

UNIVERSITÀ
DI PAVIA

FACOLTA' DI INGEGNERIA
DIPARTIMENTO DI INGEGNERIA INDUSTRIALE E DELL'INFORMAZIONE

CORSO DI LAUREA MAGISTRALE IN BIOINGEGNERIA

TESI DI LAUREA

Tecniche di intelligenza artificiale per l'estrazione strutturata di informazioni da testi clinici per il supporto all'aderenza delle prescrizioni terapeutiche

Candidato: Giacomo Marzero

Relatrice: Prof.ssa Lucia Sacchi
Correlatore: Prof. Enea Parimbelli

A.A. 2025/26

Riassunto

Lo scompenso cardiaco con frazione di eiezione ridotta (HFrEF) rappresenta una delle principali sfide per la sanità pubblica, caratterizzato da elevata morbilità e frequenti riospedalizzazioni. Le linee guida ESC raccomandano la quadruplice terapia farmacologica, ma nella pratica clinica persistono inerzia terapeutica e scarsa aderenza ai dosaggi target; la natura non strutturata delle lettere cliniche ostacola inoltre una valutazione sistematica della qualità delle cure. Nell'ambito del progetto GUIDE-AI, questa tesi propone una pipeline in tre fasi per supportare l'aderenza alle linee guida a partire da lettere di dimissione in lingua italiana. Nella prima fase le raccomandazioni ESC vengono formalizzate in workflow decisionali deterministici, con il coinvolgimento di un cardiologo (approccio *Expert-in-the-loop*). Nella seconda, un Large Language Model (LLM) estrae le informazioni rilevanti dal testo libero in uno schema JSON definito con Pydantic, sulla base della formalizzazione eseguita nella fase precedente. Nella terza, un modulo deterministico calcola automaticamente tre score di aderenza (mHFC, OPS e C4P). Il sistema è stato validato su 20 lettere reali anonimizzate dell'ICS Maugeri di Pavia, confrontando l'output di cinque LLM, sia *open-weight* sia proprietari, erogati tramite servizi cloud, con i dati registrati dai cardiologi in un form di riferimento. Quattro modelli su cinque hanno mostrato ottime prestazioni, con criticità residue isolate a singoli campi; le discrepanze negli score sono dipese esclusivamente da informazioni estratte in modo errato, mentre il calcolo deterministico è risultato corretto ogni volta che l'input lo era. I risultati rappresentano una prima verifica di fattibilità dell'approccio e individuano nell'estrazione il principale collo di bottiglia, evidenziando la necessità di un paradigma *Human-in-the-Loop* per l'impiego clinico. La struttura modulare facilita l'estensione ad altre patologie, lingue e centri.

Abstract

Artificial intelligence techniques for structured information extraction from clinical texts to support adherence to therapeutic prescriptions

Heart failure with reduced ejection fraction (HFrEF) is a major public health challenge, characterised by high morbidity and frequent rehospitalisations. The ESC guidelines recommend a four-pillar pharmacological therapy, yet in clinical practice therapeutic inertia and poor adherence to target doses persist, and the unstructured nature of clinical letters hampers a systematic assessment of the quality of care. Within the GUIDE-AI project, this thesis proposes a three-stage pipeline to support guideline adherence starting from discharge letters written in Italian. First, the ESC recommendations are formalised into deterministic decision workflows, with the involvement of a cardiologist (expert-in-the-loop approach). Second, a Large Language Model (LLM) extracts the relevant information from free text into a JSON schema defined with Pydantic, based on the formalisation carried out in the previous stage. Third, a deterministic module automatically computes three adherence scores (mHFC, OPS and C4P). The system was validated on 20 real, anonymised letters from ICS Maugeri (Pavia), comparing the output of five LLMs, both open-weight and proprietary, accessed through cloud services, with the data recorded by cardiologists in a reference form. Four out of five models showed excellent performance, with residual issues isolated to specific fields; discrepancies in the scores were due exclusively to incorrectly extracted information, whereas the deterministic calculation was correct whenever its input was. The results provide a first proof of feasibility of the approach and identify information extraction as the main bottleneck, highlighting the need for a human-in-the-loop paradigm for clinical use. The modular architecture facilitates extension to other diseases, languages and clinical centres.

Indice

Riassunto 2
Abstract 3
Indice 4
1. Introduzione e inquadramento del progetto 6
1.1. Il progetto *GUIDE-AI* 6
1.2 Obiettivi del lavoro di tesi 8
1.3 Struttura dell'elaborato 10
2. Stato dell'arte e background tecnologico 12
2.1 Scompenso cardiaco con frazione di eiezione ridotta 12
2.2 NLP e LLM nella cartella clinica elettronica 13
2.3 Estrazione strutturata da testi clinici 15
3. Modellazione del workflow Farmacologico 18
3.1 Mappatura delle linee guida europee 18
3.2 Definizione e formalizzazione del workflow 24
3.3 Consulenza clinica e perfezionamento del modello 37
4. Estrazione dei dati clinici tramite LLM 41
4.1 Descrizione del dataset di lettere cliniche 41
4.2 Progettazione dello Schema JSON 43
4.3 Architettura dell'estrattore basato su LLM 47
5. Calcolo automatico degli score di aderenza 54
5.1 Score di aderenza 54
5.2 Selezione degli score di aderenza 56
5.3 Estrazione multimodale dei dosaggi dalle linee guida 59
5.4 Algoritmo di calcolo automatico degli score 60
6. Risultati e discussione 63
6.1 Prestazioni del processo di estrazione di informazioni 63
6.2 Valutazione del sistema di scoring dell'aderenza 69
6.3 Esempio illustrativo: dalla lettera clinica agli score di aderenza 71
6.4 Analisi e discussione dei risultati 74

7. Conclusioni e sviluppi futuri 77
7.1 Sintesi del lavoro svolto 77
7.2 Integrazione del lavoro nelle fasi successive del progetto GUIDE-AI 79
7.3 Limiti del lavoro 82
7.4 Sviluppi futuri: verso la generazione automatica di raccomandazioni terapeutiche 83
8. Bibliografia 85
Appendice A: Form di annotazione di riferimento 90
Tabella A.1 90
Appendice B: prompt usati nella fase di estrazione 93
B.1 Prompt di sistema 93
B.2 Prompt per le informazioni demografiche 93
B.3 Prompt per le intolleranze 94
B.4 Prompt per gli esami di laboratorio 94
B.5 Prompt per le comorbidità 95
B.6 Prompt per disturbi e condizioni mediche 96
B.7 Prompt per l'estrazione di variabili binarie 96
B.8 Prompt per la data di dimissione 98
B.9 Prompt per le terapie farmacologiche 98
B.9.3 Specifica delle istruzioni sui singoli campi 99
Appendice C: Tabelle integrali degli score di aderenza 102
Tabella C.1 102
Tabella C.2 103
Tabella C.3 104

1. Introduzione e inquadramento del progetto

1.1. Il progetto *GUIDE-AI*

Ad oggi, le linee guida orientano la pratica clinica per la maggior parte delle patologie. Elaborati da società scientifiche ed esperti di settore sulla base delle migliori evidenze disponibili, questi documenti raccomandano i regimi farmacologici più idonei per ciascuna patologia, con l'obiettivo di standardizzare le cure e migliorare la qualità di vita del paziente. La documentazione cresce a ritmi sempre più sostenuti. L'intensa proliferazione delle direttive terapeutiche, spesso condensate in volumi di centinaia di pagine in costante aggiornamento, rende arduo l'apprendimento tempestivo da parte del personale medico, generando un divario marcato tra le raccomandazioni teoriche e la loro reale applicazione nel contesto assistenziale.

La sanità digitale ha risposto a tale disallineamento con i cosiddetti Sistemi di Supporto alle Decisioni Cliniche (*Clinical Decision Support Systems*, CDSS) [1], che nascono per affiancare il medico nel processo prescrittivo, mettendo a confronto il profilo dell'anamnesi del singolo individuo con le evidenze consolidate della letteratura scientifica. L'algoritmo assiste l'esperto senza mai sostituirlo. Operando come un co-pilota cognitivo, il sistema mira a contenere l'incidenza dell'errore umano ed elevare l'uniformità delle prestazioni erogate, salvaguardando l'autonomia decisionale del professionista.

I CDSS tradizionali mostrano limiti strutturali. Fondati su basi di conoscenza deterministiche e su rigide regole d'inferenza di tipo *if-then* programmate manualmente, i modelli basati sulla conoscenza storici faticano a stare al passo con domini clinici in continua evoluzione, soffrono di una ridotta scalabilità e risultano incapaci di elaborare il dato testuale non strutturato [2]. Tali caratteristiche sono proprio quelle che caratterizzano i testi clinici. L'impossibilità di processare la massa critica di testo libero che caratterizza le linee guida e le cartelle cliniche quotidiane ha spinto il settore verso l'intelligenza artificiale, il machine learning e i modelli linguistici di grandi dimensioni (Large Language Models, LLM) fondati su tecniche avanzate di Elaborazione del Linguaggio Naturale (NLP).

L'intelligenza artificiale è capace di estrapolare le informazioni utili dal testo. Convertendo la prosa clinica delle lettere di dimissione in strutture dati formalizzate come schemi JSON, le nuove architetture consentono di verificare l'effettiva aderenza della terapia somministrata rispetto alle raccomandazioni vigenti, contestualizzandola sul profilo specifico del paziente. Lo specialista resta al centro della decisione. Questo approccio, infatti, si sviluppa secondo la logica *Human-in-the-Loop*, in cui l'elaboratore fornisce suggerimenti e motivazioni lasciando intatta la responsabilità prescrittiva in capo all'operatore sanitario. Tuttavia, la pratica anticipa spesso la regolamentazione. Nonostante l'assenza di un formale riconoscimento da parte degli enti regolatori o delle società scientifiche, si osserva una diffusione spontanea di strumenti generativi generalisti (non specifici per il dominio medico) come ChatGPT e Gemini, affiancati da piattaforme verticali per la medicina, sia tra i pazienti sia tra i clinici [3].

In questo contesto, il progetto GUIDE-AI (Guiding your treatment with current evidence), finanziato dall'iniziativa IHI (Innovative Health Initiative) nell'ambito del programma Horizon Europe dell'Unione Europea [4], si pone l'obiettivo di sviluppare soluzioni tecnologiche d'avanguardia per la gestione delle patologie croniche ad alto impatto sociosanitario. L'obiettivo finale consiste nello sfruttare l'IA generativa per incrementare la percentuale di pazienti trattati in conformità con le direttive ufficiali. Nasce la figura dei Guideline Navigator. Tali sistemi avanzati di supporto verificano l'aderenza delle prescrizioni farmacologiche alle linee guida e mettono a disposizione moduli di comunicazione per spiegare al paziente, in linguaggio comune, il razionale alla base delle scelte terapeutiche.

L'iniziativa si snoda su un arco quadriennale guidato da tre pilastri strategici:

- **Obiettivo A [*people-centricity*]**: La centralità della persona guida la progettazione. Coinvolgendo attivamente pazienti e medici nella co-progettazione dell'architettura, questo obiettivo mira a identificare gli ostacoli all'aderenza terapeutica e a realizzare un'interfaccia trasparente capace di promuovere la comprensione condivisa delle scelte di cura.

- **Obiettivo B [Innovate & Integrate]**: L'integrazione tecnologica richiede validazione e robustezza. Il focus risiede nello sviluppo di Guideline Navigator basati su LLM e architetture di Retrieval-Augmented Generation (RAG) [5], direttamente integrabili nei sistemi di cartella clinica elettronica; la valutazione dell'accuratezza comprende il benchmarking su dati retrospettivi e sessioni di *red teaming* per saggiare la resistenza ad attacchi avversariali.
- **Obiettivo C [evaluate in real world]**: L'impatto clinico viene testato sul campo. Attraverso uno studio prospettico, randomizzato e multicentrico in contesti ospedalieri e ambulatoriali, il progetto misura come risultato primario la percentuale di aderenza terapeutica e, come variabili secondarie, le variazioni di schema posologico, l'incidenza di eventi avversi, la qualità della vita, l'esperienza d'uso e la mortalità.

La fase iniziale mappa tre patologie croniche. Il framework concentra le proprie attività preliminari su un nucleo ad alto impatto sociosanitario che comprende la broncopneumopatia cronica ostruttiva, lo scompenso cardiaco cronico e l'insufficienza renale cronica [6], per poi estendersi gradualmente ad altre condizioni cliniche o recepire i periodici aggiornamenti delle linee guida esistenti. L'indagine svolta all'interno della presente tesi circoscrive il proprio raggio d'azione alla gestione dello scompenso cardiaco con frazione di eiezione ridotta (*Heart Failure with reduced Ejection Fraction*, HFrEF), identificandolo quale caso di studio per l'implementazione dei modelli sviluppati.

1.2 Obiettivi del lavoro di tesi

Nel contesto del progetto GUIDE-AI, l'HFrEF rappresenta un banco di prova ideale. La spiccata complessità che ne caratterizza la gestione farmacologica rende questa patologia lo scenario d'elezione per saggiare sul campo le capacità analitiche e l'affidabilità del sistema sviluppato. La terapia si fonda sul paradigma della Terapia Medica Guidata dalle Linee Guida (*Guideline-Directed Medical Therapy*, GDMT) [7], un approccio combinato basato sull'azione sinergica dei quattro pilastri farmacologici che compongono la cosiddetta quadruplice terapia.

L'attuazione clinica incontra enormi ostacoli. La necessità di modulare continuamente la titolazione dei dosaggi, il monitoraggio costante del quadro emodinamico e l'incessante aggiornamento della letteratura scientifica alimentano un divario marcato tra le raccomandazioni teoriche e la reale pratica prescrittiva. I dati reali confermano questo problema. Le evidenze raccolte nello studio CHAMP-HF testimoniano l'entità di questo scarto, rivelando come appena l'1% dei pazienti riceva la combinazione farmacologica prevista dalla GDMT ai dosaggi pienamente titolati [8].

La tesi affronta questa criticità. Inquadrato nell'ambito del progetto GUIDE-AI, il presente lavoro sviluppa una pipeline automatizzata concepita per convertire le informazioni non strutturate presenti nei referti clinici in metriche quantitative e misurabili, al fine di stimare con precisione l'aderenza della pratica medica reale rispetto alle raccomandazioni ufficiali. Le tre fasi della pipeline sono strettamente collegate: dalla formalizzazione della linea guida si ricavano le informazioni utili alla gestione farmacologica della patologia, che a loro volta definiscono i dati da estrarre dalle lettere cliniche, a partire dai quali vengono calcolati gli score.

Nello specifico, il lavoro di tesi si articola nei seguenti tre obiettivi operativi:

1. **Modellizzazione della Linea Guida**: tradurre le raccomandazioni contenute nella linea guida europea di riferimento per la terapia farmacologica dell'HFrEF in un workflow strutturato, contenente tutte le informazioni di interesse in forma sintetizzata. Questa rappresentazione è stata affinata e validata tramite il consulto diretto con medici specialisti in cardiologia, in modo tale da garantire il rispetto delle esigenze della reale pratica clinica. Il workflow ha inoltre la funzione di individuare le informazioni da estrarre dalle lettere cliniche nella fase successiva del lavoro.
2. **Estrazione strutturata dei dati clinici tramite LLM**: focalizzare l'analisi sulla quadruplice terapia ed estrarre informazioni strutturate contenute nelle lettere cliniche di pazienti con diagnosi di HFrEF tramite l'applicazione di diversi LLM. Tali informazioni vengono estratte secondo uno specifico schema strutturato (JSON), definito a partire dalle informazioni individuate nel workflow dell'obiettivo precedente.

3. **Calcolo degli score di aderenza terapeutica**: utilizzare le informazioni estratte in forma strutturata nello step precedente per calcolare degli score di aderenza, calcolati sulla base delle categorie di farmaci inclusi nella terapia e delle relative quantità somministrate rispetto alle dosi riportate nella linea guida. Questo punteggio fornisce al medico un indicatore quantitativo sintetico e immediato per valutare in tempo reale la qualità della terapia prescritta al paziente.

In sintesi, l'integrazione di questi tre passaggi operativi costituisce una soluzione al divario tra la teoria delle linee guida e la pratica clinica reale. L'intera pipeline si pone come uno strumento di supporto decisionale volto a ridurre l'inerzia terapeutica e a guidare l'ottimizzazione della cura del paziente.

1.3 Struttura dell'elaborato

L'elaborato è strutturato in 6 capitoli (senza considerare il presente capitolo di introduzione) che illustrano il lavoro di ricerca svolto:

- **Capitolo 2 - Stato dell'arte e background tecnologico**: approfondisce il tema dell'HFrEF, le relative linee guida europee e lo stato dell'arte riguardante gli LLM in ambito sanitario.
- **Capitolo 3 - Modellazione del workflow farmacologico**: descrive nel dettaglio la formalizzazione del workflow di trattamento farmacologico dell'HFrEF, a partire dalle relative linee guida europee.
- **Capitolo 4 - Estrazione strutturata dei dati clinici**: presenta la struttura dello schema JSON, definita a partire dalle variabili individuate nel Capitolo 3, scelta per l'estrazione dei dati in formato strutturato a partire dalle lettere cliniche, gli LLM selezionati per l'estrazione e l'algoritmo utilizzato.
- **Capitolo 5 - Calcolo degli score di aderenza**: approfondisce il tema degli score di aderenza per l'HFrEF, presenta gli score selezionati per questa parte del lavoro di tesi e descrive l'algoritmo utilizzato per il calcolo automatico a partire dallo schema JSON estratto nello step precedente.
- **Capitolo 6 - Risultati**: analizza i risultati ottenuti, soprattutto quelli relativi all'estrazione dei dati dalle lettere cliniche e al calcolo dei relativi score di

aderenza, confrontando le performance dei diversi LLM utilizzati per l'estrazione.

- **Capitolo 7 - Conclusioni e sviluppi futuri**: evidenzia i punti di forza e le limitazioni del sistema sviluppato, a partire dai risultati ottenuti, e presenta i possibili sviluppi futuri nel quadro del progetto GUIDE-AI.

Il codice sviluppato per questo lavoro di tesi (estrazione delle informazioni, estrazione multimodale della tabella delle dosi target e calcolo degli score di aderenza) è disponibile pubblicamente su GitHub [9], nei notebook *TESI_IE*, *TESI_multimodal_extraction* e *TESI_GDMT_scores*.

2. Stato dell’arte e background tecnologico

2.1 Scompenso cardiaco con frazione di eiezione ridotta

Lo scompenso cardiaco configura una sindrome eterogenea. Tale condizione si manifesta attraverso un quadro clinico articolato in cui sintomi cardinali, quali dispnea e affaticamento, si combinano a riscontri obiettivi che includono aumento della pressione venosa, rantoli polmonari e edema periferico. L'origine può essere strutturale oppure funzionale. Tali alterazioni compromettono la dinamica cardiaca, determinando l’aumento delle pressioni intracardiache e/o una riduzione della gittata cardiaca. Comunemente lo scompenso cardiaco è causato da una disfunzione miocardica. Tuttavia, il quadro eziologico include anche valvulopatie, patologie pericardiche ed endocardiche, nonché aritmie e disturbi della conduzione che possono innescare autonomamente la sindrome o aggravarne il decorso [7].
L'impatto epidemiologico rimane globale e devastante. I dati evidenziano circa 64 milioni di casi annui a livello mondiale, con una popolazione affetta che per circa la metà presenta il fenotipo a frazione di eiezione ridotta. La prognosi dell'HFrEF è particolarmente severa. In questo sottoinsieme di pazienti, il decorso clinico mostra tassi di mortalità estremamente elevati, prossimi al 75% a distanza di cinque anni dalla diagnosi [10], [11].

La stratificazione fenotipica dell'insufficienza cardiaca si fonda storicamente sulla misurazione della frazione di eiezione ventricolare sinistra (*Left Ventricular Ejection Fraction*, LVEF), parametro ecocardiografico che definisce diversi profili clinici [7]. In particolare, tre fenotipi distinti guidano la diagnosi. Lo scompenso con frazione di eiezione preservata (*Heart Failure with preserved Ejection Fraction*, HFpEF) combina la manifestazione sintomatologica con una LVEF pari o superiore al 50%. La frazione di eiezione appare conservata. Nel fenotipo con frazione di eiezione lievemente ridotta (Heart Failure with mildly reduced Ejection Fraction, HFmrEF) invece, i valori campionano l'intervallo compreso tra il 41% e il 49%. L'HFrEF segna il quadro peggiore. Questa condizione, caratterizzata da una LVEF inferiore o uguale al 40%, incarna la forma più aggravata della patologia.

L'HFrEF costituisce la condizione più difficile da trattare. Negli ultimi anni, la gestione farmacologica di questa variante ha attraversato un vero e proprio cambio di paradigma, culminato con la formalizzazione nelle linee guida della European Society of Cardiology (ESC) della quadruplice terapia, fondata sulla titolazione simultanea e parallela di quattro distinte classi farmacologiche. La pratica clinica delude le aspettative. A fronte di un netto e comprovato beneficio in termini di riduzione della mortalità, appena l'1% dei pazienti riceve attualmente la combinazione terapeutica prevista dagli standard internazionali [8].

2.2 NLP e LLM nella cartella clinica elettronica

Le cartelle cliniche racchiudono l'intera storia del paziente attraverso flussi informativi fortemente eterogenei, distribuiti tra un nucleo di dati strutturati, come le codifiche ICD (*International Classification of Diseases*) o gli esiti di laboratorio, e un'importante componente informale. La componente narrativa prevale nella cartella. Si stima infatti che il testo libero costituisca circa l'80% dell'intero patrimonio informativo clinico [12], manifestandosi in verbali d'esame, referti specialistici e, in prima linea, nelle lettere di dimissione. Questi documenti sintetizzano l'esperienza clinica. In essi sono contenuti la caratterizzazione del quadro patologico, il monitoraggio temporale dei parametri strumentali e le prescrizioni farmacologiche elaborate dal medico. Tuttavia, il formato narrativo ostacola l'analisi. Ciò che garantisce espressività clinica costituisce al tempo stesso un ostacolo primario per le architetture di estrazione automatizzata dei dati.

L'analisi del linguaggio medico si scontra con una pronunciata eterogeneità sintattica e con l'uso frequente di abbreviazioni non standardizzate. Ogni medico adotta una propria impostazione stilistica. Ciò genera variazioni lessicali e strutturali profonde tra documenti del tutto sovrapponibili per contenuto clinico. Inoltre, l'ambiguità delle sigle amplifica il problema. Un medesimo acronimo può assumere significati diversi in base al contesto semantico o alla specializzazione: la sigla «FA», ad esempio, si riferisce sia a fibrillazione atriale sia a frequenza arteriosa. Infine, il gergo di reparto complica il quadro. L'impiego capillare di abbreviazioni informali, ristrette alla prassi locale del singolo contesto operativo, priva il testo di riferimenti univoci e invalida le

regole d'interpretazione convenzionali.

