2016* (pp. 47–51).
43. Seal, D., Roy, U. K., & Basak, R. (2020). *Sentence-level emotion detection from text based on semantic rules*. In M. Tuba, S. Akashe, & A. Joshi (Eds.), *Information and Communication Technology for Sustainable Development* (Advances in Intelligent Systems and Computing, Vol. 933). Springer. https://doi.org/10.1007/978-981-13-7166-0_42
44. Palmer, M., Gildea, D., & Kingsbury, P. (2005). *The proposition bank: An annotated corpus of semantic roles*. Computational Linguistics, 31(1), 71–106.
45. Lai, G., Xie, Q., Liu, H., Yang, Y., & Hovy, E. (2017). *Race: Large-scale reading comprehension dataset from examinations*. arXiv preprint arXiv:1704.04683.
46. Thorne, J., Vlachos, A., Christodoulopoulos, C., & Mittal, A. (2018). *Fever: A large-scale dataset for fact extraction and verification*. arXiv preprint arXiv:1803.05355.
47. Frieder, S., Pinchetti, L., Griffiths, R.-R., Salvatori, T., Lukasiewicz, T., Petersen, P. C., Chevalier, A., & Berner, J. (2023). *Mathematical capabilities of ChatGPT*. Retrieved from https://arxiv.org/abs/2301.13867

48. Azaria, A. (2022). *ChatGPT usage and limitations*. arXiv.
49. Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P. d. O., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., et al. (2021). *Evaluating large language models trained on code*. arXiv preprint arXiv:2107.03374.
50. Hendy, A., Abdelrehim, M., Sharaf, A., Raunak, V., Gabr, M., Matsushita, H., Kim, Y. J., Afify, M., & Awadalla, H. H. (2023). *How good are GPT models at machine translation? A comprehensive evaluation*. arXiv preprint arXiv:2302.09210.
51. Jiao, W., Wang, W., Huang, J.-t., Wang, X., & Tu, Z. (2023). *Is ChatGPT a good translator? A preliminary study*. arXiv preprint arXiv:2301.08745.
52. Valmeekam, K., Olmo, A., Sreedharan, S., & Kambhampati, S. (2022). *Large language models still can't plan (a benchmark for LLMs on planning and reasoning about change)*. arXiv preprint arXiv:2206.10498.
53. Nadeem, M., Bethke, A., & Reddy, S. (2020). *Stereoset: Measuring stereotypical bias in pretrained language models*. arXiv preprint arXiv:2004.09456.
54. Liang, P. P., Wu, C., Morency, L.-P., & Salakhutdinov, R. (2021). *Towards understanding and mitigating social biases in language models*. In *International Conference on Machine Learning* (pp. 6565–6576). PMLR.
55. Vig, J., Gehrmann, S., Belinkov, Y., Qian, S., Nevo, D., Singer, Y., & Shieber, S. (2020). *Investigating gender bias in language models using causal mediation analysis*. Advances in Neural Information Processing Systems, 33, 12388–12401.
56. Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., Brown, A. R., Santoro, A., Gupta, A., Garriga-Alonso, A., et al. (2022). *Beyond the imitation game: Quantifying and extrapolating the capabilities of language models*. arXiv preprint arXiv:2206.04615.
57. Qin, C., Zhang, A., Zhang, Z., Chen, J., Yasunaga, M., & Yang, D. (2023). *Is ChatGPT a general-purpose natural language processing task solver?* arXiv preprint arXiv:2302.06476.
58. Guo, B., Zhang, X., Wang, Z., Jiang, M., Nie, J., Ding, Y., Yue, J., & Wu, Y. (2023). *How close is ChatGPT to human experts? Comparison corpus, evaluation, and detection*. arXiv preprint arXiv:2301.07597.

59. Welbl, J., Glaese, A., Uesato, J., Dathathri, S., Mellor, J., Hendricks, L. A., Anderson, K., Kohli, P., Coppin, B., & Huang, P.-S. (2021). *Challenges in detoxifying language models*. arXiv preprint arXiv:2109.07445.
60. Zhuo, T. Y., Huang, Y., Chen, C., & Xing, Z. (2023). *Exploring AI ethics of ChatGPT: A diagnostic analysis*. arXiv preprint arXiv:2301.12867.
61. Borji, A. (2023). *A categorical archive of ChatGPT failures*. arXiv preprint arXiv:2302.03494.
62. Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., et al. (2023). *Sparks of artificial general intelligence: Early experiments with GPT-4*. arXiv preprint arXiv:2303.12712.
63. Davis, E. (2023). *Benchmarks for automated commonsense reasoning: A survey*. Retrieved from https://arxiv.org/abs/2302.04752
64. Mehring, J., & Leis, A. (Eds.). (2018). *Innovations in flipping the language classroom: Theories and practices*. Springer.
65. Voss, E., & Kostka, I. (2019). *Flipping academic English language learning: Experiences from an American university*. Springer Nature Singapore.
66. van Alten, D. C. D., Phielix, C., Janssen, J., & Kester, L. (2019). *Effects of flipping the classroom on learning outcomes and satisfaction: A meta-analysis*. Educational Research Review, 28, 100281.
67. Låg, T., & Sæle, R. G. (2019). *Does the flipped classroom improve student learning and satisfaction? A systematic review and meta-analysis*. AERA Open, 5, 3.
68. Mehring, J. (2016). *Present research on the flipped classroom and potential tools for the EFL classroom*. Computers in the Schools, 33, 1–10.
69. Mehring, J. (2018). *The flipped classroom*. In J. Mehring & A. Leis (Eds.), *Innovations in flipping the language classroom: Theories and practices* (pp. 1–10). Springer.
70. Davis, N. L. (2016). *Anatomy of a flipped classroom*. Journal of Teaching in Travel & Tourism, 16, 228–232.
71. Milman, N. B. (2012). *The flipped classroom strategy: What is it and how can it best be used?* Distance Learning, 9, 85–87.

72. Lomicka, L., & Lord, G. (2016). *Social networking in language learning*. In F. Farr & L. Murray (Eds.), *The Routledge handbook of language learning and technology* (pp. 225–268). Routledge.
73. Mori, Y., Omori, M., & Sato, K. (2016). *The impact of flipped online Kanji instruction on written vocabulary learning for introductory and intermediate Japanese language students*. Foreign Language Annals, 49, 729–749.
74. Lin, C.-J., & Hwang, G.-J. (2018). *A learning analytics approach to investigating factors affecting EFL students' oral performance in a flipped classroom*. Journal of Educational Technology & Society, 21, 205–219.
75. Lin, C.-J., Hwang, G.-J., Fu, Q.-K., & Chen, J.-F. (2018). *A flipped contextual game-based learning approach to enhancing EFL students' English business writing performance and reflective behaviors*. Journal of Educational Technology & Society, 21, 117–131.
76. Acha, J. (2009). *The effectiveness of multimedia programmes in children's vocabulary learning*. British Journal of Educational Technology, 40(1), 23–31. https://doi.org/10.1111/j.1467-8535.2007.00800.x
77. Bisson, M. J., van Heuven, W. J., Conklin, K., & Tunney, R. J. (2015). *The role of verbal and pictorial information in multimodal incidental acquisition of foreign language vocabulary*. Quarterly Journal of Experimental Psychology, 68(7), 1306–1326. https://doi.org/10.1080/17470218.2014.979211
78. Boers, F., Warren, P., He, L., & Deconinck, J. (2017). *Does adding pictures to glosses enhance vocabulary uptake from reading?* System, 66, 113–129. https://doi.org/10.1016/j.system.2017.03.017
79. Warren, P., Boers, F., Grimshaw, G., & Siyanova-Chanturia, A. (2018). *The effect of gloss type on learners' intake of new words during reading: Evidence from eye-tracking*. Studies in Second Language Acquisition, 40(4), 883–906. https://doi.org/10.1017/S0272263118000177
80. Lee, H., & Mayer, R. E. (2015). *Visual aids to learning in a second language: Adding redundant video to an audio lecture*. Applied Cognitive Psychology, 29(3), 445–454. https://doi.org/10.1002/acp.3123