L'esigenza di strutturare la componente testuale ha guidato l'evoluzione dell'Elaborazione del Linguaggio Naturale (*natural language processing*, NLP) dai sistemi deterministici verso i modelli probabilistici e generativi di grande scala. L'approccio iniziale fu strettamente deterministico. Tra gli anni '90 e i primi anni 2000, i sistemi clinici si basavano su grammatiche formali, espressioni regolari e dizionari ontologici integrati come UMLS (*Unified Medical Language System*) e SNOMED CT. Architetture storiche come cTAKES (*clinical Text Analysis and Knowledge Extraction System*) [13] combinavano pipeline di analisi grammaticale con algoritmi euristici per l'individuazione della negazione, tra cui NegEx [14]. Questi sistemi offrivano totale interpretabilità. L'assenza di allucinazioni e il comportamento deterministico rappresentavano punti di forza certi. Tuttavia, la rigidità che li caratterizzava ne decretò il limite. L'infrastruttura richiedeva un imponente lavoro di codifica manuale, rivelando una marcata fragilità di fronte alle deviazioni stilistiche, ai refusi testuali e ai costrutti sintattici non previsti dall'algoritmo.

L'avvento dell'architettura *Transformer* ha ridefinito la disciplina introducendo i meccanismi di *self-attention*, capaci di catturare le dipendenze contestuali a lungo raggio nel testo. In questi modelli il testo viene scomposto in *token*, corrispondenti a singole parole o a porzioni di esse, in modo tale da ricondurre anche i termini rari, come il nome di un farmaco, a componenti già note al modello; dopodiché, ogni token è convertito in un vettore (*embedding*); tale rappresentazione numerica ne racchiude il significato. Tramite il meccanismo della *self-attention*, per ogni token viene calcolato quanto ogni altro elemento della sequenza sia rilevante ai fini della sua interpretazione, e il vettore associato viene aggiornato di conseguenza. In questo modo si ottiene una rappresentazione contestuale. La stessa parola, infatti, può essere associata a vettori diversi in base al contesto nel quale è collocata. In questo meccanismo consiste un netto cambio di paradigma rispetto agli approcci che associavano ad ogni parola una rappresentazione fissa e indipendente dal contesto [15]. Su questa base si è affermato BERT (*Bidirectional Encoder Representations from Transformers*), capace di estrarre rappresentazioni semantiche complesse

mediante un pre-addestramento su imponenti corpora [16]. Il dominio biomedico ha richiesto adattamenti. Lo sviluppo di modelli specializzati, pre-addestrati su letteratura scientifica come BioBERT [17], o direttamente su note cliniche come ClinicalBERT [18], ha garantito prestazioni elevate nei task svolti a valle del pre-addestramento. Tali architetture eccellono nel riconoscimento di entità nominate (*Named Entity Recognition*, NER) per l'identificazione di patologie e farmaci, oltre che nell'estrazione di relazioni (*Relation Extraction*, RE) per la mappatura dei concetti correlati. I dati annotati restano un ostacolo. L'efficacia operativa rimane vincolata alla disponibilità di ampi dataset etichettati, indispensabili per la fase di fine-tuning specifica per ciascun compito.

L'attuale stato dell'arte nell'elaborazione del linguaggio naturale è rappresentato dagli LLM [3], modelli dotati di miliardi di parametri, addestrati con tecniche di modellazione linguistica autoregressiva (*Auto-Regressive Language Modeling*) e rifiniti successivamente tramite apprendimento per rinforzo (*Reinforcement Learning*). Questi modelli, tramite il paradigma dell'apprendimento contestuale (*In-Context Learning*), grazie ad un'istruzione in linguaggio naturale (prompt) e al testo da analizzare, sono in grado di eseguire diversi compiti, senza un addestramento specifico per ciascuno. In ambito clinico le possibili applicazioni sono numerose [3]: dal fornire risposte legate a documenti clinici a compiti di sintesi, dalla classificazione di pazienti alla traduzione, fino all'estrazione strutturata di informazioni, ovvero l'information extraction (IE), su cui si basa questo lavoro di tesi.

2.3 Estrazione strutturata da testi clinici

L'IE consiste nell'estrarre le informazioni contenute nel testo libero e nell'organizzarle in un formato strutturato specifico, in modo da poterle riutilizzare in un secondo momento [19]. Si distingue dal semplice riconoscimento di entità nominate (NER), che si limita a individuare nel testo i termini di interesse, perché richiede anche di associare tra loro le informazioni correlate. Con gli LLM questo compito può essere eseguito in due modalità [20]: *zero-shot*, con la quale l'estrazione viene eseguita fornendo unicamente l'istruzione formale e il testo della lettera clinica, oppure *few-shot*, arricchendo il prompt con esempi rappresentativi di corrispondenza tra l'input

fornito e l'output desiderato, guidando così l'LLM nell'estrazione.

Per quanto riguarda la gestione dell'HFrEF nello specifico, l'LLM deve essere in grado di estrarre informazioni strettamente correlate tra loro. Per esempio, per ogni farmaco somministrato devono essere estratti nome, principio attivo, numero di somministrazioni giornaliere ecc.
Mentre l'NLP tradizionale risolveva questo problema separando la fase di estrazione da quella di associazione delle relazioni tra le entità, gli LLM consentono di eseguire l'intera operazione di IE in un unico passaggio, utilizzando come input il testo clinico e il prompt fornito dall'utente (inclusa la struttura di estrazione dei dati) e fornendo in output un oggetto strutturato secondo lo schema sintattico e semantico richiesto, che, nel caso di questo lavoro di tesi, consiste in uno specifico schema JSON.

Un vincolo essenziale per l'utilizzo di un LLM all'interno di una pipeline automatizzata è che l'output del modello rispetti la struttura imposta dall'utente. Elementi non strutturati infatti renderebbero impossibile l'elaborazione automatica da parte dei moduli successivi all'estrazione. A tale scopo, è importante descrivere il formato dell'output desiderato direttamente all'interno del prompt, imponendo esplicitamente di evitare di generare testo introduttivo o conclusivo che potrebbe complicare l'utilizzo automatico dell'output generato. In alternativa, è possibile interrogare gli LLM tramite API (*Application Programming Interface*), ovvero interfacce di programmazione che permettono ad un software di inviare direttamente il prompt al modello e di riceverne la risposta senza la necessità dell'interfaccia di chat pensata per l'interazione con un utente umano. Le API più recenti supportano modalità di generazione strutturata, quali *JSON Mode* e *Structured Outputs* [21]. Nel primo caso, si garantisce che l'output sia in formato JSON valido, ma non che rispetti necessariamente una determinata struttura; nel secondo invece, si fornisce al modello uno schema strutturato che definisce esattamente la forma che deve avere la risposta. In ambito clinico, l'operazione di IE ad opera di LLM è già stata valutata, in particolare sui referti radiologici, come sintetizzato da una recente rassegna che ha mappato i principali approcci di estrazione [22]. In particolare, Adams et al. hanno valutato GPT-4 nella conversione di referti TC e RM in report strutturati [23]. Woźnicki et al. hanno

invece utilizzato un modello open-source eseguito in locale (Llama-2-70B-chat), controllandone l'output per garantire la validità dei report strutturati e ottenendo un'accuratezza prossima a quella di lettori umani su referti radiografici in inglese e tedesco [24]. Tuttavia, l'incapacità degli LLM di garantire sistematicamente la generazione di strutture JSON prive di errori [25] ha spinto la letteratura recente verso l'adozione di tecniche di generazione vincolata [21].

All'interno della pipeline di tesi, i modelli LLM giocano un ruolo centrale nell'estrazione dei dati utili dalle lettere cliniche e nella loro trasformazione in un oggetto JSON valido. La struttura di questo schema è volta a contenere tutte le informazioni di interesse relative al trattamento farmacologico indicato nelle linee guida ESC per l'HFrEF e funge da ponte tra l'informazione destrutturata contenuta nel testo libero e l'algoritmo di calcolo degli score di aderenza, consentendo a quest'ultimo di operare su dati strutturati e privi delle ambiguità tipiche del linguaggio naturale.

3. Modellazione del workflow Farmacologico

3.1 Mappatura delle linee guida europee

Le linee guida rilasciate dalla European Society of Cardiology (ESC) fino al 2016 [26] strutturavano il trattamento su un modello sequenziale, prevedendo l'avvio prioritario di un inibitore dell'enzima di conversione dell'angiotensina (*Angiotensin-Converting Enzyme inhibitor*, ACE-i), o di un antagonista del recettore dell'angiotensina II (*Angiotensin Receptor Blocker*, ARB) in caso di intolleranza, affiancato da un beta-bloccante, entrambi da titolare fino alla dose target o alla massima dose tollerata dal paziente. L'introduzione delle altre classi farmacologiche era differita nel tempo. Gli antagonisti del recettore dei mineralcorticoidi (*Mineralocorticoid Receptor Antagonists*, MRA), ad esempio, subentravano solo di fronte a una sintomatologia persistente e a una riduzione severa della funzione di eiezione, con LVEF≤35%.

Questo schema generava ritardi. La necessità di completare ciascuno step prima di passare al successivo dilatava la messa a regime della terapia per diversi mesi, privando i pazienti di un'adeguata protezione e lasciandoli esposti a un elevato rischio di complicanze.

La svolta determinata dalle ultime versioni delle linee guida è stata radicale [7]. Le direttive ESC del 2021 [7], consolidate dal Focused Update del 2023 [27], hanno abbandonato la sequenzialità a favore della somministrazione simultanea. I trial clinici hanno infatti dimostrato l'impatto decisivo degli MRA [28], [29] e degli inibitori del cotrasportatore sodio-glucosio di tipo 2 (*Sodium-Glucose Co-Transporter 2 Inhibitors*, SGLT2i) [30], [31] nell'abbattere la mortalità e arrestare la progressione della patologia, promuovendo l'introduzione tempestiva dei quattro pilastri farmacologici (inibitori del recettore dell'angiotensina e della neprilisina, *Angiotensin Receptor-Neprilysin Inhibitors*, ARNI, in alternativa ACE-i o ARB; beta-bloccanti; MRA; SGLT2i) subito dopo la diagnosi, per poi condurre la titolazione in parallelo fino al target desiderato.

La gestione è diventata complessa. Se l'adozione della quadruplice terapia ha ridotto sensibilmente i ricoveri e la mortalità, la sua implementazione impone un carico decisionale elevato. Il clinico deve orchestrare la titolazione contemporanea di quattro molecole, garantendo un monitoraggio stretto della pressione arteriosa, della potassiemia e della velocità di filtrazione glomerulare per adattare i dosaggi alle risposte emodinamiche e renali del singolo individuo.

I quattro pilastri agiscono in sinergia:

- **Inibitori dell'enzima di conversione dell'angiotensina** (*Angiotensin-Converting Enzyme Inhibitors*, ACE-i), inibitori del recettore dell'angiotensina e della neprilisina (*Angiotensin Receptor-Neprilysin Inhibitors*, ARNI) e antagonisti del recettore dell'angiotensina II (*Angiotensin Receptor Blockers*, ARB): fanno parte della classe degli inibitori del sistema renina-angiotensina-aldosterone (*Renin-Angiotensin-Aldosterone System inhibitors*, RAASi). La categoria più efficace è quella degli ARNI. Questa classe riduce le resistenze vascolari e protegge il miocardio agendo sull'asse renina-angiotensina. Storicamente, gli ACE-i sono stati i primi farmaci a dimostrare una riduzione tangibile della mortalità e un miglioramento sintomatico [32], [33], [34], [35]. Gli ARNI ne hanno successivamente superato le prestazioni, abbattendo ulteriormente le riospedalizzazioni e i decessi [36], preservando la velocità di filtrazione glomerulare stimata (*estimated Glomerular filtration Rate*, eGFR) [37], riducendo l'incidenza di iperkaliemia [38] e limitando persino la necessità di insulina nei pazienti diabetici [39]. Gli ARB rappresentano l'alternativa di seconda scelta in presenza di intolleranza [40].

  Il protocollo prevede l'impiego di una sola opzione all'interno della triade, accordando massima priorità agli ARNI, seguiti dagli ACE-i e, laddove non tollerati, dagli ARB.
- **Beta-bloccanti** (*Beta-blockers*, BBs): i beta-bloccanti rallentano pressione e frequenza cardiaca. Così facendo, questa classe allevia la sintomatologia [41] e garantisce un chiaro beneficio in termini di sopravvivenza se somministrata in combinazione con i RAASi [42], [43].
- **Antagonisti del recettore dei mineralcorticoidi** (*Mineralocorticoid Receptor*

Antagonists, MRA): gli MRA contrastano la fibrosi. Bloccando i recettori per l'aldosterone e per gli altri ormoni steroidei, queste molecole limitano i processi infiammatori miocardici e sono raccomandate in associazione a RAASi e Beta-bloccanti per attenuare i sintomi e ridurre i ricoveri per insufficienza cardiaca [44], [45].

- **Inibitori del cotrasportatore sodio-glucosio di tipo 2** (*Sodium-Glucose Co-Transporter 2 Inhibitors*, SGLT2i): Gli SGLT2i modulano l'escrezione renale. Favorendo l'eliminazione di glucosio e sodio per via urinaria, questi farmaci riducono il carico volemico e la pressione arteriosa. Lo studio DAPA-HF ne ha dimostrato l'impatto con dapagliflozin nel ridurre la mortalità globale e migliorare la capacità funzionale dei pazienti sintomatici [30], mentre lo studio EMPEROR-Reduced ha confermato che empagliflozin riduce in modo significativo il rischio di eventi cardiovascolari avversi e le ospedalizzazioni, migliorando contestualmente la qualità di vita [46], [31].

La Tabella 1, riportata di seguito, presenta tutti i farmaci previsti dalle linee guida ESC per ognuna delle categorie della quadruplice terapia, con le relative dosi di partenza e dosi di titolazione target:

Classe / Principio Attivo	**Dose Iniziale**	**Dose Target**
ACE-i		
Captopril	6.25 mg *t.i.d.*	50 mg *t.i.d.*
Enalapril	2.5 mg *b.i.d.*	10–20 mg *b.i.d.*
Lisinopril	2.5–5 mg *o.d.*	20–35 mg *o.d.*
Ramipril	2.5 mg *b.i.d.*	5 mg *b.i.d.*
Trandolapril	0.5 mg *o.d.*	4 mg *o.d.*
ARNI		
Sacubitril / Valsartan	49/51 mg *b.i.d.*	97/103 mg *b.i.d.*
Beta-bloccanti		
Bisoprololo	1.25 mg *o.d.*	10 mg *o.d.*
Carvedilolo	3.125 mg *b.i.d.*	25 mg *b.i.d.*
Metoprololo succinato (CR/XL)	12.5–25 mg *o.d.*	200 mg *o.d.*

Classe / Principio Attivo	Dose Iniziale	Dose Target
Nebivololo	1.25 mg *o.d.*	10 mg *o.d.*
MRA		
Eplerenone	25 mg *o.d.*	50 mg *o.d.*
Spironolattone	25 mg *o.d.*	50 mg *o.d.*
Inibitori SGLT2		
Dapagliflozin	10 mg *o.d.*	10 mg *o.d.*
Empagliflozin	10 mg *o.d.*	10 mg *o.d.*
Classe / Principio Attivo	**Dose Iniziale**	**Dose Target**
Altri Agenti		
Candesartan	4 mg *o.d.*	32 mg *o.d.*
Losartan	50 mg *o.d.*	150 mg *o.d.*
Valsartan	40 mg *b.i.d.*	160 mg *b.i.d.*
Ivabradina	5 mg *b.i.d.*	7.5 mg *b.i.d.*
Vericiguat	2.5 mg *o.d.*	10 mg *o.d.*
Digossina	62.5 µg *o.d.*	250 µg *o.d.*
Idralazina / Isosorbide Dinitrato	37.5 mg *t.i.d.* / 20 mg *t.i.d.*	75 mg *t.i.d.* / 40 mg *t.i.d.*

Tabella 1. *Farmaci della quadruplice terapia: dosi iniziali e target per il trattamento dell'HFrEF. Adattata dalle linee guida ESC 2021 [7].*

Le informazioni relative ai dosaggi contenute in Tabella 1 sono fondamentali per le fasi successive del lavoro di tesi. La presenza nella terapia di determinate categorie di farmaci e la discrepanza tra la dose effettivamente prescritta nella lettera clinica e la dose target definita in tabella costituiscono le variabili in base alle quali verranno calcolati gli score di aderenza (Capitolo 5).

Si noti come la classe SGLT2i sia l'unica per cui non avviene la titolazione, bensì viene somministrata una dose fissa. La dose iniziale e la dose target mostrate in Tabella 1 infatti sono le stesse.

Definito il quadro clinico della quadruplice terapia, la sua traduzione computazionale all'interno del progetto GUIDE-AI richiede di formalizzare le indicazioni operative

contenute nelle linee guida ESC 2021 [7] e nel relativo *Focused Update* del 2023 [27]. L'obiettivo di questa mappatura è quello di costruire un diagramma di flusso decisionale strutturato utile a sintetizzare la complessa casistica descrittiva e a semplificare gli step successivi del lavoro di tesi, individuando le variabili utili da estrarre nel passaggio di IE che dovranno essere utilizzate successivamente per calcolare gli score di aderenza.

Rispetto all'ampio dominio coperto dai documenti ESC, che include l'intero spettro dell'insufficienza cardiaca (compresi i fenotipi HFmrEF ed HFpEF, trattati nel Sottocapitolo 2.1) e spazia dal percorso diagnostico alle terapie elettriche e interventistiche, il progetto GUIDE-AI è circoscritto al solo trattamento farmacologico delle patologie trattate. In particolare, il presente lavoro di tesi si concentra sulla gestione della quadruplice terapia dell'HFrEF. La formalizzazione, dunque, si focalizza sulla gestione della terapia medica, con particolare attenzione alla gestione dei farmaci appartenenti alle classi dei quattro pilastri farmacologici.

Come anticipato nel capitolo 2.1, con il passaggio da un approccio sequenziale a step all'introduzione dei quattro pilastri, la terapia farmacologica dell'HFrEF prevede la gestione della somministrazione di quattro classi farmacologiche in parallelo. Per ciascuna è riportata una guida pratica di utilizzo sotto forma di tabella supplementare all'interno di un file di supporto alla linea guida principale [7]. Ognuna di queste tabelle contiene specifiche sezioni, che affrontano diverse sfide legate alla gestione della terapia. Alcune di queste, riportate di seguito, sono il riferimento centrale per la comprensione della terapia farmacologica e del relativo workflow:

- ***IN WHOM AND WHEN?***: elenca le possibili problematiche legate all'inizio della somministrazione, da tenere in considerazione prima che questa sia iniziata. In particolare, queste situazioni si dividono in due ulteriori gruppi:
    - ***Contraindications***: controindicazioni assolute, sufficienti da sole a sconsigliare l'inizio della terapia.
    - ***Cautions/seek specialist advice***: è consigliato il parere dello specialista; starà a quest'ultimo decidere se la terapia possa incominciare o meno.

- ***HOW TO USE?***: descrive come somministrare il farmaco nel corso della terapia; vengono indicati i tempi tra un esame di laboratorio e il successivo così come i tempi tra una titolazione e l'altra. Tuttavia, l'architettura GUIDE-AI mira all'assistenza immediata. Gran parte dei dettagli contestuali risulta quindi ridondante per l'operatività del sistema. La sezione conserva tuttavia un valore fondamentale in rari casi limite in cui vengono fornite informazioni preziose riguardo a come procedere nel caso in cui i risultati degli esami di laboratorio eccedano specifici valori soglia.
- ***PROBLEM SOLVING***: questa sezione affronta le complicanze. Il modulo orienta il clinico nella gestione degli eventi avversi insorti durante il trattamento, come la comparsa di ipotensione arteriosa. Le risposte terapeutiche previste spaziano dal de-scaling posologico all'introduzione di farmaci correttivi, fino alla sospensione definitiva della terapia nelle situazioni più severe. La tempistica definisce la differenza. A differenza del blocco *IN WHOM AND WHEN?*, incentrato sulle criticità da risolvere prima di avviare il trattamento, le strategie di *PROBLEM SOLVING* intervengono esclusivamente a corso terapeutico già avviato.

Cinque parametri guidano il monitoraggio. Oltre alle controindicazioni proprie di ciascuna classe molecolare, le istruzioni operative per la gestione dei farmaci si basano su un set ristretto di variabili fisiologiche:

- **Pressione arteriosa sistolica** (systolic blood pressure, SBP): questo parametro costituisce l'indicatore emodinamico primario per calibrare e modulare l'impiego delle molecole ad azione vasodilatatrice. Valori di SBP < 90 mmHg o la presenza di ipotensione sintomatica costituiscono il principale vincolo di diminuzione o arresto della titolazione.
- **Frequenza Cardiaca** (*heart rate*, HR): parametro critico per la classe dei BBs; Il riscontro di bradicardia (FC<50 b.p.m.) porta a suggerire il parere di un esperto.
- **Potassiemia (K+)**: Concentrazione di potassio presente nel sangue; principale marcatore di sicurezza per gli MRA e i RAASi (ARNI/ACE-i/ARB). Il

superamento di determinati valori soglia può consigliare il parere di uno specialista e/o portare ad una diminuzione o l'interruzione della terapia.

- **Tasso di Filtrazione Glomerulare Stimato** (*estimated Glomerular Filtration Rate*, eGFR): Costituisce la metrica primaria di funzione renale utilizzata dalle linee guida per definire la sicurezza di SGLT2i, MRA e ARNI/ACE-i/ARB.
- **Creatininemia**: Parametro di laboratorio utilizzato per stimare la GFR (eGFR) e utilizzato in concomitanza di tale indice per identificare episodi di danno renale acuto o incrementi repentini dei valori rispetto al baseline (>50%), che giustificano la riconsiderazione temporanea dei dosaggi.

Un'ulteriore questione centrale nella modellazione della terapia farmacologica dell'HFrEF riguarda la gerarchia prescrittiva all'interno della classe degli inibitori del sistema renina-angiotensina-aldosterone (ARNI, ACE-i e ARB). Le linee guida infatti presentano la classe ARB come una seconda scelta rispetto ad ACE-i, preferibile solo in caso di intolleranza a quest'ultima. ARNI invece viene consigliata nel caso in cui un paziente trattato con ACE-i rimanga sintomatico; tuttavia, allo stesso tempo viene specificato che il trattamento con ARNI è preferibile all'ACE-i come terapia di prima linea. Tale ambiguità testuale, in base alle sole linee guida, lascia dunque parzialmente irrisolto l'ordine di preferenza tra ACE-i e ARNI, che costituisce uno degli argomenti trattati e risolti successivamente nella fase di affinamento del modello del Sottocapitolo 3.3.

3.2 Definizione e formalizzazione del workflow

Le linee guida condensano la conoscenza scientifica. Strutturate attorno a evidenze rigorose, esse orientano l'operatore sanitario lungo l'intero percorso assistenziale, identificando nella GDMT il cardine del trattamento farmacologico per le patologie croniche al fine di ottimizzare i risultati clinici e preservare la qualità di vita del paziente. La narrazione ostacola l'applicazione reale. L'estensione dei volumi, unita alla dispersione dei contenuti e ai ritmi stringenti della pratica clinica, rende la consultazione estemporanea un processo critico, favorendo l'insorgere dell'inerzia terapeutica di fronte al continuo susseguirsi di aggiornamenti della letteratura.

L'informatica biomedica propone una svolta strutturale. Il paradigma delle *Computer-Interpretable Guidelines* [1] nasce precisamente per superare le barriere del testo libero, prefiggendosi la conversione di raccomandazioni qualitative e debolmente formalizzate in sequenze deterministiche di condizioni ed azioni esplicite, direttamente elaborabili dai sistemi informatici.
A tal proposito, la formalizzazione tramite diagrammi di flusso operativi costituisce uno degli approcci più efficaci. Un diagramma di flusso clinico schematizza il percorso assistenziale come una rete di nodi e archi direzionati. Ogni nodo può rappresentare uno snodo logico del processo (nodo decisionale), che determina la direzione del percorso terapeutico in base ai parametri del paziente, oppure un'azione da compiere sulla terapia, come le prescrizioni farmacologiche o le titolazioni dei dosaggi. Gli archi direzionali invece definiscono la sequenza temporale e/o causale del flusso.
La trasposizione delle raccomandazioni in un diagramma di flusso permette di tradurre regole cliniche complesse in una serie di proposizioni condizionali di tipo *if-then*. Questo livello di astrazione si adatta particolarmente bene a protocolli farmacologici ben definiti, come nel caso della quadruplice terapia, consentendo di modellizzare in modo più trasparente gli snodi in cui la condotta del medico deve discostarsi dallo standard teorico, come, per esempio, nel caso di intolleranze o controindicazioni.

Tuttavia, nel presente lavoro di tesi la formalizzazione in questa forma non viene impiegata per calcolare direttamente gli score di aderenza. Essa viene invece sfruttata come architettura concettuale di riferimento, che pone le basi per la fase di estrazione semantica: attraverso la mappatura logica delle raccomandazioni, il workflow costruito permette di isolare le variabili di interesse, aiutando a definire lo schema dei dati che dovrà essere estratto dalle lettere cliniche tramite LLM.

Tradizionalmente, prima delle ultime versioni delle linee guida ESC 2021/2023, la gestione farmacologica dell'HF si basava su un modello sequenziale a gradini, con il quale bisognava procedere all'introduzione e alla successiva titolazione di un farmaco prima di aggiungerne un altro alla terapia. Rispetto a questa impostazione, le raccomandazioni più recenti, con l'introduzione della quadruplice terapia, hanno imposto un cambio di paradigma. I quattro pilastri farmacologici (ARNI/ACE-i/ARB,

Beta-bloccanti, MRA e SGLT2i) infatti devono essere somministrati simultaneamente e gestiti in parallelo, al fine di massimizzare i benefici del paziente.

In termini di ingegneria del software, questa premessa preclude la possibilità di tradurre questo cambio di paradigma in un unico diagramma di flusso sequenziale. Questo tipo di workflow, di fronte alla necessità di valutare in un'unica sequenza tutte le combinazioni di parametri emodinamici, controindicazioni e stato dei dosaggi, porterebbe ad un'esplosione combinatoria di nodi decisionali. Optando per questo tipo di approccio, si otterrebbe un modello rigido, fragile di fronte ad eventuali dati mancanti e di complessa manutenzione.

Per far fronte a queste limitazioni, nel presente lavoro di tesi è stata progettata un'architettura modulare, con un canale parallelo per ognuna delle classi farmacologiche della quadruplice terapia. Questa scelta di modellazione determina alcuni vantaggi rispetto ad un diagramma di flusso unificato:

- **Indipendenza funzionale e tolleranza ai dati mancanti**: il flusso logico di una classe può essere completato indipendentemente da criticità o dati mancanti legati ad altre classi.
- **Scalabilità e manutenibilità del modello**: un eventuale aggiornamento o modifica del modello non richiede l'aggiornamento di tutti i moduli, ma solo di quello in cui è stata effettuata tale operazione.
- **Adesione al modello clinico e spiegabilità**: la struttura a canali paralleli ricalca la valutazione per singola classe effettuata dal cardiologo nel contesto assistenziale, così come la struttura delle linee guida. Ciò consente di ottenere un output altamente interpretabile, in grado di motivare la terapia prescritta per ogni farmaco senza interferire con l'ottimizzazione degli altri pilastri terapeutici effettuata in parallelo.

Al fine di garantire la massima usabilità del workflow nel contesto assistenziale, indipendentemente dalla fase clinica del paziente (primo riscontro della malattia, fase di titolazione o *follow-up* cronico), i diagrammi di flusso di tutte le classi

farmacologiche condividono la stessa macrostruttura logica. In particolare, tutti i percorsi sono strutturati in quattro fasi sequenziali:

1. **Verifica delle controindicazioni assolute**: Rappresenta il primo nodo di controllo per ogni canale. Viene verificata la presenza di controindicazioni assolute all'inizio della terapia (per esempio allergie note, eccedenza di parametri emodinamici o patologie specifiche), secondo quanto riportato nelle *Contraindications* della sezione *IN WHOM AND WHEN?* nella tabella supplementare di ogni farmaco nella linea guida, come riportato nel Sottocapitolo 3.1. In caso di riscontro positivo, il flusso operativo si conclude immediatamente, traducendosi nella sospensione immediata o nel mancato avvio della terapia.
2. **Discriminazione dello stato terapeutico iniziale**: In assenza di controindicazioni assolute, viene valutato lo stato di assunzione del farmaco. In particolare, si distinguono i pazienti già in trattamento (fase di titolazione/mantenimento) da quelli *naive* per la classe considerata (fase di potenziale avvio).
3. **Gestione delle Criticità Cliniche e Assessment Specialistico**:
    - Per i pazienti già in terapia: il workflow valuta la presenza di specifici stati patologici (per esempio ipotensione o iperkaliemia) o di alterazione dei parametri emodinamici, in base al contenuto della sezione *PROBLEM SOLVING* (e in parte *HOW TO USE?*) delle tabelle supplementari, descritte nel Sottocapitolo 3.1. L'esito determina se la terapia possa continuare oppure debba essere ridotta o sospesa.
    - Per i pazienti non ancora in terapia: l'algoritmo valuta la necessità di richiedere il parere di uno specialista sullo stato del paziente, come indicato nella sottosezione *Cautions/seek specialist advice* della sezione *IN WHOM AND WHEN?* delle tabelle supplementari (Sottocapitolo 3.1).
4. **Determinazione dell'azione terapeutica**:
    - In caso di terapia già avviata: in assenza di criticità, viene consigliata la prosecuzione della titolazione verso la dose target. Qualora emergano problematiche, l'azione varia in modo diverso e specifico in base al

problema (stasi della dose attuale, de-escalation o sospensione).

- In caso di terapia non avviata: Nel caso in cui non sia richiesta una valutazione specialistica e non sussistano ostacoli clinici, il sistema raccomanda l'avvio della terapia a partire dalla dose iniziale definita dalle linee guida (riportata nella Tabella 1 del Capitolo 3).

Al termine di ogni diagramma di flusso tutti i rami convergono verso un unico nodo di chiusura, denominato «Continue». Questo nodo formalizza la natura ciclica e dinamica della gestione dell'HFrEF: la valutazione dell'aderenza non è un evento statico, ma un processo ricorsivo che si riavvia dall'inizio a ogni successiva visita di controllo.

Sebbene la classe SGLT2i non richieda una fase di titolazione graduale e venga introdotta nella terapia ad una dose target fissa, il relativo workflow è stato realizzato seguendo le quattro fasi descritte precedentemente. Nonostante la fase di titolazione risulti superflua per questa classe, questa struttura garantisce, oltre ad un'architettura omogenea di gestione dei pilastri farmacologici, la possibilità di integrare facilmente aggiornamenti futuri delle linee guida. A conferma di questo approccio metodologico unificato, la Tabella 1 del Capitolo 3 riporta esplicitamente sia la dose di partenza sia la dose target per la classe degli SGLT2i, nonostante esse coincidano.

A monte del flusso di gestione dei vasodilatatori RAASi è necessario un ulteriore blocco decisionale, finalizzato a scegliere il farmaco da somministrare alternativamente tra le tre opzioni (ARNI/ACE-i/ARB). In particolare, viene applicata la seguente sequenza di priorità deterministica, mostrata in Figura 1:

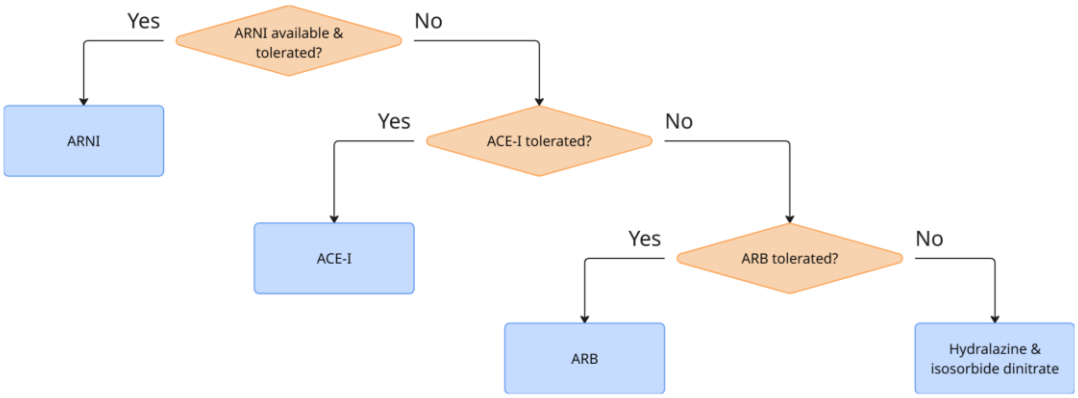


Figura 1. *Sistema gerarchico di selezione dei farmaci RAASi.*

Viene valutata prioritariamente la disponibilità e la tollerabilità legate alla classe ARNI. Come anticipato nel Capitolo 3.1 e approfondito nella fase di rifinitura (Sottocapitolo 3.3), la precedenza di ARNI rispetto ad ACE-i è stata formalizzata per allineare il workflow alle più recenti indicazioni della pratica clinica che ne promuovono l'impiego come trattamento di prima linea. Qualora sussistano controindicazioni o vincoli all'uso di ARNI (per esempio, disponibilità o costi), la scelta ricade sulla classe degli ACE-i. Infine, la classe ARB viene somministrata come ultima alternativa nel caso di un'intolleranza documentata agli ACE-i.

Di seguito vengono presentati in modo più dettagliato i workflow costruiti per modellizzare la gestione della terapia secondo la quadruplice terapia, evidenziando per ciascuna classe farmacologica la logica di controllo e i parametri fisiologici di riferimento:

- **Canale ARNI**: illustrato in Figura 2, applica il pattern a quattro stadi presentato precedentemente valutando la tollerabilità emodinamica e renale. Dopo aver valutato le controindicazioni, l'algoritmo procede in base ai valori assunti dai parametri emodinamici di K+ e eGFR. In assenza di criticità, guida la progressione dalla dose di partenza alla dose target; in caso contrario, instrada la decisione verso la sospensione della terapia e l'eventuale sostituzione con ARB.

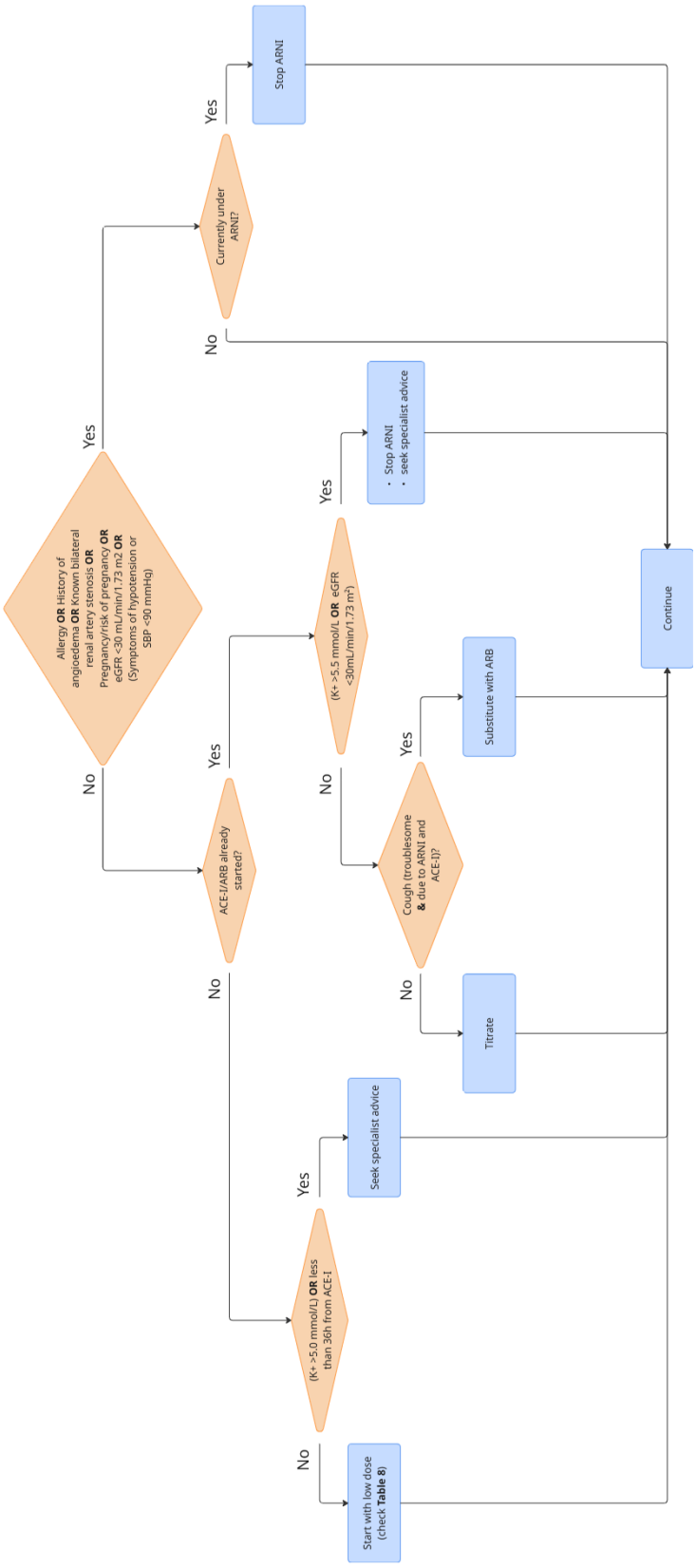


Figura 2. *Diagramma di flusso per la gestione farmacologica e la titolazione della classe ARNI, modellato sulle linee guida ESC.*

- **Canale ACE-i/ARB**: Nonostante ACE-i e ARB costituiscano due categorie farmacologiche distinte, la dinamica della loro gestione clinica e i parametri di sicurezza stabiliti dalle linee guida ESC sono identici. Di conseguenza, le due classi sono state modellate all'interno di un unico diagramma di flusso (Figura 3). Tale diagramma, inoltre, ha sostanzialmente la stessa struttura del canale riservato alla classe ARNI; la differenza rispetto alla gestione di questa classe farmacologica consiste nell'entità e nei valori soglia dei parametri emodinamici utilizzati nei blocchi decisionali.

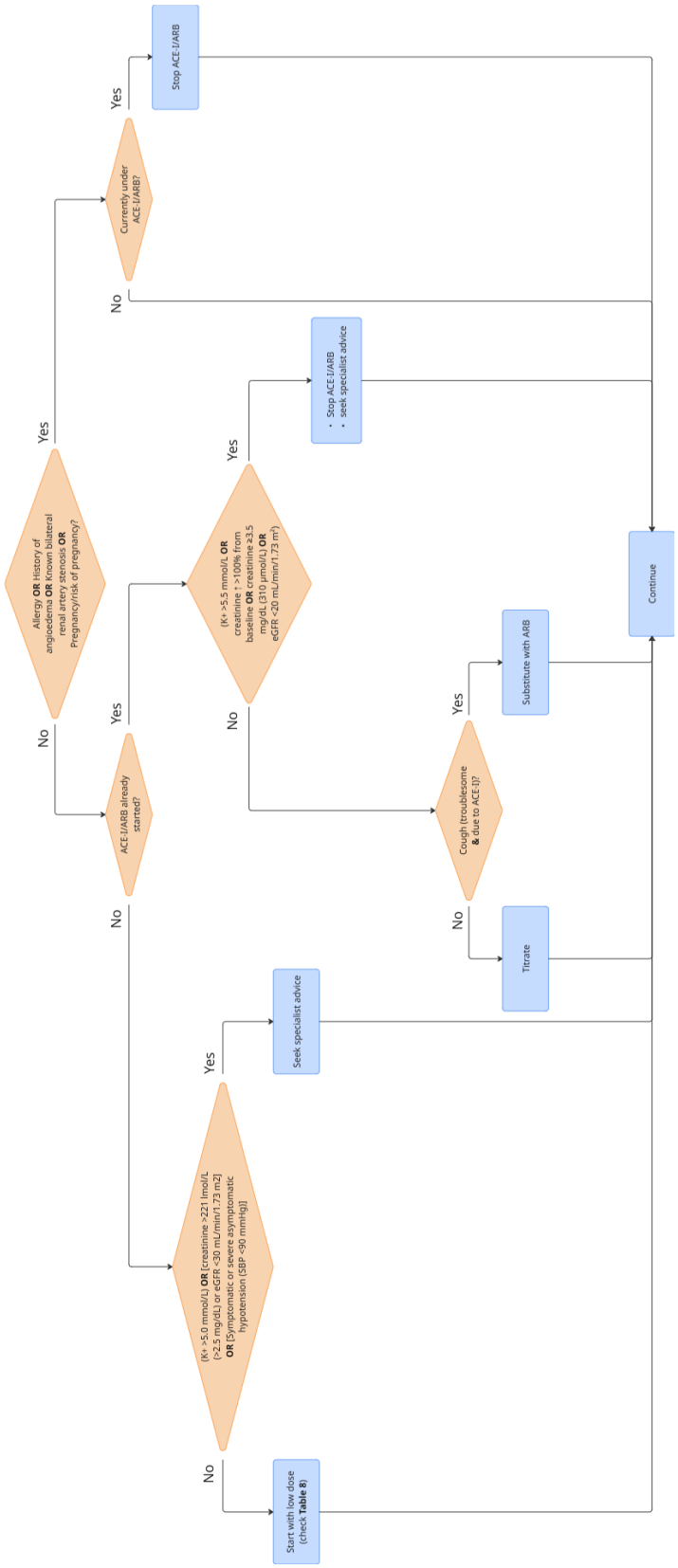


Figura 3. *Diagramma di flusso per la gestione farmacologica e la titolazione della classe ACE-i/ARB, modellato sulle linee guida ESC.*

- **Canale BBs**: Il workflow dedicato ai BBs, mostrato in Figura 4, risulta senza dubbio il più complesso, riflettendo la complessità della sezione di *PROBLEM SOLVING* delle linee guida ESC. Oltre al solito blocco decisionale delle precauzioni, infatti, seguono numerosi parametri e condizioni da verificare, tra cui la frequenza cardiaca, lo stato di deterioramento e la pressione sanguigna. In base alle condizioni del paziente può essere consigliato di dimezzare la terapia, sospenderla, oppure, in assenza di complicanze, di procedere alla titolazione. L'ostacolo principale legato alla modellazione della gestione dei BBs è stato individuare il giusto ordine gerarchico di gestione delle problematiche, che nelle linee guida non veniva esplicitato. A questo proposito, è stato fondamentale l'intervento dell'esperto nella fase di rifinitura del modello.

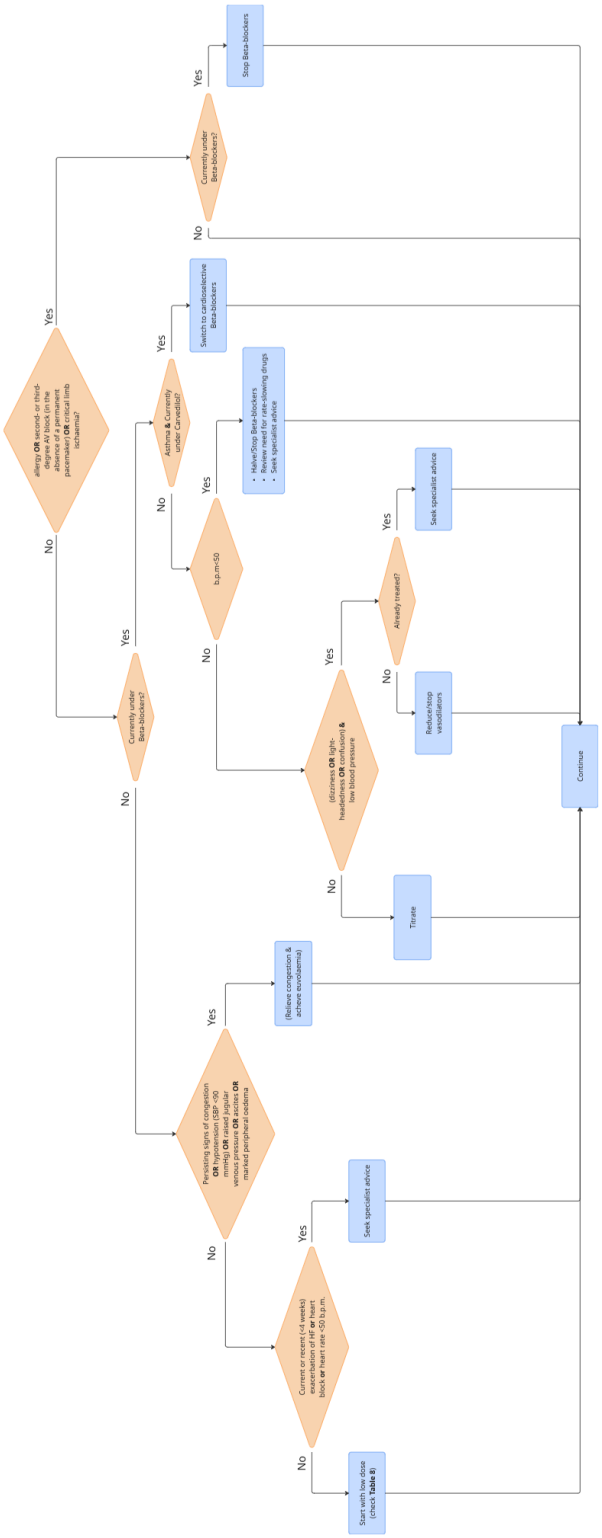


Figura 4. *Diagramma di flusso per la gestione farmacologica e la titolazione della classe BBs, modellato sulle linee guida ESC.*

Canale MRA: questo workflow prevede la verifica della presenza di allergie note e successivamente indirizza la terapia sulla base di specifiche soglie legate ai tre parametri emodinamici K+, eGFR e creatininemia. In base ad essi, l'algoritmo stabilisce se ridurre la dose, sospendere la somministrazione o proseguire con l'ottimizzazione della terapia. Tale diagramma di flusso è mostrato in Figura 5.