81. Yang, H. Y. (2014). *Does multimedia support individual differences? EFL learners' listening comprehension and cognitive load*. Australasian Journal of Educational Technology, 30(6), 699–713. https://doi.org/10.14742/ajet.639
82. Lin, J. J., Lee, Y. H., Wang, D. Y., & Lin, S. S. (2016). *Reading subtitles and taking e-notes while learning scientific materials in a multimedia environment: Cognitive load perspectives on EFL students*. Journal of Educational Technology & Society, 19(4), 47–58.
83. Montero-Perez, M., Peters, E., Clarebout, G., & Desmet, P. (2014). *Effects of captioning on video comprehension and incidental vocabulary learning*. Language Learning & Technology, 18(1), 118–141.
84. Chun, D. M., & Plass, J. L. (1996). *Effects of multimedia annotations on vocabulary acquisition*. The Modern Language Journal, 80(2), 183–198.
85. Ramezanali, N., & Faez, F. (2019). *Vocabulary learning and retention through multimedia glossing*. Language Learning & Technology, 23(2), 105–124.
86. Teng, F., & Zhang, D. (2021). *The associations between working memory and the effects of multimedia input on L2 vocabulary learning*. International Review of Applied Linguistics in Language Teaching. https://doi.org/10.1515/iral-2021-0130
87. Yoshii, M. (2006). *L1 and L2 glosses: Their effects on incidental vocabulary learning*. Language Learning & Technology, 10(3), 85–101.
88. Yoshii, M., & Flaitz, J. (2002). *Second language incidental vocabulary retention: The effect of text and picture annotation types*. CALICO Journal, 20(1), 33–58.
89. Mayer, R. E. (2001). *Multimedia learning*. Cambridge University Press.
90. Paivio, A. (1972). *Imagery and verbal processes*. Holt, Rinehart & Winston.
91. Paivio, A. (1986). *Mental representations*. Oxford University Press.
92. Paivio, A. (1990). *Mental representation: A dual-coding approach*. Oxford University Press.
93. Ramezanali, N., Uchihara, T., & Faez, F. (2021). *Efficacy of multimodal glossing on second language vocabulary learning: A meta-analysis*. TESOL Quarterly, 55, 105–133. https://doi.org/10.1002/tesq.579
94. Dornyei, Z. (2001). *Motivational strategies in the language classroom*. Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511667343

95. Alsubaie, S. S., & Alabbad, A. M. (2020). *The effect of Japanese animation series on informal third language acquisition among Arabic native speakers*. English Language Teaching, 13(8), 91–119.
96. Noels, K. A., Lou, N. M., Vargas Lascano, D. I., Chaffee, K. E., Dincer, A., Zhang, Y. S. D., & Zhang, X. (2020). *Self-determination and motivated engagement in language learning*. In *The Palgrave handbook of motivation for language learning* (pp. 95–115). Springer International Publishing.
97. Zhang, R., & Zou, D. (2021). *A state-of-the-art review of the modes and effectiveness of multimedia input for second and foreign language learning*. Computer Assisted Language Learning, 35, 1–27. https://doi.org/10.1080/09588221.2021.1896555
98. Aldera, A. S., & Mohsen, M. A. (2013). *Annotations in captioned animation: Effects on vocabulary learning and listening skills*. Computers & Education, 68, 60–75. https://doi.org/10.1016/j.compedu.2013.04.018
99. Peters, E. (2019). *The effect of imagery and on-screen text on foreign language vocabulary learning from audiovisual input*. TESOL Quarterly, 53(4), 1008–1032. https://doi.org/10.1002/tesq.531
100. Teng, F. (2019). *Incidental vocabulary learning for primary school students: The effects of L2 caption type and word exposure frequency*. The Australian Educational Researcher, 46(1), 113–136. https://doi.org/10.1007/s13384-018-0279-6
101. Lee, M., & Révész, A. (2020). *Promoting grammatical development through captions and textual enhancement in multimodal input-based tasks*. Studies in Second Language Acquisition. Retrieved from https://discovery.ucl.ac.uk/id/eprint/10097884/3/Revesz_Main%20document_Final%20Submission.pdf
102. Montero-Perez, M., Peters, E., Clarebout, G., & Desmet, P. (2014). *Effects of captioning on video comprehension and incidental vocabulary learning*. Language Learning & Technology, 18(1), 118–141.
103. Montero-Perez, M., Peters, E., & Desmet, P. (2018). *Vocabulary learning through viewing video: The effect of two enhancement techniques*. Computer Assisted Language Learning, 31(1–2), 1–26.

104. Winke, P., Gass, S., & Syodorenko, T. (2010). *The effects of captioning videos used for foreign language listening activities*. Language Learning & Technology, 14(1), 65–86.

105. Peters, E. (2019). *The effect of imagery and on-screen text on foreign language vocabulary learning from audiovisual input*. TESOL Quarterly, 53(4), 1008–1032. https://doi.org/10.1002/tesq.531

106. Lee, H., & Mayer, R. E. (2015). *Visual aids to learning in a second language: Adding redundant video to an audio lecture*. Applied Cognitive Psychology, 29(3), 445–454. https://doi.org/10.1002/acp.3123

107. Godwin-Jones, R. (2024). *Generative AI, pragmatics, and authenticity in second language learning*. arXiv preprint arXiv:2410.14395. https://doi.org/10.48550/arXiv.2410.14395

108. Grice, P. (1989). *Studies in the way of words*. Harvard University Press.

109. Gubelmann, R. (2024). *Large language models, agency, and why speech acts are beyond them (for now) – A Kantian-cum-pragmatist case*. Philosophy & Technology, 37(1), 32. https://doi.org/10.1007/s13347-024-00696-1