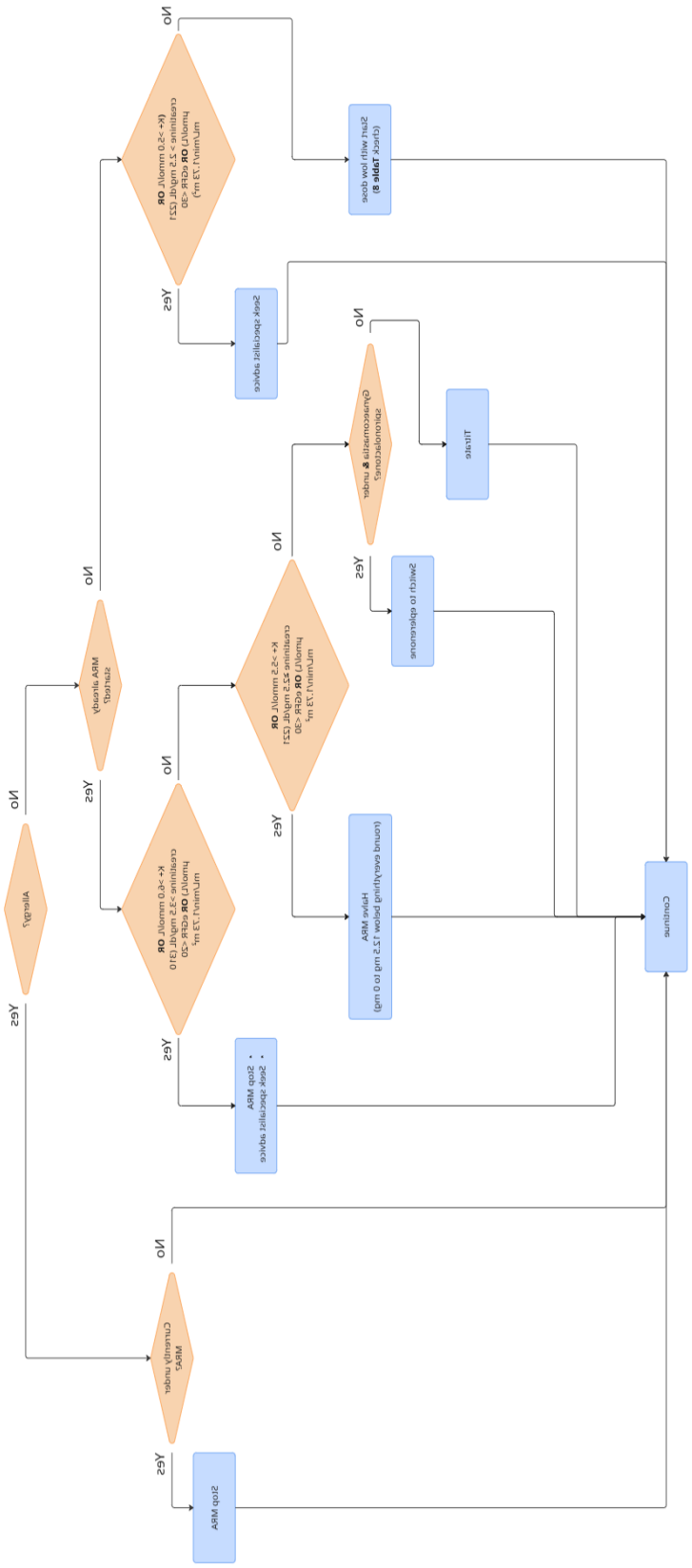


Figura 5. *Diagramma di flusso per la gestione farmacologica e la titolazione della classe MRA, modellato sulle linee guida ESC.*

- **Canale SGLT2i**: In conformità con il pattern a quattro stadi definito precedentemente, il workflow per gli SGLT2i mantiene la stessa architettura, pur gestendo una terapia farmacologica a dose fissa. Il flusso logico di questa classe risulta il più elementare e prevede solamente le controindicazioni come unico blocco decisionale (Figura 6). In assenza di tali ostacoli clinici, il flusso raccomanda direttamente l'avvio o il mantenimento della terapia alla dose fissa target prescritto dalle linee guida, senza fasi intermedie di titolazione.

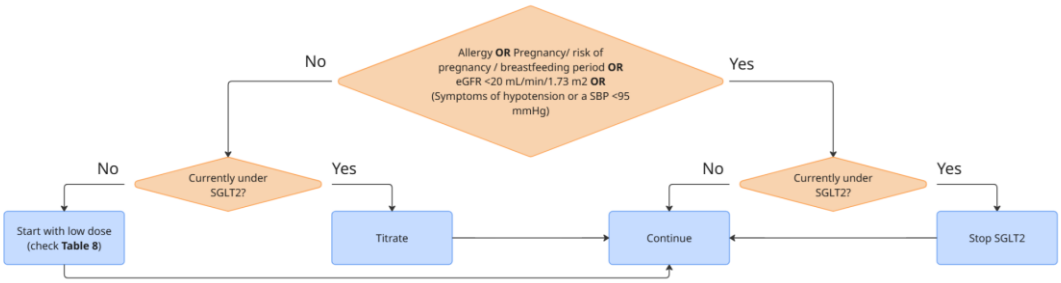


Figura 6. *Diagramma di flusso per la gestione farmacologica e la titolazione della classe SGLT2, modellato sulle linee guida ESC.*

3.3 Consulenza clinica e perfezionamento del modello

La traduzione puramente letterale di una linea guida in un algoritmo deterministico presenta delle criticità strutturali. Queste, infatti, tendono a descrivere uno scenario ideale di trattamento, e spesso non riescono a modellizzare completamente la variabilità reale.

Pertanto, risulta fondamentale perfezionare il workflow costruito a partire dalle linee guida tramite un approccio basato sul coinvolgimento iterativo degli esperti (*Expert-in-the-Loop*), nel quale il diagramma di flusso teorico viene sottoposto a cicli iterativi di revisione e affinamento in collaborazione diretta con specialisti del settore (nello specifico, medici cardiologi esperti nello scompenso cardiaco).

A tal fine, la prima versione dei diagrammi di flusso è stata sottoposta ad un processo iterativo di rifinitura con l'aiuto del Dr. Tobias Röschl, medico specialista in

cardiologia dell'Ospedale Charité – Universitätsmedizin Berlin, Berlino, Germania (ospedale partner nel progetto GUIDE-AI). Questo processo ha permesso di adattare gli step del percorso terapeutico alla prassi consolidata nelle strutture sanitarie e di risolvere le ambiguità presenti nel testo, definendo, per esempio, soglie operative univoche per raccomandazioni che nella linea guida rimangono aperte all'interpretazione.

In particolare, il coinvolgimento iterativo degli esperti ha consentito di risolvere le seguenti criticità emerse durante la costruzione dei workflow:

- **Formalizzazione della gerarchia delle problematiche**: Nella pratica clinica in contesto assistenziale, l'eventuale insorgenza di più complicanze prevede la gestione della criticità preminente e il rinvio delle ulteriori ottimizzazioni ai successivi controlli di *follow-up*, nei quali il workflow verrà ripercorso dall'inizio. A tale scopo, è di fondamentale importanza definire un ordine gerarchico di gestione, che, nella maggior parte dei casi, nelle linee guida non viene definito esplicitamente.
- **Conflitti logici**: La natura qualitativa che spesso assumono le parti più discorsive delle linee guida può generare ambiguità concettuali, che, nel contesto di un workflow decisionale, si traducono in blocchi condizionali sovrapposti o in raccomandazioni tra loro incompatibili.
- **Elementi superflui**: Le linee guida contengono frequentemente ridondanze, casistiche marginali descritte in modo approssimativo e qualitativo che non vengono considerate nella routine ambulatoriale o indicazioni estese a molecole accessorie non facenti parte dei farmaci appartenenti alla quadruplice terapia.

In particolare, i pilastri farmacologici per i quali il processo di rifinitura si è reso maggiormente necessario sono stati gli inibitori del sistema RAAS (ARNI/ACE-i/ARB) e i BBs.

La criticità principale legata ai RAASi consiste in un conflitto logico nella sezione *PROBLEM SOLVING* delle relative tabelle supplementari: sia per ARNI sia per ACE-

i/ARB viene presentata la stessa strategia farmacologica incentrata su specifici valori soglia dei parametri emodinamici. Se tali parametri eccedono un primo valore critico, il flusso può portare verso il dimezzamento della dose; se invece viene oltrepassato un secondo valore soglia più alto del primo, la terapia viene sospesa. L'elemento di conflitto si presenta nel caso specifico della classe ARNI per la quale tali valori sembrano coincidere. In particolare, per questa classe vengono definite le seguenti soglie di tollerabilità:

- **Valori accettabili di norma**: eGFR ≥ 30 mL/min/1.73m^2 e K+ ≤ 5.5 mmol/L.
- **Condizione di dimezzamento della terapia**: Incrementi della Creatinina o della Potassiemia superiori ai limiti sopra descritti.
- **Condizione di sospensione del farmaco**: eGFR < 30 mL/min/1.73m^2 o K+ > 5.5 mmol/L

Questa formulazione genera un'incompatibilità condizionale. Le due proposizioni logiche condividono l'esatto confine di attivazione: superata la soglia critica (K+ > 5.5 o eGFR < 30), il sistema si trova di fronte a due istruzioni mutuamente esclusive e attivate dalla medesima condizione. Nel contesto di un diagramma di flusso deterministico, questa sovrapposizione impedisce di stabilire in modo univoco quale debba essere la traiettoria da seguire.

Il consulto con l'esperto di dominio ha portato alla risoluzione di questo stallo concettuale allineando il modello alla pratica clinica. In particolare, nella routine ambulatoriale, l'insorgenza di un'iperkaliemia severa (K+ > 5.5 mmol/L) o di un decadimento marcato della funzione renale (eGFR < 30 mL/min/1.73m^2) costituisce un segnale di allarme prioritario che non lascia spazio ad aggiustamenti intermedi incerti, ma impone la tempestiva sospensione della terapia RAASi e il consulto specialistico. Dunque, grazie a questa fase di rifinitura, entrambi i canali ARNI e ACE-i/ARB hanno subito una semplificazione, basata sul superamento della sola soglia di sicurezza maggiormente critica.

Per quanto riguarda invece la classe dei BBs, la modellizzazione del canale ad essa associato è stata la più complessa dell'intero progetto, principalmente a causa dell'elevato numero di sintomi ed eventi avversi descritti nella sezione di *PROBLEM*

SOLVING delle tabelle supplementari ESC. Nel testo originale delle linee guida, diverse raccomandazioni fanno riferimento a descrittori qualitativi, tra cui condizioni di deterioramento indicate genericamente come «*severe*» o «*serious deterioration*», senza associare a tali definizioni valori numerici o parametri fisiologici oggettivi.
Anche in questo caso, in fase di rifinitura, è stata apportata una parziale semplificazione del workflow. Tali elementi sono infatti stati esclusi dal flusso logico, trattandosi di condizioni non definite in modo oggettivo.

La seconda criticità sostanziale risiedeva nell'assenza di un ordine di precedenza esplicito all'interno delle linee guida ESC. Nel testo di riferimento, le diverse problematiche vengono elencate senza specificare una gerarchia di trattamento. Tuttavia, nella pratica ambulatoriale in contesto assistenziale, tali complicanze possono presentarsi in modo concomitante, imponendo al medico una scelta prioritaria poiché una singola visita di controllo consente l'attuazione di un'unica modifica alla terapia in corso.
Grazie al processo iterativo di revisione da parte degli esperti, è stato possibile stabilire e convalidare il corretto ordine gerarchico di gestione delle problematiche, assicurando che il workflow risponda in modo univoco e riproducibile anche di fronte a quadri clinici articolati.

4. Estrazione dei dati clinici tramite LLM

4.1 Descrizione del dataset di lettere cliniche

L'architettura di Information Extraction (IE) poggia su un dataset clinico reale. I dati riflettono contesti assistenziali effettivi. Il corpus, comprendente lettere di dimissione e referti ambulatoriali, proviene dagli Istituti Clinici Scientifici Maugeri (ICS Maugeri) di Pavia, partner di progetto nell'iniziativa GUIDE-AI e centro di riferimento per la validazione sul campo delle tecnologie sviluppate.

La tutela dei dati è stata prioritaria. Prima di consentire l'accesso al materiale testuale, l'ente ospedaliero ha sottoposto la documentazione a un rigoroso protocollo di anonimizzazione per garantire la piena conformità con le normative vigenti sulla protezione della privacy. Ogni identificatore del paziente è stato rimosso. Nomi, codici fiscali, recapiti e riferimenti di cartella clinica sono stati sostituiti con stringhe alfanumeriche fittizie (quali «XYZ» o «PK-00000»), un'operazione che ha assicurato la massima protezione delle informazioni personali senza compromettere la struttura sintattica originale, la cui integrità risulta essenziale per l'efficacia degli algoritmi di estrazione.

Per quanto riguarda l'acquisizione del dataset, quest'ultima si è articolata in due fasi temporali distinte, per un totale di 20 lettere cliniche. In una prima fase è stato utilizzato un campione ridotto di cinque lettere cliniche, che ha svolto una funzione essenziale per la comprensione della struttura sintattica dei report, la mappatura delle varianti terminologiche e la prima fase di ingegnerizzazione dei prompt. Successivamente, il campione è stato integrato con ulteriori 15 lettere, raggiungendo il numero complessivo di 20 documenti. Questa seconda fase ha permesso di valutare le performance del sistema di estrazione di fronte ad una maggiore variabilità nell'impostazione stilistica adottata dal medico e di valutare quantitativamente gli LLM utilizzati nel processo di IE (come descritto nel Sottocapitolo 6.1).

I report ambulatoriali di cui è composto il dataset hanno un formato semi-strutturato, caratterizzato dalla compresenza di parti in testo libero, elenchi sintetici e tabelle. La

struttura dei documenti è costruita secondo l'ordine cronologico degli eventi, dai meno ai più recenti. In particolare, le sezioni di interesse per la fase di estrazione sono le seguenti:

- **Anagrafica**: Costituisce la parte introduttiva del documento e vi sono contenute le informazioni anagrafiche del paziente, dalle quali si possono derivare possibili fattori di rischio, tra cui il sesso, la provenienza o l'età.
- **Anamnesi patologica remota e prossima**: Questa sezione racchiude l'evoluzione temporale della patologia sotto forma di testo libero. Vi sono contenute informazioni longitudinali, tra le quali il trend storico della frazione di eiezione o gli esami di laboratorio pregressi. La natura non strutturata di questa sezione richiede la capacità, da parte degli LLM, di estrarre correttamente i dati aggiornati più recenti.
- **Esame Obiettivo ed Esami Strumentali**: Sezioni contenenti i parametri fisiologici immediati rilevati durante la visita. La forma sintetica e abbreviata delle variabili misurate (ad esempio, PA corrisponde a pressione arteriosa), impone al sistema di estrazione la capacità di interpretare correttamente gli acronimi e le abbreviazioni.
- **Conclusioni diagnostiche**: Sintetizza il fenotipo osservato nel paziente, presentando, oltre alla diagnosi di HF, eventuali comorbidità o condizioni che potrebbero influire sulla gestione della quadruplice terapia (ad esempio, la cardiomiopatia).
- **Tabelle terapeutiche**: Contengono le informazioni relative alle prescrizioni farmacologiche divise in due tabelle che mappano rispettivamente lo stato della terapia all'ingresso e all'uscita dalla visita.

Le tabelle dedicate alla terapia farmacologica rappresentano la fonte di informazione primaria per il calcolo degli score di aderenza del Capitolo 5. La suddivisione dei dati farmacologici si articola su due tabelle distinte: la prima traccia l'anamnesi prescrittiva antecedente alla visita, mentre la seconda formalizza la terapia da proseguire o avviare al termine del riscontro clinico. L'architettura informativa rimane identica. Entrambi i prospetti condividono infatti una struttura speculare che registra, per ogni singola voce, il nome commerciale, il principio attivo, la via di somministrazione e il dosaggio

dell'agente terapeutico. I farmaci introdotti ex novo nella terapia vengono distinti da un apposito marcatore di stato (ad esempio, «NUOVO»).

L'analisi automatizzata di queste sezioni presenta delle complessità strutturali che rendono inefficaci i parser di testo basati su espressioni regolari (*regular expression*, regex), giustificando l'utilizzo di LLM. Tra queste criticità c'è la necessità di associare univocamente ad ogni farmaco la corretta classe della quadruplice terapia (ARNI, BBs, MRA, SGLT2i) a partire dalle informazioni contenute nel nome commerciale e nel principio attivo. A tale complessità si aggiunge l'utilizzo sistematico di abbreviazioni ed espressioni contratte nella definizione del dosaggio, la cui corretta interpretazione da parte del modello è fondamentale per la successiva fase di calcolo automatico degli score.
In sintesi, la corretta conversione di tali informazioni in un formato strutturato costituisce il presupposto per la fase successiva del presente lavoro di tesi.

4.2 Progettazione dello Schema JSON

Il testo semi-strutturato delle lettere ambulatoriali è stato convertito in un oggetto JSON con una struttura tale da contenere tutte le informazioni di interesse per il trattamento dell'HFrEF formalizzate nei workflow descritti nel Capitolo 3.
Parallelamente alla costruzione di questo schema, è stato predisposto e compilato un form di annotazione da parte dei medici cardiologi dell'ICS Maugeri, al fine di confrontare le informazioni estratte automaticamente dagli LLM con l'annotazione di riferimento ottenuta dai moduli compilati per ciascuna delle 20 lettere cliniche raccolte. Tale form è riportato integralmente nella Tabella A.1 dell'Appendice A.

A tale scopo, la progettazione di tale schema dati è stata guidata dalla necessità di rispecchiare fedelmente la struttura del form. In particolare, lo schema è stato articolato riproducendo le 8 macro-sezioni del form di riferimento:

- **Inquadramento demografico (Sezione A):** Estrazione delle variabili anagrafiche essenziali (sesso ed età).
- **Intolleranze farmacologiche (Sezione B):** Elenco strutturato dei principi attivi e delle sostanze verso cui il paziente presenta controindicazioni

documentate.

- **Esami di laboratorio (Sezione C):** Rappresentazione mediante un array di oggetti contenente il nome dell'esame, il valore numerico, l'unità di misura e la data di rilevazione, garantendo la tracciabilità dei principali parametri clinici (K^+, eGFR, creatinina, pressione arteriosa sistolica [PAS], frequenza cardiaca [FC], LVEF e QRS)
- **Comorbilità e condizioni cliniche (Sezioni D ed E):** Tracciamento delle patologie concomitanti e dei sintomi acuti registrati nell'anamnesi (es. dispnea, fibrillazione atriale).
- **Criteri fenotipici e flag di sicurezza per l'HFrEF (Sezione F):** Mappatura booleana di indicatori clinici specifici per lo scompenso cardiaco (es. riacutizzazione recente, uso di diuretici dell'ansa, presenza di edemi periferici, eziologia ischemica/non ischemica).
- **Terapia farmacologica corrente e alla dimissione (Sezioni G e H):** Elenco dettagliato dei farmaci prescritti (nome commerciale, principio attivo, dose giornaliera e classe farmacologica), distinguendo lo stato prescrittivo precedente alla visita da quello concordato alla fine di tale visita.

Lo schema JSON è stato implementato in ambiente Python (versione 3.14.5) sfruttando la libreria Pydantic. L'uso di questa libreria consente di imporre vincoli di tipizzazione specifici sui dati estratti, mediante tipi scalari (*float*, *int*, *bool*) e enumerazioni (*Enum*) per le classi farmacologiche e vettori di oggetti.
Di seguito, in tabella 2, viene riportato un estratto sintetico della formalizzazione dello schema target in ambiente Python:

Campo	Tipo di Dato	Descrizione
demographic_info	*List[str]*	Informazioni demografiche (età, sesso, etnia, ecc.).
drug_intolerances	*List[str]*	Farmaci o classi non tollerati dal paziente.

Campo	Tipo di Dato	Descrizione
lab_values	*List[LabEntry]*	Misurazioni di laboratorio e parametri vitali più recenti.
comorbidities	*List[str]*	Diagnosi concomitanti afferenti all'HFrEF.
disorders_and_conditions	*List[str]*	Diagnosi, sintomi o condizioni patologiche non classificabili come comorbilità.
worsening_heart_failure	*bool*	Vero se il paziente ha avuto un ricovero o una visita per HF negli ultimi 6 mesi.
loop_diuretic	*bool*	Vero se il paziente assumeva un diuretico dell'ansa prima della visita.
peripheral_edema	*bool*	Vero se è documentata la presenza di edemi periferici.
hfrEF_90_days	*bool*	Vero se la diagnosi di HFrEF risale ad almeno 3 mesi prima.
comfort_care_only	*bool*	Vero in presenza di sole cure palliative.
hfrEF_etiology	*Enum*	Eziologia dell'HFrEF: *ISCHEMIC* oppure *NON_ISCHEMIC*.
discharge_date	*str*	Data dell'ultima dimissione per HF.
current_therapy	*List[TherapyEntry]*	Elenco dei farmaci assunti prima della visita.
discharge_therapy	*List[TherapyEntry]*	Regola terapeutica completa prescritta alla dimissione.

Tabella 2. *Schema JSON target: campi di primo livello della classe HFrEFPatientSchema, con tipo di dato e descrizione (estratto sintetico).*

Ad ogni elemento della lista *lab_values* vengono associate le seguenti informazioni, riportate in Tabella 3:

Campo	Tipo di Dato	Descrizione e Regole di Estrazione
lab_exam	*str*	Nome dell'esame di laboratorio o del parametro clinico (es. Creatinina, Potassio, FC, LVEF).
lab_value	*str*	Valore numerico esatto associato al risultato.
lab_unit	*str*	Unità di misura della variabile (es. mg/dL, mEq/L, %)
lab_date	*str*	Data della misurazione (formato gg.mm.aaaa).
notes	*str*	Informazioni contestuali aggiuntive sul parametro estratto.

Tabella 3. *Struttura della classe LabEntry: campi associati a ciascun elemento della lista lab_values.*

Per ciascun farmaco individuato nel documento, ovvero per ogni elemento delle liste *current_therapy* e *discharge_therapy*, vengono riportate le seguenti informazioni (Tabella 4):

Campo	Tipo di Dato	Descrizione e Regole di Estrazione
drug_name	*str*	Nome esatto del farmaco come riportato nella lettera clinica.
generic_drug_name	*str*	Denominazione comune internazionale (DCI) o codice ATC tradotto in inglese.
brand_drug_name	*str*	Nome commerciale del farmaco.
active_ingredient	*str*	Dosaggio del principio attivo e unità di misura (es. 20 mg o 97/103 mg).

Campo	Tipo di Dato	Descrizione e Regole di Estrazione
dose_units_number	*float*	Conteggio totale delle unità fisiche (es. compresse) assunte nell'arco delle 24 ore.
drug_class	*Enum*	Mappatura automatica della classe terapeutica: *ACE_I*, *ARB*, *ARNI*, *BETA_BLOCKER*, *MRA*, *SGLT2_INHIBITOR*, *LOOP_DIURETIC*.
notes	*str*	Note su titolazioni, modifiche di dose, date di inizio/variazione dello schema posologico.

Tabella 4. *Struttura della classe TherapyEntry: campi associati a ciascun farmaco delle liste current_therapy e discharge_therapy.*

Dal confronto diretto tra il form compilato dagli specialisti e lo schema JSON target del processo di estrazione si può osservare come essi condividano una struttura comune a eccezione di un campo presente nel form di annotazione ma assente nello schema Pydantic. Si tratta del campo «Critically ill - HFrEF», un valore booleano che indica l'ospedalizzazione per shock cardiogeno. Tale informazione, non essendo utile né al calcolo degli score di aderenza né alla gestione della quadruplice terapia in generale, è stata esclusa dallo schema di estrazione finale per evitare di appesantire inutilmente il processo di estrazione.

In conclusione, la struttura dati così formalizzata assicura un flusso di informazioni omogeneo e deterministico, capace di registrare le informazioni utili alla gestione farmacologica dell'HFrEF e direttamente interpretabile dall'algoritmo di calcolo automatico degli score di aderenza (Capitolo 5).

4.3 Architettura dell'estrattore basato su LLM

L'estrazione strutturata delle informazioni dalle lettere cliniche richiede un controllo rigoroso sul formato dei dati generati dagli LLM nel processo di IE. Come anticipato nel Sottocapitolo 4.2, per garantire la conformità degli output allo schema predefinito, è stata utilizzata la libreria Python Pydantic.
In particolare, la struttura adottata prende forma sotto la classe radice

HFrEFPatientSchema, la quale aggrega sottosezioni specializzate, tra cui, per esempio, *TherapyEntry* per la modellazione delle prescrizioni e *LabEntry* per i parametri ematochimici e strumentali. Attraverso la funzione *model_json_schema()*, la definizione Pydantic viene convertita in uno schema JSON standardizzato, in modo tale da garantire la conformità sintattica degli oggetti generati.

Le istruzioni impartite al modello si articolano su due livelli complementari, in modo tale da minimizzare le ambiguità interpretative del linguaggio clinico:

- **Prompt di sistema**: Definisce il contesto e le direttive generali. Definisce il ruolo del modello come estrattore dati specializzato in ambito cardiologico, impone il divieto assoluto di inferenza per dati non esplicitati (assegnando il valore *null*) e fissa le regole di formattazione dell'output (Sottocapitolo B.1 dell'Appendice B).
- **Descrizioni di campo**: Descrizioni specifiche di ogni variabile del modello. Sfruttando l'argomento *description* nativo di Pydantic, in queste descrizioni vengono specificate indicazioni specifiche per l'informazione da estrarre, per esempio sinonimi clinici, abbreviazioni ammesse o la forma con cui estrarre i dati (Sottocapitoli B.2-B.9 dell'appendice B).

Per mitigare ulteriormente il rischio di allucinazioni e garantire la verificabilità del dato estratto, è stato utilizzato il costrutto generico *Extracted[T]*. In questo modo, ogni campo all'interno dello schema non restituisce solamente le informazioni estratte, ma una struttura composta dal dato vero e proprio convertito nel tipo finale predefinito (*float*, *int*, *Enum* ecc.) e il relativo *source_text*, ovvero la porzione di testo originale dal quale è stata derivata l'informazione estratta.
Ciò permette una fase di validazione-post estrazione accurata, permettendo di verificare istantaneamente la veridicità delle estrazioni effettuate dal modello e determinando la necessità di effettuare modifiche ai prompt.

La pipeline supporta sia modelli ospitati localmente sia servizi cloud; nella sperimentazione finale tutti i modelli sono stati erogati tramite servizi cloud (Ollama Cloud e AWS Bedrock). La selezione dei modelli ha avuto come obiettivo principale il mantenimento della massima riproducibilità e il rispetto dei vincoli strutturali

definiti dallo schema JSON..

Nella fase preliminare di sperimentazione, i modelli eseguiti localmente sono stati ospitati presso il Laboratorio di Informatica Biomedica “Mario Stefanelli” dell’Università di Pavia. L'infrastruttura di inferenza è stata progettata con un’impostazione client-server distribuita su rete locale, in modo tale da disaccoppiare la pipeline di gestione Python dall'esecuzione dei modelli. In particolare, l'interfacciamento con gli LLM è stato realizzato tramite chiamate HTTP tramite API REST dirette (tramite la funzione *call_llm_port*) verso l'endpoint */api/chat* dei server Ollama. Per ottimizzare le risorse di calcolo e bilanciare il carico in base alla complessità dei modelli, sono stati impiegati nodi di rete dedicati su indirizzi IP distinti. I modelli ospitati localmente non sono stati però selezionati per la sperimentazione finale poiché non rispettavano i criteri di selezione descritti di seguito, in particolare la conformità allo schema JSON. La stessa interfaccia è stata mantenuta nella sperimentazione finale, in cui Gemma 4 31B e GPT-OSS 120B sono stati richiamati tramite Ollama Cloud: il modello da utilizzare è infatti selezionato dal parametro *model* della richiesta, per cui lo stesso codice consente di interrogare sia modelli ospitati localmente sia modelli cloud.

Oltre ai modelli serviti tramite Ollama, la pipeline di IE è stata collegata al servizio AWS Bedrock tramite l’SDK *boto3*, in modo tale da estendere il processo di estrazione a modelli avanzati erogati in modalità serverless (tra cui le ultime versioni delle famiglie Anthropic Claude, DeepSeek e Qwen).

Il confronto tra modelli eseguiti in locale e modelli erogati in cloud evidenzia vantaggi e svantaggi complementari. L’esecuzione in locale garantisce un maggiore controllo sulla versione del modello utilizzato, senza richiedere costi legati al numero di richieste; inoltre, evita che eventuali dati sensibili vengano trasferiti ad un fornitore terzo. Tuttavia, richiede hardware dedicato e limita, di norma, la dimensione dei modelli utilizzabili e quindi le prestazioni raggiungibili.
Al contrario, il cloud permette di accedere, senza infrastruttura propria, a modelli di grandi dimensioni e allo stato dell’arte; comporta però il trasferimento dei dati a un

fornitore terzo, con le relative implicazioni normative e contrattuali, costi proporzionali all'utilizzo e il rischio che le versioni dei modelli vengano aggiornate o dismesse, a scapito della riproducibilità. In questo studio il ricorso al cloud è stato possibile perché le lettere cliniche sono state anonimizzate e pseudonimizzate (Sottocapitolo 4.1); nell'impiego reale su dati non anonimizzati, la scelta dovrà tenere conto di questi vincoli.

La selezione dei modelli idonei alla sperimentazione finale è avvenuta sulla base di un processo di filtraggio basato sui seguenti parametri:

- **Conformità allo schema JSON**: Si tratta della caratteristica più importante da rispettare; senza una struttura conforme allo schema predefinito, è impossibile procedere alla fase di calcolo automatico degli score di aderenza. Molti dei modelli più semplici testati, ovvero con un numero di parametri ridotto (< 10B) hanno mostrato questo limite, venendo di conseguenza scartati.
- **Massima riproducibilità**: È stata premiata la capacità di fornire output simili tra loro nel corso di estrazioni ripetute, senza la presenza di artefatti o dati mancanti o discordanti tra un'estrazione e l'altra.
- **Correttezza dei dati estratti**: Oltre a fornire dati simili tra un'estrazione e l'altra, è importante che i modelli selezionati riescano ad estrarre le corrette informazioni contenute nelle lettere cliniche, evitando allucinazioni, dati mancanti o inferenze prive di basi solide nel testo di riferimento.

Sulla base di questi principi, la fase di sperimentazione finale si è focalizzata su cinque architetture rappresentative: due erogate tramite Ollama Cloud, ovvero Gemma 4 31B e GPT-OSS 120B, e tre eseguite tramite AWS Bedrock, ovvero Qwen 3 32B, Claude Sonnet 4.6 e DeepSeek v3.2. Le caratteristiche note principali per ogni modello sono sintetizzate di seguito in Tabella 5:

Modello	Ambiente di utilizzo	Numero di parametri	Applicazioni
Gemma 4 31B	Ollama (Ollama Cloud)	~31 miliardi	Elaborazione multimodale (testo e immagini) e integrazione con strumenti esterni.
GPT-OSS 120B	Ollama (Ollama Cloud)	~117–120 miliardi	Ragionamento logico-matematico complesso e gestione di flussi operativi strutturati.
Qwen 3 32B	AWS Bedrock	~32,8 miliardi	Conversazione diretta, ragionamento esteso e generazione di output formattati/JSON.
Claude Sonnet 4.6	AWS Bedrock	Non dichiarato	Elaborazione di documenti/progetti a contesto esteso (fino a 1M token), ragionamento adattivo e automazione.
DeepSeek v3.2	AWS Bedrock	Non specificato	Sviluppo di agenti autonomi, scrittura e analisi di codice e orchestrazione di strumenti esterni.

Tabella 5. *Caratteristiche principali degli LLM applicati alle lettere cliniche nella fase di estrazione.*

L'estrazione di informazioni in forma strutturata a partire da lettere cliniche di natura non strutturata, concisa e spesso ambigua presenta diverse criticità, che in questa fase del lavoro di tesi sono state affrontate sia tramite i prompt dati in input agli LLM sia tramite la struttura dello schema Pydantic. Le problematiche principali incontrate durante l'implementazione della struttura di IE sono le seguenti:

- **Disambiguazione tra comorbidità e condizioni cliniche**: I modelli tendono a confondere la condizione patologica principale (HF), le comorbidità e le altre condizioni cliniche, inserendole in modo indistinto in un'unica lista. Per evitare questo tipo di ambiguità, è stata imposta una netta separazione tra i concetti di comorbidità (*comorbidities*) e disordini e condizioni cliniche (*disorders_and_conditions*) dello schema JSON. Attraverso le descrizioni

Pydantic, è stato specificato di escludere tassativamente lo scompenso cardiaco da entrambi i campi, in quanto patologia target la cui diagnosi viene considerata a priori. Inoltre, per entrambe le proprietà è stato specificato di non includere patologie o condizioni già incluse in un altro campo, in modo tale da evitarne la sovrapposizione. Tuttavia, questa parte dello schema rimane tuttora la più problematica, dal momento che LLM diversi possono selezionare alternativamente come *comorbidities* o *disorders_and_conditions* diverse condizioni.

- **Inquadramento temporale delle terapie**: La distinzione tra la terapia corrente antecedente alla visita e quella aggiornata al termine della stessa è di fondamentale importanza per la fase di calcolo automatico degli score di aderenza, che hanno l'obiettivo di quantificare la qualità della terapia aggiornata. In una prima fase, i campi *current_therapy* e *discharge_therapy* venivano spesso mal interpretati dagli LLM, con il rischio di non distinguere correttamente le due terapie oppure di includere nella terapia aggiornata solamente i farmaci prescritti ex-novo. Per evitare questo inconveniente, è stato reso esplicito nel prompt di includere in *current_therapy* tutti e soli i farmaci prescritti antecedentemente alla visita, mentre è stato specificato di includere tutti i farmaci prescritti al termine della visita in *discharge_therapy*, sia quelli per cui la terapia deve essere proseguita da prescrizioni precedenti sia quelli di nuova prescrizione.
- **Calcolo della dose giornaliera**: Anche la corretta estrazione legata alla dose giornaliera risulta fondamentale per il calcolo esatto degli score di aderenza. L'ampia varietà nella sintassi legata alle prescrizioni farmaceutiche (ad esempio, «1 cpr ore 8-20», «1 cpr 3 volte al giorno», «1 cpr a giorni alterni») porta i modelli a compiere errori nel distinguere il frazionamento della singola unità fisica dal numero di somministrazioni giornaliere, generando errori nell'estrazione della quantità quotidiana complessiva. Per mitigare questo problema è stato fondamentale strutturare il prompt del campo *dose_units_number* con un approccio *few-shots*, fornendo all'LLM gli esempi necessari ad interpretare correttamente tutte le possibili formulazioni presenti nei testi clinici.

- **Classificazione farmacologica**: Un'ultima informazione di interesse per il calcolo degli score di aderenza è la classe di appartenenza dei farmaci inclusi nella terapia. Al fine di evitare che, a causa di numerosi nomi commerciali e principi attivi equivalenti, i modelli non riescano a ricondurre correttamente i farmaci alle classi terapeutiche raccomandate dalle linee guida ESC per la quadruplice terapia, è stata utilizzata una classe *Enum* Pydantic per vincolare i valori ammessi nel campo *drug_class* alle principali categorie: ARNI, ACE-i, ARB, Beta-bloccanti, MRA, SGLT2i, Diuretici e «Altro» (*Other* nel codice), che raccoglie i farmaci non appartenenti alla quadruplice terapia. In aggiunta, nella descrizione relativa a questo campo è stato ribadito di attenersi strettamente a tali categorie farmacologiche, in modo tale da minimizzare ulteriormente la possibilità di errore.

L'output generato dagli LLM nella di IE costituiscono la base di partenza per lo step successivo del lavoro di tesi, ovvero il calcolo degli score di aderenza, presentato nel capitolo successivo.

5. Calcolo automatico degli score di aderenza

5.1 Score di aderenza

L'aderenza terapeutica misura la conformità della terapia. Definita come il grado di allineamento tra la gestione clinica praticata dall'operatore e le evidenze scientifiche consolidate nelle linee guida di riferimento, essa persegue l'ottimizzazione dell'efficacia e della sicurezza nei percorsi di cura, promuovendo l'appropriatezza prescrittiva e l'uso razionale delle risorse. La quantificazione dell'aderenza mappa il divario tra la terapia prescritta e i regolamenti. L'analisi sistematica di tali scostamenti consente di identificare con precisione le discrepanze esistenti tra la prassi quotidiana e i protocolli teorici, fornendo informazioni essenziali per elevare la qualità dell'assistenza sanitaria.

L'HFrEF ridefinisce questo paradigma. Con l'affermazione della quadruplice terapia quale standard di cura [7], la valutazione dell'appropriatezza prescrittiva ha superato la semplice presenza formale dei farmaci nella terapia, imponendo la verifica del raggiungimento dei dosaggi target stabiliti per ciascuno dei quattro pilastri fondamentali: ARNI/ACE-i/ARB, BBs, MRA e SGLT2i.

Il dosaggio somministrato impatta sul beneficio. La letteratura clinica correla direttamente la titolazione ottimale alla riduzione della mortalità per tutte le cause e all'abbattimento dei ricoveri per scompenso acuto [8], dimostrando che la sottotitolazione ingiustificata genera un incremento del rischio clinico assimilabile all'omissione completa di una classe molecolare. Tuttavia, l'inerzia terapeutica frena l'applicazione reale. A dispetto delle evidenze disponibili, la pratica medica quotidiana registra la persistenza di condotte prescrittive conservative, con una quota rilevante di pazienti trattata attraverso regimi farmacologici incompleti o sottodimensionati a livello di dosaggio.

Per quantificare l'aderenza alle linee guida per l'HFrEF in modo oggettivo, la letteratura ha proposto diversi indici sintetici di valutazione. Tra questi, lo *Heart Failure collaboratory* (HFC) *Score*, introdotto originariamente da Abraham et al.

(2020) [47], è stato uno dei primi tentativi di quantificazione dell'aderenza della terapia alle raccomandazioni ESC. Tale punteggio prendeva in considerazione i quattro pilastri della quadruplice terapia, a eccezione della classe SGLT2i, e l'idralazina-nitrato, attribuendo alla presenza nella terapia di ogni classe di farmaci un punteggio specifico. Tuttavia, la valutazione del dosaggio somministrato era riservata esclusivamente alle classi degli ACE-i/ARB e dei Beta-bloccanti. In particolare, questa quantificazione non avveniva in modo continuo, bensì tramite una discretizzazione binaria che prevedeva l'assegnazione di due soli valori di punteggio: uno più elevato qualora la posologia superasse la soglia del 50% della dose target di linea guida, e uno inferiore in caso contrario. La somma dei singoli contributi di tutti i farmaci definiva il punteggio finale, variabile all'interno di un intervallo compreso tra 0 e 6.

Questo punteggio risulta piuttosto grezzo sotto diversi aspetti. Per la maggior parte delle classi farmacologiche, infatti, non viene soppesata la dose somministrata rispetto alla dose target, e per le classi in cui ciò accade, l'adozione di due sole finestre discrete anziché di una scala continua penalizza la precisione della misura. Inoltre, essendo antecedente alle evidenze dei benefici legati agli SGLT2i, questo punteggio non calcola il contributo di tale classe di farmaci, risultando così incompleto per la valutazione dell'approccio moderno basato sulla quadruplice terapia.

In seguito alla nascita dello score HFC originale, sono stati ideati numerosi sistemi di misurazione dell'aderenza terapeutica, concepiti per superare la rigidità del modello originario e introdurre gli aggiornamenti introdotti dalle più recenti linee guida ESC. In particolare, a partire dal 2020 sono stati individuati in letteratura 25 score GDMT distinti, 13 dei quali derivati dallo score HFC originale [48].

Nel contesto del presente lavoro di tesi, sono state selezionate alcune di queste metriche candidate (presentate nel sottocapitolo 5.2) per valutare l'aderenza della terapia indicata nelle lettere cliniche a partire dai dati precedentemente estratti nello step di IE ad opera degli LLM. A tal fine, lo schema JSON contenente le informazioni estratte dal testo è stato opportunamente strutturato in modo tale da riassumere, per ogni farmaco, le informazioni necessarie al calcolo di questi punteggi, ovvero la classe

di appartenenza, il principio attivo, la dose singola e il numero di assunzioni giornaliere. Questa formalizzazione consente ad un sistema a regole (deterministico) di individuare automaticamente la dose giornaliera totale somministrata, confrontarla con i dosaggi target di linea guida e calcolare gli score di aderenza in modo oggettivo e riproducibile.

5.2 Selezione degli score di aderenza

Come anticipato nel Capitolo 2, in seguito alla nascita dello score HFC originario, sono state presentate numerose metriche finalizzate a quantificare l'aderenza della terapia farmacologica per l'HFrEF alle relative linee guida ESC. Tuttavia, ognuno di questi punteggi presenta delle peculiarità in termini di farmaci tenuti in considerazione, sensibilità della misura e contesto di applicazione.

Nell'ambito del presente lavoro di tesi, a partire dai numerosi candidati, gli indici da adottare sono stati selezionati sulla base delle seguenti caratteristiche:

- **Valutazione istantanea dell'aderenza**: la capacità dell'indice di fotografare l'aderenza della terapia nel momento specifico della dimissione ospedaliera (il momento in cui viene redatta la lettera di dimissione dalla quale vengono estratti i dati per il calcolo dei punteggi).
- **Inclusione ed esclusività dei farmaci della quadruplice terapia**: la focalizzazione mirata ed esaustiva dei quattro pilastri farmaceutici (ARNI/ACE-i/ARB, BBs, MRA e SGLT2i), escludendo la presenza di altri farmaci non primariamente impattanti sulla mortalità e la ri-ospedalizzazione.
- **Granularità del contributo dei dosaggi**: nonostante i contributi di ogni classe farmacologica vengano calcolati per la maggior parte su scala discreta, viene data la priorità agli score calcolati tramite una valutazione continua, o quantomeno il più dettagliata possibile, dei dosaggi rispetto alle dosi target di linea guida.

I punteggi selezionati in base a queste caratteristiche nel corso del lavoro di tesi sono i seguenti:

- ***Modified Heart Failure Collaboratory*** **(mHFC)** [49]: aggiornamento dello

score HFC originario; ne esistono due versioni, entrambe caratterizzate da un sistema di punteggio a contributi discreti. La prima integra le classi dei beta-bloccanti, degli MRA, dei RAASi e degli SGLT2i, mentre la seconda associa un contributo allo score finale anche alla presenza di ivabradina, vericiguat e idralazina/nitrati. Nel presente lavoro di tesi è stata calcolata la prima delle due versioni; la scelta è motivata da diverse ragioni: in primo luogo, i farmaci aggiuntivi della versione estesa non fanno parte dei quattro pilastri su cui si concentra la fase di estrazione; inoltre, di tutte le lettere utilizzate come dataset per la fase di IE e la successiva fase di calcolo degli score, solamente tre documenti contengono una prescrizione di ivabradina, mentre vericiguat e idralazina/nitrati non compaiono mai all'interno del dataset di riferimento. Pertanto, considerata la limitata presenza nel dataset dei farmaci aggiuntivi previsti dalla seconda versione e il conseguente impatto marginale sul punteggio finale, si è scelto di utilizzare la prima versione del mHFC.

Il contributo di ogni farmaco, compresi quelli facenti parte della versione estesa del punteggio, è mostrato nel dettaglio in Figura 7:

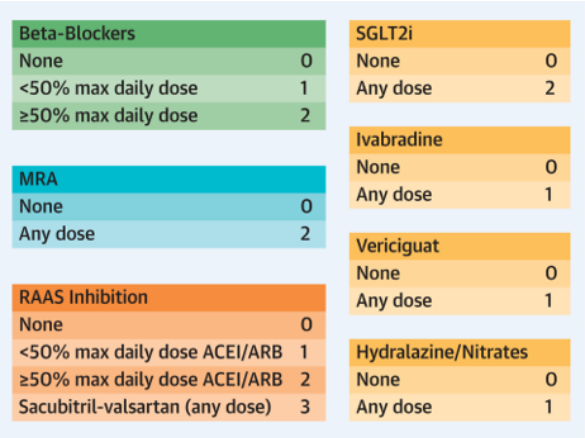


Figura 7. *Contributi specifici relativi a tutti i farmaci considerati dallo score mHFC (estratto dall'articolo di riferimento per questo score di aderenza [49]).*

- ***Optimization performance score*** **(OPS)** [50]: indicatore mirato ed esaustivo per la quadruplice terapia; per quasi tutti i farmaci presenta dei punteggi

discreti, con due sole finestre per il calcolo del contributo di ognuno rispetto alla dose target. L'unico per il quale non viene considerata la dose è SGLT2i, del quale viene valutata unicamente la presenza/assenza con un contributo fisso, a prescindere dalla dose somministrata. Tuttavia, tale differenza è coerente con la Tabella 1 proveniente dalla linea guida ESC, che non prevede una variazione nel tempo della dose somministrata.