110. Tao, Y., Agrawal, A., Dombi, J., Sydorenko, T., & Lee, J. I. (2024). *ChatGPT role-play dataset: Analysis of user motives and model naturalness*. arXiv preprint arXiv:2403.18121. https://doi.org/10.48550/arXiv.2403.18121

111. Ruis, L., Khan, A., Biderman, S., Hooker, S., Rocktäschel, T., & Grefenstette, E. (2024). *The Goldilocks of pragmatic understanding: Fine-tuning strategy matters for implicature resolution by LLMs*. Advances in Neural Information Processing Systems, 36. Retrieved from https://arxiv.org/abs/2210.14986

112. Gilmore, A. (2019). *Materials and authenticity in language teaching*. In *The Routledge handbook of English language teacher education* (pp. 299–318). Routledge. https://doi.org/10.4324/9781315659824-25

113. Toncic, J. (2020, August 1). *Teachers, AI grammar checkers, and the newest literacies: Emending writing pedagogy and assessment*. Digital Culture & Education. Retrieved from https://doaj.org/article/69fe64fef90e46588b7e47e15dc1ba3c

114. Al-Ahdal, A. (2020, May 4). *Using computer software as a tool of error analysis: Giving EFL teachers and learners a much-needed impetus*. SSRN. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3570619

115. Ruder, S. (2022). *NLP-progress*.

116. Nguyen, H. U., Duong, L. N., & Pham, V. P. (2022). *Written corrective feedback strategies applied by Van Lang University's EFL lecturers in teaching online*. AsiaCALL OnlineJournal, 13(2), 21–41. https://doi.org/10.54855/acoj.221322

117. Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020). *wav2vec 2.0: A framework for self-supervised learning of speech representations*. arXiv preprint arXiv:2006.11477.

118. Zhang, Y., Park, D. S., Han, W., Qin, J., Gulati, A., Shor, J., Jansen, A., Xu, Y., Huang, Y., Wang, S., et al. (2021). *BigSSL: Exploring the frontier of large-scale semi-supervised learning for automatic speech recognition*. arXiv preprint arXiv:2109.13226.

119. Rogers, C. L., Dalby, J. M., & DeVane, G. (1994). *Intelligibility training for foreign-accented speech: A preliminary study*. The Journal of the Acoustical Society of America, 96(5), 3348. https://doi.org/10.1121/1.410623

120. Inceoglu, S., Lim, H., & Chen, W.-H. (2020). *ASR for EFL pronunciation practice: Segmental development and learners' beliefs*. The Journal of Asia TEFL, 17(3), 824–840.

121. Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023). *Robust speech recognition via large-scale weak supervision*. In *Proceedings of the International Conference on Machine Learning* (pp. 28492–28518).

122. Javed, T., Joshi, S., Nagarajan, V., Sundaresan, S., Nawale, J., Raman, A., Bhogale, K., Kumar, P., & Khapra, M. M. (2023). *Svarah: Evaluating English ASR systems on Indian accents*. In *Proceedings of Interspeech* (pp. 5087–5091). https://doi.org/10.21437/Interspeech.2023-2588

123. Chanethom, V., & Henderson, A. (2023). *Alignment in ASR and L1 listeners' recognition of L2 learner speech: French EFL learners & dictation*. Research in Language, 21, 245–266. https://doi.org/10.18778/1731-7533.21.3.03

124. Ballier, N., Arnold, T., Méli, A., Thurston, T., & Yunès, J.-B. (2024). *Whisper for L2 speech scoring*. International Journal of Speech Technology, 27, 923–934. https://doi.org/10.1007/s10772-024-10141-5

125. Ortega, L. (2009). *Understanding second language acquisition*. Routledge.

126. Seliger, H. W. (1977). *Does practice make perfect? A study of interaction patterns and L2 competence*. Language Learning, 27, 263–278.

127. Gardner, R. C. (2007, December 15). *Motivation and second language acquisition* [Conference presentation]. Seminario Sobre Plurilingüismo: Las Aportaciones Del Centro Europeo de Lenguas Modernas de Graz, Universidad de Alcalá, Spain.

128. Oxford, R., & Nyikos, M. (1989). *Variables affecting choices of language learning strategies by university students*. Modern Language Journal, 73, 291–300.

129. Victori, M., & Lockhart, W. (1995). *Enhancing metacognition in self-directed language learning*. System, 23, 223–234.

130. Graham, S. J. (2007). *Learner strategies and self-efficacy: Making the connection*. The Language Learning Journal, 35, 81–93.

131. Mercer, S., & Ryan, S. (2010). *A mindset for EFL: Learners' beliefs about the role of natural talent*. ELT Journal, 64, 436–444.

132. Lasagabaster, D., Doiz, A., & Sierra, J. M. (2014). *Introduction*. In D. Lasagabaster, A. Doiz, & J. M. Sierra (Eds.), *Motivation and foreign language learning: From theory to practice*. John Benjamins Publishing Company.

133. Colman, A. M. (2015). *Jost's law*. In *A dictionary of psychology*.

134. Ebbinghaus, H. (1885). *Über das Gedächtnis: Untersuchungen zur experimentellen Psychologie*. Duncker & Humblot.

135. Dempster, F. N. (1988). *The spacing effect: A case study in the failure to apply the results of psychological research*. American Psychologist, 43(8), 627–634. https://doi.org/10.1037/0003-066X.43.8.627

136. Hintzman, D. L. (1974). *Theoretical implications of the spacing effect*. In *Theories in cognitive psychology: The Loyola Symposium* (pp. 77–99). Lawrence Erlbaum.

137. Melton, A. W. (1970). *The situation with respect to the spacing of repetitions and memory*. Journal of Verbal Learning and Verbal Behavior, 9(5), 596–606. https://doi.org/10.1016/S0022-5371(70)80107-4

138. Underwood, B. J. (1970). *A breakdown of the total-time law in free-recall learning*. Journal of Verbal Learning and Verbal Behavior, 9(5), 573–580. https://doi.org/10.1016/S0022-5371(70)80104-9

139. Von Wright, J. (1971). *Effects of distributed practice and distributed recall tests on later recall of paired associates*. Journal of Verbal Learning and Verbal Behavior, 10(3), 311–315. https://doi.org/10.1016/S0022-5371(71)80060-9

140. Spitzer, H. F. (1939). *Studies in retention*. Journal of Educational Psychology, 30(9), 641–656. https://doi.org/10.1037/h0063404

141. Cain, L. F., & Willey, R. D. V. (1939). *The effect of spaced learning on the curve of retention*. Journal of Experimental Psychology, 25(2), 209–214. https://doi.org/10.1037/h0054640

142. Nakata, T. (2011). *Computer-assisted second language vocabulary learning in a paired-associated paradigm: A critical investigation of flashcard software*. Computer Assisted Language Learning, 24(1), 17–38. http://dx.doi.org/10.1080/09588221.2010.520675

143. Hirschel, R., & Fritz, E. (2013). *Learning vocabulary: CALL program versus vocabulary notebook*. System, 41(3), 639–653. https://doi.org/10.1016/j.system.2013.07.016

144. Bower, J. V., & Rutson-Griffiths, A. (2016). *The relationship between the use of spaced repetition software with a TOEIC word list and TOEIC score gains*. Computer Assisted Language Learning, 29(7), 1238–1248. http://dx.doi.org/10.1080/09588221.2016.1222444

145. Ono, T. (2017). *Vocabulary learning through computer assisted language learning*. Hitotsubashi Journal of Arts and Sciences, 58, 67–72.

146. Jorgensen, C. (2024). *Introducing spaced repetition software (SRS) for vocabulary acquisition in a university-level Arabic language course: A case study*. Journal of Language Teaching, 4(2), 33–42. https://doi.org/10.54475/jlt.2024.01

147. Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., … & He, Y. (2025). *Deepseek-r1: Incentivizing reasoning capability in LLMs via reinforcement learning*. arXiv preprint arXiv:2501.12948.

148. OpenAI. (2024a). *Hello GPT-4o*. Retrieved from https://openai.com/index/hello-gpt-4o/

149. OpenAI. (2024b). *Learning to reason with LLMs*. Retrieved from https://openai.com/index/learning-to-reason-with-llms/

150. Tabibian, B., Upadhyay, U., De, A., Zarezade, A., Schölkopf, B., & Gomez-Rodriguez, M. (2019). *Enhancing human learning via spaced repetition optimization*. Proceedings of the National Academy of Sciences of the United States of America, 116(10), 3988–3993. https://doi.org/10.1073/pnas.1815156116

151. Robinson, R., & Schneider, E. (2023). *Using an automated spaced repetition algorithm to enhance learning in Caribbean medical students: A pilot study*. In *EDULEARN23 Proceedings* (pp. 6380–6383). IATED.

152. Baillifard, A., Gabella, M., Banta Lavenex, P., & Martarelli, C. S. (2023). Implementing learning principles with a personal AI tutor: A case study. *arXiv preprint arXiv:2309.13060*.