Di seguito, in Figura 8, viene mostrata la tabella riassuntiva per il calcolo di tale indice:

Optimization Potential Score (OPS)			
	Points		
	None	**Low Dose**	**Target Dose**
ACEI/ARB/ARNI	0	1	2
Beta Blocker	0	1	2
MRA	0	1	2
ARNI	0	1	2
SGLT2i	0	-	2

Figura 8. *Tabella contenente i contributi dei farmaci della quadruplice terapia per il calcolo dello score OPS (estratta dall'articolo di riferimento per questo score di aderenza [50]).*

- ***Modified Continuous 4 pillar* (mC4P)** [51]: metrica evoluta derivata dall'HFC originale; include tutti e solamente i farmaci della quadruplice terapia e calcola dei contributi continui per ognuno di essi, tenendo in considerazione sia la presenza delle diverse classi farmacologiche sia la dose somministrata rispetto alla dose target specificata nelle linee guida.

  Il calcolo dei diversi contributi avviene nel seguente modo:

  - **Per ACE-i/ARB, BBs, MRA e SGLT2**: 2 + 2 × (% dose target)
  - **Per ARNI**: 4 + 2 × (% dose target)

5.3 Estrazione multimodale dei dosaggi dalle linee guida

Il calcolo degli score di aderenza, come illustrato nel sottocapitolo precedente, si basa frequentemente sulla percentuale della dose somministrata calcolata rispetto alla dose target indicata nelle linee guida. Al fine di effettuare questo calcolo autonomamente, l'algoritmo necessita di un archivio di riferimento contenente le dosi d'attacco e le dosi target raccomandate per ciascun principio attivo. Invece di cablare manualmente queste informazioni nel codice, si è optato per un approccio più dinamico, basato su modelli linguistici multimodali, in grado di interpretare direttamente le immagini e di ricavarne le informazioni necessarie. Tale approccio garantisce la riutilizzabilità e scalabilità del sistema nel caso di futuri aggiornamenti delle linee guida di riferimento o di estensione del sistema ad altre patologie.

In Tabella 1, proveniente dalle linee guida ESC, sono schematizzate tutte le informazioni necessarie, ovvero la dose d'attacco e la dose target per ogni farmaco previsto dal regolamento. Elaborando tale tabella in formato immagine attraverso il modello Qwen 2.5 VL (invocato tramite API Amazon Bedrock) e sfruttando la tecnica di *Structured Outputs* vincolata tramite schemi Pydantic come per la fase di IE descritta nel capitolo 4, in un unico passaggio sono state estratte le seguenti informazioni per ogni farmaco:

- La classe farmacologica di appartenenza (*drug_class*).
- Il nome generico del principio attivo (*generic_drug_name*).
- Il dosaggio di partenza e target (*starting_dose* e *target_dose*).
- La frequenza di somministrazione giornaliera (*starting_dose_daily_intakes* e *target_dose_daily_intakes*), facendo convertire automaticamente all'LLM le relative abbreviazioni mediche (per esempio, o.d. corrisponde a 1, b.d. corrisponde a 2 ecc.).

Il risultato di questa estrazione viene salvato in due formati nei file *HFrEF_medication_table.json* e *HFrEF_medication_table.csv*, che fungono come base di conoscenza relativa alle linee guida per la successiva fase di calcolo degli score.

Il formato con il quale sono salvate le informazioni è lo stesso utilizzato nelle linee guida di riferimento; eventuali normalizzazioni e trasformazioni avvengono subito prima della successiva fase di calcolo degli indici.

5.4 Algoritmo di calcolo automatico degli score

Dopo aver estratto in forma strutturata sia le informazioni legate alle dosi di partenza e target per ogni farmaco dalla linea guida sia la terapia del paziente a partire dalle lettere cliniche, il modulo *GDMT_scores* esegue il calcolo vero e proprio degli indici di aderenza terapeutica secondo le formule dettagliate nel Sottocapitolo 5.2.

In questo contesto risulta fondamentale soffermarsi sui possibili formati con i quali sono annotati i dosaggi nella linea guida e su come questi siano stati gestiti nella fase di calcolo. In particolare, in Tabella 1 compaiono queste tre possibili formulazioni:

- **Valore singolo**: per esempio, «50 mg» o «6,25 mg»; rappresenta il caso più semplice e viene convertito direttamente in un valore scalare che rappresenta la dose giornaliera, ottenuta moltiplicando la dose per la frequenza prescritta.
- **Terapie di combinazione**: per esempio, «97/103 mg» nel caso di sacubitril/valsartan oppure «37,5 mg/20 mg» nel caso di idralazina/isosorbide. In questo caso, le dosi dei singoli principi attivi vengono sommate.
- **Range di dosaggio**: per esempio, «10–20 mg» o «2,5–5 mg»; con questo tipo di notazione, il range viene salvato sotto forma di tupla (D_{min},D_{max}).

Quest'ultimo formato, non essendo trasformato in un singolo valore scalare, ha la necessità aggiuntiva di essere gestito diversamente dagli altri casi anche per quanto riguarda il calcolo della percentuale della dose somministrata rispetto alla dose target ($\%_{target}$). Ciò avviene in base alla seguente logica:

$$\begin{cases} \%_{target} = 100\% \; se \; D_{min} \leq D_{paziente} \\ \%_{target} = \dfrac{D_{paziente}}{D_{min}} \; se \; D_{paziente} < D_{min} \end{cases}$$

Un'altra sfida operativa nel calcolo degli score risiede nella variabilità terminologica

dei nomi dei farmaci: le linee guida ESC riportano la nomenclatura internazionale in inglese (ad esempio, Bisoprolol e Metoprolol succinate), mentre una lettera clinica può riportare gli stessi principi attivi in altre lingue, in base alla provenienza del documento, come nel caso delle 20 lettere utilizzate nel presente lavoro, redatte in italiano (ad esempio, Bisoprololo e Metoprololo succinato). Questa differenza può determinare un mancato allineamento tra i nomi dei farmaci riportati nei due documenti, con conseguente erronea riduzione dello score di aderenza.

Per superare questo ostacolo, la funzione *lookup_target_dose* implementa un algoritmo di corrispondenza a due livelli, basato sul confronto tra stringhe: per prima cosa si verifica la corrispondenza esatta delle due stringhe normalizzate in minuscolo; successivamente, se il match diretto fallisce, le stringhe estratte dalla lettera e dal database di riferimento vengono scomposte in singoli termini tramite espressioni regolari, scartando la punteggiatura e i termini con lunghezza inferiore a quattro caratteri. Per ciascuna coppia di parole (w_1, w_2), l'algoritmo incrementa uno score di similarità:

- Se le due parole sono identiche, viene associato un contributo pari a 1: $\Delta score = 1 \; se \; w_1 = w_2$ (ad esempio, «metoprolol» = «metoprolol»).
- Se invece non sono identiche, si valuta la possibilità che una sia inclusa nell'altra (ad esempio, «Bisoprolol» in «Bisoprololo»), imponendo una lunghezza minima di cinque caratteri, e viene calcolato un punteggio ponderato sulla base della sovrapposizione: $\Delta Score = 0.8 \times \frac{min(w_1,w_2)}{max(w_1,w_2)}$.

Infine, la miglior corrispondenza con i farmaci elencati dalla linea guida viene considerata valida solo se il punteggio complessivo supera la soglia di sicurezza ($Score \geq 0.7$). Questo approccio garantisce una tolleranza alle variazioni linguistiche elevata, evitando allo stesso tempo i falsi positivi legati a nomi corti o radicali comuni. Nel contesto del matching approssimato tra stringhe [52], si tratta di uno score riconducibile ad una misura di similarità basata sulle lunghezze, calcolato come rapporto normalizzato tra le lunghezze delle stringhe. I parametri dell'algoritmo sono stati scelti in modo euristico, con l'obiettivo di tollerare solamente le variazioni tra le denominazioni italiane e inglesi dei principi attivi, evitando abbinamenti diversi tra farmaci diversi. In particolare, il peso 0,8 combinato con la soglia 0,7 impone che il rapporto tra le lunghezze delle due parole sia almeno pari a 0,875.

Un’ultima situazione che lo scenario ideale non tiene in considerazione è il caso in cui un farmaco presente nella lettera clinica venga correttamente classificato all'interno di una classe GDMT, ma per motivi di mancata corrispondenza nel database non sia possibile recuperare la dose target esatta. Un esempio molto comune all’interno del dataset è rappresentato dal Canrenone, un farmaco appartenente alla classe degli MRA ma non riportato nella linea guida ESC (Tabella 1).
In questo contesto, l’algoritmo adotta un approccio conservativo assegnando il punteggio minimo previsto dallo score specifico per quella classe. Contestualmente, l'evento viene tracciato inserendo un messaggio di avviso nel campo *note* dell’oggetto *DrugScoreDetail*, in modo tale da garantire la trasparenza dell’output fornito al clinico.

Gli score calcolati per ogni lettera e a partire dai dati estratti da ognuno dei cinque modelli utilizzati nella fase di IE sono riportati integralmente nell’Appendice C.

6. Risultati e discussione

6.1 Prestazioni del processo di estrazione di informazioni

La valutazione delle prestazioni del modulo di estrazione delle informazioni basato su LLM è stata effettuata su un dataset di 20 lettere cliniche reali di pazienti affetti da HFrEF, rese anonime e messe a disposizione dall'ICS Maugeri di Pavia. Questi documenti hanno la forma semi-strutturata tipica dei testi ospedalieri, caratterizzati dall'utilizzo di acronimi, abbreviazioni e sintassi eterogenea.

Per valutare la qualità delle informazioni estratte, sono stati utilizzati come riferimento i form compilati manualmente dai medici cardiologi per ciascuna lettera. Tali form rispecchiano esattamente la struttura dello schema JSON utilizzato nello step di estrazione e descritto nel Capitolo 4 (comprendente dati anagrafici, parametri ematochimici ed emodinamici, comorbilità ed elenco dei farmaci relativi a terapie d'ingresso e di dimissione). Ciò garantisce un confronto diretto campo per campo tra il dato estratto automaticamente dai modelli e quello validato dallo specialista. In particolare, le performance degli LLM nella fase di estrazione sono state valutate sulla base del numero di corrispondenze corrette tra il dato estratto e l'informazione corrispondente riportata nel form rispetto alle 20 lettere totali; questa operazione è stata eseguita per ogni sezione dello schema JSON e per ognuno dei cinque modelli utilizzati.

Per contestualizzare la valutazione delle prestazioni del modulo di IE, è opportuno chiarire la logica metodologica alla base del calcolo degli indici di accuratezza.
Dato che, a parte la terapia di dimissione, utile al calcolo degli score di aderenza, in questa fase del progetto GUIDE-AI non è ancora richiesto di estrarre le informazioni in un formato stringente, questi punteggi sono stati calcolati sulla base della presenza e correttezza semantica delle informazioni estratte rispetto al riferimento. In particolare, sono stati seguiti questi principi:

- **Prevalenza del contenuto informativo su formato e lingua**: La valutazione si è basata sull'entità semantica del dato clinico, senza contare come errori le

difformità dovute alla lingua o alla formattazione, a patto che il significato dell'informazione estratta corrispondesse a quella registrata nel form dallo specialista.

- **Criterio di inclusione per le sezioni a lista**: Per i campi strutturati come insiemi di entità (quali esami di laboratorio, sintomi o patologie concomitanti), il punteggio è stato assegnato in base alla capacità del modello di rilevare tutti gli elementi target indicati nella scheda di annotazione medica. La presenza di elementi aggiuntivi rispetto a quelli individuati dallo specialista non è stata penalizzata, dal momento che si tratta di domini descritti in modo qualitativo e i cui confini non sono strettamente delimitati a priori. Inoltre, la totale assenza di allucinazioni durante il processo di estrazione garantisce che tali elementi aggiuntivi rappresentino esclusivamente un dettaglio contestuale ridondante, ininfluente ai fini del calcolo dell'aderenza e del tutto privo di impatti negativi sui futuri sviluppi del sistema.
- **Logica booleana e gestione dei valori nulli**: In conformità con le informazioni fornite agli LLM tramite il prompt di sistema («If a field is not mentioned or cannot be inferred, set it to *null*», in italiano «Se un campo non è menzionato o non può essere dedotto, impostalo a *null*») per evitare di assegnare valori alle variabili di output non basate sulle evidenze, la presenza del valore *null* nello schema JSON è stata interpretata come valore booleano *false*.
- **Logica di valutazione della terapia farmacologica**: Per quanto riguarda la singola dose di assunzione e il numero di assunzioni giornaliere, non sono stati penalizzati errori commessi su farmaci non facenti parte della terapia legata all'HF, sulla base della quale devono essere calcolati gli score di aderenza. Queste informazioni, infatti, nel contesto del presente lavoro di tesi, hanno la sola funzione di calcolare tali indici e i prompt legati alla loro estrazione è basato sulla modalità di assunzione specifica (numero di compresse al giorno).

Di seguito, in Tabella 6, vengono mostrati gli indici di accuratezza calcolati per ogni sezione dello schema JSON e del form (sulle righe), e per ognuno dei cinque LLM utilizzati nel processo di IE (sulle colonne):

Sezione / Campo Schema JSON	DeepSeek v3.2	Gemma 4 (31B)	GPT-OSS 120B	Qwen 3 (32B)	Claude 4.6 Sonnet
demographic_info	19/20	**20/20**	18/20	**20/20**	**20/20**
drug_intolerances	**20/20**	**20/20**	**20/20**	**20/20**	**20/20**
lab_values	16/20	15/20	18/20	3/20	**20/20**
comorbidities	18/20	18/20	15/20	17/20	**19/20**
disorders_and_conditions	17/20	18/20	**19/20**	18/20	17/20
worsening_heart_failure	19/20	**20/20**	18/20	18/20	14/20
loop_diuretic	18/20	19/20	18/20	17/20	**20/20**
peripheral_edema	17/20	17/20	17/20	17/20	**18/20**
hfrEF_90_days	14/20	14/20	5/20	**17/20**	14/20
comfort_care_only	**20/20**	**20/20**	**20/20**	**20/20**	**20/20**
hfrEF_etiology	**19/20**	**19/20**	**19/20**	**19/20**	**19/20**
discharge_date	**20/20**	19/20	19/20	14/20	18/20
current_therapy	15/20	14/20	13/20	12/20	**17/20**
discharge_therapy	18/20	**19/20**	**19/20**	12/20	18/20

***Tabella 6**. Performance della fase di estrazione, quantificate come numero di match tra il dato estratto e l'informazione corrispondente riportata nel form rispetto alle 20 lettere totali. Per ciascuna sezione vengono evidenziati i risultati dell'estrazione dei modelli con le performance migliori.*

Per ciascuna sezione, sulla base della Tabella 6, si possono fare le seguenti osservazioni:

- ***demographic info***: Il processo di anonimizzazione dei documenti ha reso questa sezione povera in contenuto a priori; l'unica informazione di importanza clinica rilevante presente in modo esplicito in tutte le lettere è l'età. Nonostante venga esplicitato nel relativo prompt di estrarre l'età e non la data di nascita, quasi sempre, a eccezione di rari casi nell'utilizzo di Qwen e GPT, il dato è stato estratto nella seconda forma. Ad ogni modo, il contenuto semantico dell'informazione estratta rimane lo stesso.
- ***Drug intolerance***: Nonostante l'accuratezza assoluta degli output, bisogna

considerare l'assenza totale di intolleranze da parte di tutti i pazienti. Dunque, non è stato possibile verificare a pieno la capacità degli LLM nell'estrarre correttamente le intolleranze, quando presenti.

- ***Lab values***: Il modello che spicca per delle performance particolarmente negative è Qwen, che nella maggior parte dei casi non ha estratto alcun valore oppure ha registrato i test senza popolare i campi legati al valore, all'unità di misura o alla data. Per tutti gli altri modelli invece gli errori sono dovuti a sporadici test mancanti nei dati estratti. Inoltre, spesso sono stati estratti test e misure aggiuntive non annotate dagli specialisti all'interno del form (ad esempio, peso e altezza). Ciò è dovuto all'approccio conservativo adottato nei relativi prompt, che favorisce la presenza di elementi aggiuntivi rispetto all'assenza di test utili alla valutazione dell'aderenza terapeutica.
- ***Comorbidities e Disorders and Conditions***: Gli errori sono dovuti all'assenza sporadica di elementi o al confondimento della classe di appartenenza; infatti, condizioni classificate come comorbidità nei form di riferimento sono state classificate come disordini dagli LLM, e viceversa. Inoltre, sono stati estratti elementi non annotati dai medici ma comunque presenti nelle lettere cliniche. Questa variabilità nel processo di estrazione è sintomo della natura poco delineata di queste due sezioni, non solo per i modelli utilizzati, ma anche per gli specialisti che si sono occupati di annotarle.
- ***Worsening HF***: I modelli hanno dimostrato ottime performance, a eccezione di Claude, che ha ottenuto il punteggio più basso (14/20).
- ***Loop diuretic***: Quasi tutti gli errori consistono in falsi positivi; in particolare, solamente uno di questi risulta un falso negativo.
- ***Peripheral edema***: Vengono commessi gli stessi errori da tutti i modelli (a eccezione di Claude che commette un errore in meno rispetto agli altri).
- ***HFrEF 90 days***: Deepseek, Gemma e soprattutto GPT, che ha classificato solo un caso, hanno registrato molti *null*, mentre Claude ha commesso molti errori di classificazione. Ciò è dovuto al fatto che l'informazione relativa alla diagnosi di HFrEF è spesso poco esplicita; per esempio, può essere legata ad una diagnosi registrata nel documento molto tempo prima dell'ultima visita, oppure deve essere dedotta a partire da elementi sparsi nel documento (per

esempio dalla combinazione di diagnosi di HF e LVEF<40%). Il modello che sembra avere performance migliori è Qwen.

- ***HFrEF etiology***: Tutti i modelli sbagliano solo una volta in corrispondenza della lettera 18; il paziente, infatti, soffre di cardiopatia ischemica, ma nel form l'eziologia di HFrEF è annotata come "non ischemica".
- ***Discharge date***: Gli errori sono dovuti al confondimento da parte dei modelli della data della visita con una data di dimissione. Qwen sembra commettere questo errore con più frequenza degli altri modelli.
- ***Current therapy***: Osservando le performance legate all'estrazione di questa sezione, che risultano le più scarse dell'intera fase di IE, si potrebbe erroneamente arrivare alla conclusione che i modelli utilizzati non siano in grado di interpretare questo tipo di informazioni, compromettendo la possibilità di derivare indici di aderenza affidabili. Tuttavia, la differenza di performance tra l'estrazione di questa sezione rispetto alla successiva (*Discharge therapy*) dimostra l'esatto contrario, confermando la riuscita della fase di affinamento del prompt condotta su quest'ultima. Dal punto di vista metodologico e applicativo, infatti, la sezione di maggiore interesse per il calcolo degli *score* di aderenza e per la valutazione della GDMT è la *Discharge Therapy*, in quanto riflette il regime posologico aggiornato e validato al termine della visita o della degenza. Le informazioni contenute nella *Current therapy* sono invece superflue, riferite ad una posologia destinata a essere sovrascritta. Pertanto, nel corso del lavoro di tesi si è preferito affinare con maggiore cura i prompt relativi alla *Discharge therapy* senza estendere quelli relativi alla *Current therapy*, per evitare di appesantire il processo di estrazione. Queste due sezioni sono inoltre molto simili non solo nella struttura generale, ma anche nel contenuto. Esse, infatti, differiscono sempre per pochi farmaci (nella maggior parte dei casi si tratta dell'introduzione di un singolo nuovo farmaco al termine della visita). L'unica differenza, dunque, risiede proprio nel prompt specifico per ognuna di esse, responsabile della notevole differenza di performance. A prova di ciò, la principale fonte di errore registrata nella *Current Therapy* è l'omissione di alcuni principi attivi (verificatasi in dieci casi complessivi tra i vari modelli); nella *Discharge therapy* questo problema risulta

completamente risolto.

- ***Discharge therapy***: L'estrazione legata a questa sezione è caratterizzata da ottime performance a eccezione di Qwen, che commette alcuni errori relativi alla classificazione farmacologica e al numero di dosi somministrate giornalmente. Dato che le informazioni contenute in questa sezione sono riutilizzate nel calcolo degli score di aderenza, per comprendere a pieno le criticità dello step successivo del lavoro di tesi, gli errori commessi nella fase di estrazione per ogni paziente sono elencate di seguito:
    - **Paziente 1**: Qwen ha riportato erroneamente una quantità di 1,5 unità/giorno per Sacubitril/Valsartan (Entresto), anziché 2, come riportato nel testo di riferimento.
    - **Paziente 2**: Qwen ha estratto una quantità di 0,5 unità/giorno per il Metoprololo, anziché l'unità intera (1) riportata nella lettera di dimissione.
    - **Paziente 3**: per il farmaco Sacubitril/Valsartan (Entresto), Qwen ha calcolato una quantità giornaliera pari a 0,5 compresse, anziché 2.
    - **Paziente 4**: Claude ha estratto il valore *null* per il campo relativo alla posologia del Dapagliflozin (Forxiga), a fronte di una somministrazione effettiva di 0,5 compresse/giorno.
    - **Paziente 7**: Qwen ha estratto erroneamente la posologia di Ramipril e Bisoprololo, registrando 1 compressa/giorno per ciascun principio attivo, anziché le 2 compresse/giorno effettivamente prescritte.
    - **Paziente 13**: Qwen non ha riconosciuto il principio attivo Canrenoato di potassio (*Potassium canrenoate*) come appartenente alla classe degli MRA.
    - **Paziente 16**: analogamente al Paziente 2, Qwen ha estratto per il Metoprololo una quantità pari a 0,5 anziché 1.
    - **Paziente 17**: in questo caso si sono verificate diverse anomalie: il modello GPT non ha riconosciuto il farmaco Synjardy (Empagliflozin/Metformina) come appartenente alla classe degli SGLT2i, mentre Qwen ha sovrastimato la quantità giornaliera del diuretico Lasix e dell'MRA Lasix, oltre ad aver erroneamente attribuito

0.5 invece di 5 mg alla quantità di principio attivo contenuta nella singola dose di bisoprololo (BB).

Tuttavia, l'errore più rilevante consiste nell'errata registrazione del dosaggio relativo al farmaco Synjardy (SGLT2) da parte di tutti i modelli utilizzati. Ciò è dovuto al formato utilizzato, che indica la compresenza di due principi attivi, metformina e empagliflozin, le cui quantità vengono indicate con l'espressione «Metformina + empagliflozin 5 mg/1000 mg». La natura implicita dell'informazione e la mancata esemplificazione nei prompt di estrazione secondo l'approccio *few-shot*, dovuta alla singolare occorrenza di questa formulazione, hanno portato all'estrazione errata del dosaggio.

In conclusione, il modulo di IE costruito in questa fase del lavoro di tesi risulta avere complessivamente ottime performance, in particolare per quanto riguarda l'estrazione delle informazioni riguardanti la terapia farmacologica, che costituisce una base di partenza affidabile per il successivo calcolo automatico degli score di aderenza.

6.2 Valutazione del sistema di scoring dell'aderenza

A partire dalle informazioni estratte in forma strutturata dalle 20 lettere cliniche, la fase successiva del lavoro di tesi ha previsto il calcolo dei tre score di aderenza mHFC, OPS e mC4P, secondo le formule descritte nel Capitolo 5. Poiché in fase di annotazione non è stato richiesto ai medici di registrare esplicitamente questa informazione all'interno dei form, i valori di riferimento sono stati calcolati a partire dalle informazioni legate ai farmaci annotati dagli specialisti.

Le performance di questo modulo, coerentemente con la valutazione della fase di IE, sono state quantificate con la stessa metrica adottata nel Sottocapitolo 6.1: per ciascun modello, il sistema ha calcolato in automatico i tre score di aderenza a partire dalle informazioni estratte dagli LLM, e per ogni score è stato riportato il numero di lettere per cui il valore calcolato automaticamente coincide con il valore di riferimento, calcolato manualmente, sul totale delle 20 lettere cliniche. In Tabella 7 sono riportati i risultati di questa analisi:

Score di Aderenza	Claude 4.6 Sonnet	DeepSeek v3.2	Qwen 3 (32B)	GPT-OSS 120B	Gemma 4 (31B)
mHFC	**20/20**	**20/20**	13/20	19/20	**20/20**
OPS	**20/20**	**20/20**	17/20	19/20	**20/20**
mC4P	18/20	**19/20**	12/20	**19/20**	**19/20**

***Tabella** 7. Performance della fase di calcolo degli score di aderenza in termini di numero di corrispondenze tra lo score calcolato automaticamente e quello calcolato manualmente a partire dalle informazioni riportate nel form rispetto alle 20 lettere totali. In grassetto sono evidenziate le performance migliori per score di aderenza.*

Dall'analisi dei risultati della Tabella 7 si deduce che le discrepanze riscontrate nel calcolo degli score di aderenza non sono dovute ad errori della logica deterministica usata per calcolare i punteggi, ma costituiscono una propagazione diretta dell'errore derivante dalla precedente fase di IE.

Gli errori derivano prevalentemente dalle informazioni estratte da Qwen 3 (32B), che presenta discrepanze su tutti e tre gli score. Per gli altri modelli le discrepanze sono isolate: riguardano esclusivamente l'mC4P per Claude 4.6 Sonnet (una), DeepSeek v3.2 (una) e Gemma 4 (due), mentre GPT-OSS 120B presenta una discrepanza per ciascuno dei tre score. Facendo una sintesi delle discrepanze osservate, gli errori del modulo di estrazione che impattano direttamente sul calcolo dell'aderenza possono essere ricondotti a due macrocategorie:

- **Conteggio delle somministrazioni giornaliere**: l'errata conversione

aritmetica delle frazioni o la mancata interpretazione della frequenza di somministrazione (ad esempio, «1 cpr 8-20») determina una penalizzazione (Pazienti 2, 3, 4, 7 e 16) oppure un incremento (Paziente 17) ingiustificati dello score di aderenza globale.

- **Classificazione farmacologica**: L'errata classificazione dei farmaci porta ad una sovrastima (Paziente 10) oppure ad una sottostima (Pazienti 13 e 17) del punteggio calcolato.

In conclusione, gli indici calcolati confermano che l'algoritmo deterministico di calcolo degli score di aderenza garantisce un'accuratezza del 100% laddove l'input estratto dall'LLM sia privo di errori. Le deviazioni rispetto al riferimento riflettono i limiti dei singoli modelli nell'interpretazione del testo clinico non strutturato, evidenziando come la precisione della fase di IE rappresenti il vero collo di bottiglia per l'intera architettura.

6.3 Esempio illustrativo: dalla lettera clinica agli score di aderenza

A titolo esemplificativo, si riporta il percorso della lettera del Paziente 13 attraverso le tre fasi del sistema: lettera clinica, informazioni estratte con confronto con le annotazioni dei cardiologi e calcolo degli score di aderenza. Il caso consente di confrontare un modello che estrae correttamente la terapia (Claude 4.6 Sonnet) con uno che commette un errore di classificazione farmacologica (Qwen 3, 32B).

La tabella 8 riporta la terapia da proseguire o iniziare a partire dalla visita, così come scritta nella lettera:

Nome commerciale (Principio attivo)	**Dosaggio**	**Stato**
TRIATEC*2,5MG 28 CPR DIV. *(Ramipril 2,5 mg 28 compresse divisibili (importato dalla Germania))*	1 cp la mattina, 1 cpr la sera la sera	Nuova
FUROSEMIDE DOC*25MG 30 CPR *(Furosemide 25 mg 30 compresse)*	1+1/2 la mattina; 1 il pomeriggio	

Nome commerciale (Principio attivo)	Dosaggio	Stato
PANTOPRAZOLO ABC*20MG 14CPR *(Pantoprazolo sale sodico sesquidrato 20 mg 14 compresse gastroresistenti)*	1 cpr al giorno	
POTASSIO CAN EG*100MG 20 CPR *(Potassio canrenoato 100 mg 20 compresse rivestite con film)*	1/ 2 cpr alle 16	
FORXIGA*10MG 28CPR RIV *(Dapagliflozin 10 mg 28 compresse rivestite con film)*	1 cpr al giorno	
PROCORALAN*5MG 56CPR *(Ivabradina cloridrato 5 mg 56 compresse rivestite con film)*	1 cpr alle 8; 1 cpr alle 20	
CARDIOASPIRIN*100MG 30CPR *(Acido acetilsalicilico 100 mg 30 compresse gastroresistenti)*	1 cpr alle 12	
BISOPROLOLO ALM*1,25MG 28CPR *(Bisoprololo fumarato 1,25 mg 28 compresse rivestite con film)*	2 cpr al mattino, 1 cpr la sera	

Tabella 8. *Estratto della sezione «Terapia farmacologica da proseguire/iniziare» della lettera del Paziente 13.*

Per ciascun farmaco i modelli hanno estratto nello schema JSON il principio attivo, la dose singola, il numero di assunzioni giornaliere e la classe farmacologica. La Tabella 9 riporta le informazioni rilevanti per il calcolo degli score relative ai farmaci della quadruplice terapia. Claude ha estratto tutte le informazioni in accordo con le annotazioni dei cardiologi, mentre Qwen ha classificato il potassio canrenoato come «Altro» (*Other*) anziché come MRA. I restanti farmaci (furosemide, pantoprazolo, ivabradina e acido acetilsalicilico) sono stati estratti in modo identico dai due modelli e, nelle versioni degli score adottate, non concorrono al calcolo.

Principio attivo	Dose singola × assunzioni/giorno	Dose giornaliera	Classe (Claude)	Classe (Qwen)
Ramipril	2,5 mg × 2	5 mg	ACE-i	ACE-i
Bisoprololo	1,25 mg × 3	3,75 mg	Beta-bloccante	Beta-bloccante
Potassio canrenoato	100 mg × 0,5	50 mg	MRA	**Altro**
Dapagliflozin	10 mg × 1	10 mg	SGLT2i	SGLT2i

Tabella 9. *Informazioni estratte nella sezione discharge_therapy per i farmaci della quadruplice terapia: confronto tra l'annotazione dei cardiologi (coincidente con l'output di Claude) e l'output di Qwen. Dose singola e numero di assunzioni coincidono nei due modelli.*

A partire da queste informazioni, il modulo deterministico calcola la dose giornaliera di ciascun farmaco, la confronta con la dose target indicata in Tabella 1 e assegna i contributi secondo le regole dei tre score (Sottocapitolo 5.2). La Tabella 10 mostra il risultato: l'errata classificazione del potassio canrenoato fa sì che il contributo della classe MRA non venga conteggiato, con una sottostima di tutti e tre gli score.

Pilastro	Farmaco (dose giornaliera / target)	% dose target	mHFC	OPS	mC4P
RAASi	Ramipril (5 / 10 mg)	50%	2	1	3,00
Beta-bloccanti	Bisoprololo (3,75 / 10 mg)	37,5%	1	1	2,75
MRA	Potassio canrenoato (50 mg)	n.d.	2 (Qwen: 0)	1 (Qwen: 0)	2,00 (Qwen: 0)
SGLT2i	Dapagliflozin (10 / 10 mg)	100%	2	2	4,00
Totale			**7** (Qwen: **5**)	**5** (Qwen: **4**)	**11,8** (Qwen: **9,8**)

Tabella 10. *Contributi di ciascun farmaco agli score mHFC, OPS e mC4P per la lettera del Paziente 13, calcolati dal modulo deterministico. Tra parentesi i valori ottenuti a partire dalle informazioni estratte da Qwen. Il potassio canrenoato non compare in Tabella 1, dove per gli MRA sono riportati eplerenone e spironolattone; il contributo attribuito è quindi il minimo previsto per ciascuno degli score.*

L’esempio conferma quanto emerso nell’analisi aggregata: il modulo di calcolo applica correttamente le regole di ciascuno score ai dati ricevuti, e la discrepanza rispetto al riferimento dipende interamente dall’errore commesso nella fase di IE. Lo stesso errore, inoltre, ha un impatto diverso sui tre indici (−2 per mHFC, −1 per OPS e −2,0 per mC4P), in funzione dei contributi che ciascuno assegna alla classe MRA.

6.4 Analisi e discussione dei risultati

L’analisi comparativa condotta a partire dal dataset di 20 lettere cliniche evidenzia un certo grado di eterogeneità tra i modelli analizzati (GPT e Gemma via Ollama Cloud; Claude, Deepseek e Qwen via AWS) nelle capacità di estrazione ed elaborazione del testo clinico non strutturato.

Quasi tutti i modelli utilizzati, ovvero Claude, Deepseek, Gemma e GPT, hanno mostrato ottime performance, dimostrandosi capaci di interpretare correttamente la struttura delle lettere di dimissione, gestendo con precisione i formati numerici, le abbreviazioni farmacologiche e la sintassi tipica della manualistica medica italiana.
L’unico modello a discostarsi dalle performance di estrazione degli altri LLM è stato Qwen, per il quale è stato registrato il tasso più alto di errore. La dimensione più contenuta del modello (32B parametri) ha comportato errori ricorrenti sia nell’estrazione degli esami di laboratorio (sezione *lab_values* dello schema JSON) sia nell’estrapolazione delle informazioni legate alla terapia farmacologica del paziente (sezioni *current_therapy* e *discharge_therapy*). In quest’ultimo caso, sono stati commessi errori sia nel parsing numerico dei dosaggi, sia nell'associazione ontologica dei principi attivi.

Il confronto dei risultati dell’analisi delle performance tra lo step di IE e quello di calcolo degli score di aderenza conferma che il modulo deterministico di calcolo dei punteggi possiede un’affidabilità del 100%: quando le informazioni estratte dal documento clinico e memorizzate nello schema JSON sono corrette, lo score calcolato corrisponde al valore di riferimento. Pertanto, il vero collo di bottiglia risiede interamente nella fase di IE.

Gli errori generati dagli LLM nella fase di estrazione delle informazioni relative alla sezione *Discharge therapy*, elencate approfonditamente nel Sottocapitolo 6.1, si traducono in due distinti comportamenti anomali nel calcolo dell'aderenza, approfonditi nel Sottocapitolo 6.2: errori nel conteggio del numero di somministrazioni giornaliere e errori di classificazione dei farmaci.

In entrambi i casi l'errore può portare ad un aumento ingiustificato degli score di aderenza oppure ad una diminuzione rispetto al valore di riferimento. Nel primo caso, ciò è dovuto all'incremento del numero di dosi somministrate giornalmente rispetto a quanto riportato nel form dagli specialisti o all'associazione errata di un farmaco ad una classe appartenente alla quadruplice terapia; nel secondo caso invece il disallineamento è dovuto al decremento del numero di assunzioni o alla mancata o errata classificazione di un farmaco dal quale il calcolo degli score risulta dipendente.

In questo contesto emerge un legame tra la complessità dello score e la sua capacità di distinguere terapie farmacologiche più o meno complete. mHFC e OPS assegnano a ciascun farmaco contributi discreti (nell'OPS, per esempio, con due sole finestre di dose rispetto al target), per cui terapie diverse possono ottenere lo stesso punteggio. Il C4P, invece, è l'unico score con un contributo continuo per ogni farmaco, proporzionale alla percentuale della dose target, e ha l'intervallo di valori più ampio (da 5,5 a 15 sulle 20 lettere, contro 4–9 per mHFC e 2–8 per OPS). Assume inoltre un numero maggiore di valori distinti (15, contro 6 per mHFC e 7 per OPS) ed è quindi in grado di cogliere anche piccole differenze nella terapia, come un dosaggio inferiore al target: è lo score più sensibile (Tabella C.3 dell'Appendice C).
A una maggiore sensibilità si associa un numero maggiore di discrepanze rispetto ai valori di riferimento. Queste non dipendono dal modulo di calcolo, che è deterministico, ma dagli errori commessi dagli LLM nella fase di IE, che gli score più sensibili rilevano più facilmente. Su 100 combinazioni (20 lettere per 5 modelli), l'OPS, lo score meno granulare, presenta solo 4 discrepanze (Tabella C.2), l'mHFC 8 (Tabella C.1) e il C4P 13 (Tabella C.3), che comprendono tutte quelle rilevate dagli altri due score. Ciò conferma che gli score più sensibili sono più adatti a cogliere anche errori lievi sui dosaggi o sulla classificazione dei farmaci, e quindi a quantificare la

qualità della terapia prescritta.

Dal punto di vista del medico, in contesto assistenziale, queste tipologie di errore comportano i seguenti rischi decisionali:

- Una sottostima dello score potrebbe indurre il sistema a notificare al cardiologo un segnale di falso allarme di mancata aderenza, riducendo a lungo termine la fiducia del medico nel sistema di supporto, secondo il concetto di *alert fatigue* [53].
- Una sovrastima dello score invece potrebbe risultare ancora più pericolosa, in quanto maschererebbe una reale omissione terapeutica o un sottodosaggio, inducendo il medico a non effettuare modifiche all'attuale terapia e privando il paziente dell'opportunità di ottimizzare la quadruplice terapia.

Queste evidenze dimostrano che, nello stato attuale della tecnologia degli LLM generalisti, l'architettura non può funzionare in modo completamente autonomo. I risultati infatti convalidano la scelta progettuale di adottare un paradigma del tipo *Human-in-the-Loop*: il modello LLM deve fungere da assistente per una valutazione rapida dell'aderenza farmacologica, ma spetta al medico validare o correggere l'output ottenuto.

7. Conclusioni e sviluppi futuri

7.1 Sintesi del lavoro svolto

Il presente lavoro di tesi ha affrontato la difficoltà di valutare in modo sistematico l'aderenza alla quadruplice terapia dell'HFrEF, ostacolata dalla natura non strutturata delle lettere cliniche, sviluppando una pipeline articolata in tre fasi:

1. **Formalizzazione della Conoscenza Clinica**: il primo passaggio ha riguardato la formalizzazione delle linee guida ESC discorsive in workflow decisionali deterministici, ognuno dei quali riferito alla gestione di uno dei pilastri farmacologici della quadruplice terapia. Per risolvere i conflitti logici e le ambiguità sintattiche individuate nei documenti clinici, il flusso decisionale è stato perfezionato grazie ad un approccio del tipo *Expert-in-the-loop*, che ha permesso risolvere i dubbi relativi alla formalizzazione delle linee guida e a circoscrivere la modellizzazione ai soli elementi rilevanti nella pratica clinica.
2. **Estrazione delle informazioni da lettere cliniche**: Il secondo step consiste nella fase di IE automatizzata dei dati da lettere cliniche. In particolare, a partire dalle informazioni utili alla gestione della terapia farmacologica individuate grazie alla formalizzazione eseguita nello step precedente, è stato progettato uno schema JSON target mediante la libreria Pydantic. Sono stati utilizzati cinque LLM, sia *open-weight* sia proprietari, erogati tramite servizi cloud via API (Ollama Cloud e AWS Bedrock). Parallelamente, per ognuna delle 20 lettere usate come dataset, è stata richiesta ai medici dell'ICS Maugeri di Pavia la compilazione di un form di riferimento, con la medesima struttura dello schema utilizzato per l'estrazione. Ciò ha permesso di verificare la validità delle informazioni estrapolate dai modelli rispetto a quanto annotato dagli specialisti.
3. **Calcolo degli score di aderenza**: Il terzo e ultimo passaggio è costituito dal codice deterministico per il calcolo automatico degli score di aderenza. Dopo aver estratto le informazioni relative ai farmaci di riferimento per la patologia dalle linee guida ESC tramite un approccio multimodale, l'algoritmo valuta la presenza delle diverse classi farmacologiche e il raggiungimento del relativo dosaggio target, generando in output tre indici quantitativi di aderenza.

La pipeline è stata testata utilizzando cinque diversi modelli linguistici, valutati su un dataset di 20 lettere reali a confronto con i dati di riferimento registrati dai cardiologi nei form di annotazione. I risultati della fase di IE hanno confermato l'efficacia dell'approccio basato su output strutturato tramite Pydantic, evidenziando criticità residue isolate a campi specifici per alcuni dei modelli applicati (Tabella 6).

Per quanto riguarda invece il calcolo degli score di aderenza, affidato a un modulo deterministico che opera sulle informazioni estratte dagli LLM, le discrepanze rispetto ai valori di riferimento si sono verificate solamente quando i dati estratti erano errati (Tabella 7). Di conseguenza, gli score calcolati a partire dalle informazioni estratte da tutti i modelli, a eccezione di Qwen, hanno mostrato un'ottima corrispondenza con i valori di riferimento.
Nel complesso, nel lavoro di tesi è stata dimostrata la capacità della pipeline di gestire la forte variabilità linguistica della documentazione clinica italiana e l'impiego operativo.

La metodologia adottata nel presente lavoro di tesi offre un ulteriore vantaggio, ovvero una chiara suddivisione delle responsabilità tra la parte probabilistica dell'estrazione del testo (affidata agli LLM) e la parte deterministica responsabile del calcolo degli score di aderenza. Grazie a questo disaccoppiamento il sistema evita il rischio di errori o allucinazioni legati al calcolo diretto dei punteggi ad opera dei modelli e garantisce un maggiore grado di spiegabilità e riproducibilità dei risultati.
Inoltre, la struttura modulare facilita la manutenzione e la scalabilità del sistema: un eventuale aggiornamento futuro delle linee guida ESC o delle soglie di titolazione richiederebbe semplicemente la riprogrammazione delle modalità di calcolo del punteggio o l'aggiornamento delle informazioni relative ai farmaci di riferimento, lasciando completamente inalterata la fase di estrazione dei dati dai documenti clinici.
Infine, lo schema JSON costruito tramite Pydantic può essere riutilizzato come base per l'estensione dell'attuale framework ad altre patologie croniche.

7.2 Integrazione del lavoro nelle fasi successive del progetto GUIDE-AI

Il lavoro svolto nel corso di questa tesi si pone come un primo tassello nella realizzazione dell'obiettivo B (*Innovate and integrate*) del progetto quadriennale GUIDE-AI (Sottocapitolo 1.1). La modellazione dei workflow farmacologici, l'ingegnerizzazione dei prompt e lo sviluppo dei primi prototipi costituiscono il passaggio metodologico necessario per convertire modelli analitici isolati nell'architettura software scalabile e condivisa prevista dal consorzio.

In particolare, gli algoritmi di estrazione delle informazioni e gli output strutturati progettati nel corso di questo studio forniscono le basi per i *guideline navigators:* convertendo la lettera clinica in un profilo strutturato del paziente (terapia in corso, comorbidità, parametri clinici), producono l'input standardizzato che i *navigator*, basati su LLM e architetture RAG (Obiettivo B), potranno utilizzare per recuperare le indicazioni pertinenti delle linee guida e formulare le raccomandazioni terapeutiche. Inoltre, il corpus delle 20 lettere cliniche, corredato dai dati di riferimento compilati dai cardiologi e dai relativi score, costituisce un riferimento validato, che potrà essere utilizzato nelle fasi successive di *benchmarking* su dati retrospettivi: permette di quantificare le prestazioni del modulo di IE nel processare il testo clinico non strutturato (come fatto nel Sottocapitolo 6.1) e, una volta affiancato dalle valutazioni dei cardiologi sulla terapia attesa, anche la coerenza delle raccomandazioni generate con le linee guida.
Infine, le criticità emerse nella fase di estrazione (ad esempio, l'individuazione dei dosaggi o la distinzione tra comorbidità e altri disordini) e approfondite nel Sottocapitolo 6.1 costituiscono una raccolta di casistiche che potrà essere utilizzata per migliorare l'ingegnerizzazione dei prompt e come casi di test nel *benchmarking*, riducendo il rischio di errori e allucinazioni dei *guideline navigators* prima della loro estensione su scala reale.

In un'ottica di sviluppo all'interno del progetto, l'architettura software progettata e convalidata nel corso di questo lavoro di tesi presenta una marcata natura modulare, basata sul disaccoppiamento netto tra il livello semantico di estrazione dati e quello di calcolo deterministico degli score di aderenza. Questa separazione costituisce il

requisito fondamentale per garantire la scalabilità del sistema secondo le seguenti tre declinazioni:

- **Scalabilità multi-patologica**: nonostante il sistema sia stato ideato per il caso specifico dell'HFrEF, l'infrastruttura risulta predisposta per essere generalizzata ad altre patologie croniche ad alta complessità prescrittiva. L'adattamento a nuovi domini clinici non necessita della riscrittura dell'intero codice sorgente o dell'addestramento di nuovi modelli da zero, ma richiede solamente la ridefinizione dello schema target JSON e la formalizzazione dei relativi diagrammi di flusso decisionali derivati dalle linee guida di riferimento.
- **Scalabilità multilingue**: la scelta di condurre deliberatamente l'ingegnerizzazione dei prompt interamente in lingua inglese sfrutta le proprietà multilingue native degli LLM di ultima generazione, in modo tale da processare lettere di dimissione e verbali clinici redatti nelle diverse lingue nazionali dei partner europei del consorzio e mappare direttamente le informazioni nello schema standardizzato in inglese. In questo modo, gli step successivi di elaborazione dei dati si baseranno su una struttura unificata e indipendente dalla lingua di origine del testo.
- **Scalabilità multi-centro**: uno dei principali ostacoli all'adozione di sistemi di supporto alle decisioni nei consorzi sanitari internazionali consiste nella eterogeneità dei formati e degli stili adottati nelle redazioni dei documenti clinici. In questo scenario, risulta particolarmente efficace l'utilizzo di LLM come "traduttori semantici", capaci di interpretare la prosa clinica non strutturata e di convertire i testi in un formato strutturato e omogeneo.

Tale infrastruttura scalabile e rigorosa pone le basi per la fase successiva del progetto GUIDE-AI, rappresentata dall'Obiettivo C, *Evaluate in Real World*, che prevede il passaggio dalla validazione retrospettiva "in vitro" (condotta sul corpus di 20 lettere) alla sperimentazione prospettica "in vivo" all'interno della routine ambulatoriale e di reparto dei centri clinici partner, tra cui gli ICS Maugeri.

I *guideline navigator* dovranno seguire il paradigma Human-in-the-Loop, in cui la pipeline di IE e il motore decisionale operano non come agenti prescrittivi autonomi, ma come strumenti di supporto alle decisioni (CDSS). In questo scenario, nel corso di

una visita o della stesura di una lettera di dimissione, il sistema dovrà analizzare il testo non strutturato dato in input dal medico, estraendo i parametri rilevanti tramite lo schema JSON formalizzato e restituendo istantaneamente l'aderenza della terapia prescritta al paziente alle linee guida.
In questo scenario, il calcolo e la presentazione dei tre score di aderenza (*mHFC*, *OPS*, *modified C4P*) può risultare un'utile informazione aggiuntiva per quantificare velocemente la qualità della terapia attuale.
Le performance di tale sistema, in accordo con le direttive del progetto, verranno valutate da diversi indicatori, tra cui metriche di usabilità e accettabilità tecnologica e metriche di efficacia prescrittiva. Le prime valuteranno l'impatto sul carico cognitivo del clinico, i tempi di interazione in contesto assistenziale e il grado di fiducia espresso dai cardiologi, mentre le seconde quantificheranno la capacità del sistema di incrementare la percentuale di pazienti dimessi con una quadruplice terapia ottimizzata o correttamente titolata rispetto alla pratica clinica convenzionale.
In questo contesto, il contributo fornito da questo lavoro di tesi rappresenta il blocco concettuale necessario a garantire che lo studio di GUIDE-AI nella realtà clinica poggi su regole di sicurezza clinicamente verificate, riducendo al minimo il rischio di fallimento durante l'interazione diretta con i medici sul campo.

Infine, l'integrazione del presente lavoro di tesi nel progetto GUIDE-AI è completata dall'allineamento con l'obiettivo A del progetto (*People-Centricity*), secondo il quale il sistema sviluppato deve essere progettato sulla base dei reali bisogni di tutti gli attori coinvolti nel processo di cura. In questo contesto, lo schema dati strutturato, che associa a ciascun valore estratto la porzione di testo da cui deriva, e gli score di aderenza, indicatori quantitativi e interpretabili, rendono la terapia e il suo processo di valutazione trasparenti e ispezionabili: costituiscono quindi materiale su cui clinici e pazienti potranno essere coinvolti nelle attività di co-progettazione previste dall'obiettivo. Da un lato, la logica alla base dei flussi delle terapie farmacologiche è stata affinata con l'aiuto di un esperto del settore per assicurare l'aderenza con la pratica clinica (Sottocapitolo 3.3), dall'altro i pazienti potranno beneficiare di un percorso di titolazione più tempestivo, una minore incidenza di riospedalizzazioni per scompensamento acuto e una migliore qualità di vita complessiva.

7.3 Limiti del lavoro

Il lavoro presenta alcuni limiti che, al tempo stesso, indicano i punti della pipeline su cui è possibile intervenire.

- **Valutazione dell'estrazione**: L'accuratezza delle informazioni estratte è stata calcolata confrontando manualmente l'output dei modelli con il form compilato dai cardiologi, secondo criteri di equivalenza semantica (Sottocapitolo 6.1). Questo approccio, adeguato a un campione di 20 lettere, non è sostenibile nella fase di *benchmarking* su un numero molto maggiore di documenti. Un possibile miglioramento consiste nell'automatizzare il confronto con metriche standard dell'NLP: per i campi con valori vincolati dallo schema (booleani, *Enum*, valori numerici) il confronto diretto (*exact match*); per le sezioni a lista, *precision*, *recall* e *F1* calcolate a livello di entità, previa normalizzazione dei termini (ad esempio mappandoli su classificazioni standard); per i campi testuali, metriche basate su embedding [54] oppure l'uso di un LLM come valutatore, da validare su un sottoinsieme verificato manualmente [55].
- **Confrontabilità degli score**: I tre score di aderenza (mHFC, OPS e C4P) hanno scale e intervalli diversi e non sono direttamente confrontabili tra loro, né è immediato stabilire quando una differenza tra due punteggi sia clinicamente rilevante. Si potrebbe sviluppare, con il contributo di cardiologi esperti, una categorizzazione degli score in fasce di aderenza (ad esempio bassa, intermedia e alta), così da renderli confrontabili e da valutare se un errore di estrazione modifica la categoria assegnata al paziente e non soltanto il valore numerico.
- **Campione e generalizzabilità**: Il corpus è limitato a 20 lettere di un singolo centro, in lingua italiana: i risultati vanno quindi intesi come una prima verifica di fattibilità. Un *benchmarking* su un campione più ampio, multicentrico e annotato da più specialisti consentirebbe di stimare con maggiore robustezza le prestazioni e l'accordo tra gli annotatori.
- **Workflow decisionali**: La logica dei flussi terapeutici è stata affinata con un singolo esperto e, nel presente lavoro, non è stata implementata né valutata: la sua validazione clinica è demandata alle fasi successive del progetto.

- **Dati e infrastruttura**: La sperimentazione è stata condotta con modelli erogati in cloud su lettere pseudonimizzate. L'impiego su dati identificabili in contesto clinico reale richiederà modelli eseguiti su infrastrutture locali dell'ente, con un'eventuale riduzione delle prestazioni da valutare.

7.4 Sviluppi futuri: verso la generazione automatica di raccomandazioni terapeutiche

Per quanto riguarda gli sviluppi a breve termine, l'analisi del dataset di lettere cliniche nel contesto del lavoro di tesi ha messo in luce la grande variabilità di impostazioni stilistiche, acronimi e formulazioni utilizzabili nei testi ospedalieri (rilevata, ad esempio, per i dosaggi del Paziente 17, nel Sottocapitolo 6.1). Di conseguenza, è consigliabile l'integrazione di ulteriori testi, provenienti da altri centri e redatti in altre lingue, al fine di coprire la più vasta gamma di linguaggi utilizzati. In questo modo, infatti, utilizzando l'approccio *few-shots* in fase di ingegnerizzazione dei prompt, l'architettura di IE potrà raggiungere un grado di accuratezza maggiore in ambito multicentrico e multilingua.

Per quanto riguarda invece gli sviluppi a lungo termine, la naturale evoluzione dell'attività svolta nel presente lavoro di tesi è il passaggio da un CDSS analitico ad uno generativo. Sebbene l'architettura sviluppata sia in grado di estrarre con elevata accuratezza le informazioni strutturate di interesse per la gestione della terapia farmacologica e di determinarne l'aderenza rispetto alle linee guida ESC tramite gli score calcolati, il sistema rimane confinato a un ruolo di supporto passivo. Gli sviluppi futuri saranno indirizzati verso l'estensione ad un sistema di supporto proattivo, che non sia solamente in grado di analizzare i testi clinici, ma che sia capace di generare proattivamente raccomandazioni terapeutiche in linguaggio naturale, in base alle informazioni relative al paziente e alle linee guida aggiornate.

Dal punto di vista architetturale, questa transizione verso la generazione di raccomandazioni prescrittivi richiede l'adozione di metodologie in grado di minimizzare il rischio di allucinazioni o di obsolescenza rispetto alle versioni aggiornate delle linee guida. A tal fine, si renderà necessaria l'integrazione di sistemi

RAG, in cui il modello linguistico viene affiancato da un database dinamico contenente le versioni più aggiornate delle linee guida ESC. Parallelamente, l'architettura potrà essere estesa a sistemi multiagente, nei quali un modello generatore si dedica alla formulazione della proposta terapeutica sulla base del profilo del paziente, mentre un secondo agente validatore si occupa di verificare le eventuali controindicazioni, le interazioni farmacologiche e la sicurezza prescrittiva prima della presentazione al medico.

Un altro elemento di novità sarà l'estensione dall'analisi di un testo isolato (la lettera clinica) all'integrazione di flussi di dati multimodali. Gli sviluppi futuri prevedono infatti la capacità di processare non solo la prosa clinica, ma anche dati strutturati e segnali biomedici provenienti da referti ecocardiografici o tracciati elettrocardiografici digitali, al fine di estrarre parametri complessi, tra cui, per esempio, la frazione di eiezione precisa.

Infine, affinché questi sviluppi possano trovare una reale applicazione in contesto medico, dovranno essere rispettati i paradigmi di spiegabilità e conformità regolatoria. Ogni raccomandazione generata dal sistema dovrà essere accompagnata dalla porzione del testo clinico relativo al paziente o della linea guida volta a giustificare le scelte fatte dall'architettura, in modo tale da soddisfare i requisiti di sicurezza imposti dal Regolamento europeo sui Dispositivi Medici (MDR) e dall'*EU AI Act* per i software sanitari ad alto rischio, garantendo la massima affidabilità tecnologica e la tutela della salute del paziente.

8. Bibliografia

[1] M. Peleg *et al.*, «Comparing computer-interpretable guideline models: a case-study approach», *J. Am. Med. Inform. Assoc. JAMIA*, vol. 10, fasc. 1, pp. 52–68, 2003, doi: 10.1197/jamia.m1135.

[2] M. Peleg, «Computer-interpretable clinical guidelines: a methodological review», *J. Biomed. Inform.*, vol. 46, fasc. 4, pp. 744–763, ago. 2013, doi: 10.1016/j.jbi.2013.06.009.

[3] J. Clusmann *et al.*, «The future landscape of large language models in medicine», *Commun. Med.*, vol. 3, fasc. 1, p. 141, ott. 2023, doi: 10.1038/s43856-023-00370-1.

[4] «GUIDE-AI | IHI Innovative Health Initiative». Consultato: 19 agosto 2026. [Online]. Disponibile su: http://www.ihi.europa.eu/projects-results/project-factsheets/guide-ai

[5] K. K. Y. Ng, I. Matsuba, e P. C. Zhang, «RAG in Health Care: A Novel Framework for Improving Communication and Decision-Making by Addressing LLM Limitations», *NEJM AI*, vol. 2, fasc. 1, gen. 2025, doi: 10.1056/AIra2400380.

[6] P. Brennan, M. Perola, G.-J. van Ommen, E. Riboli, e European Cohort Consortium, «Chronic disease research in Europe and the need for integrated population cohorts», *Eur. J. Epidemiol.*, vol. 32, fasc. 9, pp. 741–749, set. 2017, doi: 10.1007/s10654-017-0315-2.

[7] T. A. McDonagh *et al.*, «2021 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure», *Eur. Heart J.*, vol. 42, fasc. 36, pp. 3599–3726, set. 2021, doi: 10.1093/eurheartj/ehab368.

[8] S. J. Greene *et al.*, «Medical Therapy for Heart Failure With Reduced Ejection Fraction: The CHAMP-HF Registry», *J. Am. Coll. Cardiol.*, vol. 72, fasc. 4, pp. 351–366, lug. 2018, doi: 10.1016/j.jacc.2018.04.070.

[9] G. Marzero, “tesi-hfref" GitHub, 2026. [Online]. Available: https://github.com/giacomomarzero/tesi-hfref. [Accessed: 14-Oct-2026].

[10] C. W. Tsao *et al.*, «Heart Disease and Stroke Statistics—2022 Update: A Report From the American Heart Association», *Circulation*, vol. 145, fasc. 8, feb. 2022, doi: 10.1161/CIR.0000000000001052.

[11] K. S. Shah *et al.*, «Heart Failure With Preserved, Borderline, and Reduced Ejection Fraction», *J. Am. Coll. Cardiol.*, vol. 70, fasc. 20, pp. 2476–2486, nov. 2017, doi: 10.1016/j.jacc.2017.08.074.

[12] T. B. Murdoch e A. S. Detsky, «The inevitable application of big data to health care», *JAMA*, vol. 309, fasc. 13, pp. 1351–1352, apr. 2013, doi: 10.1001/jama.2013.393.

[13] G. K. Savova *et al.*, «Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications», *J. Am. Med. Inform. Assoc. JAMIA*, vol. 17, fasc. 5, pp. 507–513, 2010, doi: 10.1136/jamia.2009.001560.

[14] W. W. Chapman, W. Bridewell, P. Hanbury, G. F. Cooper, e B. G. Buchanan, «A simple algorithm for identifying negated findings and diseases in discharge summaries», *J. Biomed. Inform.*, vol. 34, fasc. 5, pp. 301–310, ott. 2001, doi: 10.1006/jbin.2001.1029.

[15] A. Vaswani *et al.*, «Attention Is All You Need», 2017, *arXiv*. doi: 10.48550/ARXIV.1706.03762.
[16] J. Devlin, M.-W. Chang, K. Lee, e K. Toutanova, «BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding», in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota: Association for Computational Linguistics, 2019, pp. 4171–4186. doi: 10.18653/v1/N19-1423.
[17] J. Lee *et al.*, «BioBERT: a pre-trained biomedical language representation model for biomedical text mining», *Bioinformatics*, vol. 36, fasc. 4, pp. 1234–1240, feb. 2020, doi: 10.1093/bioinformatics/btz682.
[18] E. Alsentzer *et al.*, «Publicly Available Clinical BERT Embeddings», in *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, Minneapolis, Minnesota, USA: Association for Computational Linguistics, 2019, pp. 72–78. doi: 10.18653/v1/W19-1909.
[19] M. Agrawal, S. Hegselmann, H. Lang, Y. Kim, e D. Sontag, «Large language models are few-shot clinical information extractors», in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, 2022, pp. 1998–2022. doi: 10.18653/v1/2022.emnlp-main.130.
[20] T. B. Brown *et al.*, «Language Models are Few-Shot Learners», 2020, *arXiv*. doi: 10.48550/ARXIV.2005.14165.
[21] B. T. Willard e R. Louf, «Efficient Guided Generation for Large Language Models», 2023, *arXiv*. doi: 10.48550/ARXIV.2307.09702.
[22] D. Reichenpfader, H. Müller, e K. Denecke, «A scoping review of large language model based approaches for information extraction from radiology reports», *npj Digit. Med.*, vol. 7, fasc. 1, p. 222, 2024, doi: 10.1038/s41746-024-01219-0.
[23] L. C. Adams *et al.*, «Leveraging GPT-4 for Post Hoc Transformation of Free-Text Radiology Reports into Structured Reporting: A Multilingual Feasibility Study», *Radiology*, vol. 307, fasc. 4, p. e230725, 2023, doi: 10.1148/radiol.230725.
[24] P. Woźnicki *et al.*, «Automatic structuring of radiology reports with on-premise open-source large language models», *Eur. Radiol.*, vol. 35, fasc. 4, pp. 2018–2029, 2025, doi: 10.1007/s00330-024-11074-y.
[25] C. Niu *et al.*, «Development and Validation of a Large Language Model for Generating Fully-Structured Radiology Reports», 2024, *arXiv*. doi: 10.48550/ARXIV.2409.18319.
[26] P. Ponikowski *et al.*, «2016 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure: The Task Force for the diagnosis and treatment of acute and chronic heart failure of the European Society of Cardiology (ESC) Developed with the special contribution of the Heart Failure Association (HFA) of the ESC», *Eur. Heart J.*, vol. 37, fasc. 27, pp. 2129–2200, lug. 2016, doi: 10.1093/eurheartj/ehw128.
[27] T. A. McDonagh *et al.*, «2023 Focused Update of the 2021 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure», *Eur. Heart J.*, vol. 44, fasc. 37, pp. 3627–3639, ott. 2023, doi: 10.1093/eurheartj/ehad195.

[28] E. Gayat *et al.*, «Heart failure oral therapies at discharge are associated with better outcome in acute heart failure: a propensity-score matched study», *Eur. J. Heart Fail.*, vol. 20, fasc. 2, pp. 345–354, feb. 2018, doi: 10.1002/ejhf.932.
[29] M. G. Crespo-Leiro *et al.*, «European Society of Cardiology Heart Failure Long-Term Registry (ESC-HF-LT): 1-Year Follow-Up Outcomes and Differences Across Regions», *Eur. J. Heart Fail.*, vol. 18, fasc. 6, pp. 613–625, giu. 2016, doi: 10.1002/ejhf.566.
[30] J. J. V. McMurray *et al.*, «Dapagliflozin in Patients with Heart Failure and Reduced Ejection Fraction», *N. Engl. J. Med.*, vol. 381, fasc. 21, pp. 1995–2008, nov. 2019, doi: 10.1056/NEJMoa1911303.
[31] J. Butler *et al.*, «Empagliflozin and health-related quality of life outcomes in patients with heart failure with reduced ejection fraction: the EMPEROR-Reduced trial», *Eur. Heart J.*, vol. 42, fasc. 13, pp. 1203–1212, mar. 2021, doi: 10.1093/eurheartj/ehaa1007.
[32] CONSENSUS Trial Study Group, «Effects of enalapril on mortality in severe congestive heart failure. Results of the Cooperative North Scandinavian Enalapril Survival Study (CONSENSUS)», *N. Engl. J. Med.*, vol. 316, fasc. 23, pp. 1429–1435, giu. 1987, doi: 10.1056/NEJM198706043162301.
[33] R. Garg e S. Yusuf, «Overview of randomized trials of angiotensin-converting enzyme inhibitors on mortality and morbidity in patients with heart failure. Collaborative Group on ACE Inhibitor Trials», *JAMA*, vol. 273, fasc. 18, pp. 1450–1456, mag. 1995.
[34] M. Packer *et al.*, «Comparative effects of low and high doses of the angiotensin-converting enzyme inhibitor, lisinopril, on morbidity and mortality in chronic heart failure. ATLAS Study Group», *Circulation*, vol. 100, fasc. 23, pp. 2312–2318, dic. 1999, doi: 10.1161/01.cir.100.23.2312.
[35] SOLVD Investigators, S. Yusuf, B. Pitt, C. E. Davis, W. B. Hood, e J. N. Cohn, «Effect of enalapril on survival in patients with reduced left ventricular ejection fractions and congestive heart failure», *N. Engl. J. Med.*, vol. 325, fasc. 5, pp. 293–302, ago. 1991, doi: 10.1056/NEJM199108013250501.
[36] J. J. V. McMurray *et al.*, «Angiotensin-neprilysin inhibition versus enalapril in heart failure», *N. Engl. J. Med.*, vol. 371, fasc. 11, pp. 993–1004, set. 2014, doi: 10.1056/NEJMoa1409077.
[37] K. Damman *et al.*, «Renal Effects and Associated Outcomes During Angiotensin-Neprilysin Inhibition in Heart Failure», *JACC Heart Fail.*, vol. 6, fasc. 6, pp. 489–498, giu. 2018, doi: 10.1016/j.jchf.2018.02.004.
[38] A. S. Desai *et al.*, «Reduced Risk of Hyperkalemia During Treatment of Heart Failure With Mineralocorticoid Receptor Antagonists by Use of Sacubitril/Valsartan Compared With Enalapril: A Secondary Analysis of the PARADIGM-HF Trial», *JAMA Cardiol.*, vol. 2, fasc. 1, p. 79, gen. 2017, doi: 10.1001/jamacardio.2016.4733.
[39] J. P. Seferovic *et al.*, «Effect of sacubitril/valsartan versus enalapril on glycaemic control in patients with heart failure and diabetes: a post-hoc analysis from the PARADIGM-HF trial», *Lancet Diabetes Endocrinol.*, vol. 5, fasc. 5, pp. 333–340, mag. 2017, doi: 10.1016/S2213-8587(17)30087-6.
[40] C. B. Granger *et al.*, «Effects of candesartan in patients with chronic heart failure and reduced left-ventricular systolic function intolerant to angiotensin-

converting-enzyme inhibitors: the CHARM-Alternative trial», *Lancet*, vol. 362, fasc. 9386, pp. 772–776, set. 2003, doi: 10.1016/S0140-6736(03)14284-5.
[41] M. B. Fowler, «Effects of beta blockers on symptoms and functional capacity in heart failure», *Am. J. Cardiol.*, vol. 80, fasc. 11A, pp. 55L-58L, dic. 1997, doi: 10.1016/s0002-9149(97)00849-7.
[42] MERIT-HF Study Group, «Effect of metoprolol CR/XL in chronic heart failure: Metoprolol CR/XL Randomised Intervention Trial in Congestive Heart Failure (MERIT-HF)», *Lancet*, vol. 353, fasc. 9169, pp. 2001–2007, giu. 1999.
[43] M. Packer *et al.*, «The effect of carvedilol on morbidity and mortality in patients with chronic heart failure. U.S. Carvedilol Heart Failure Study Group», *N. Engl. J. Med.*, vol. 334, fasc. 21, pp. 1349–1355, mag. 1996, doi: 10.1056/NEJM199605233342101.
[44] B. Pitt *et al.*, «The effect of spironolactone on morbidity and mortality in patients with severe heart failure. Randomized Aldactone Evaluation Study Investigators», *N. Engl. J. Med.*, vol. 341, fasc. 10, pp. 709–717, set. 1999, doi: 10.1056/NEJM199909023411001.
[45] F. Zannad *et al.*, «Eplerenone in patients with systolic heart failure and mild symptoms», *N. Engl. J. Med.*, vol. 364, fasc. 1, pp. 11–21, gen. 2011, doi: 10.1056/NEJMoa1009492.
[46] M. Packer *et al.*, «Cardiovascular and Renal Outcomes with Empagliflozin in Heart Failure», *N. Engl. J. Med.*, vol. 383, fasc. 15, pp. 1413–1424, ott. 2020, doi: 10.1056/NEJMoa2022190.
[47] W. T. Abraham *et al.*, «Standardized Definitions for Evaluation of Heart Failure Therapies: Scientific Expert Panel from the Heart Failure Collaboratory and Academic Research Consortium», *Eur. J. Heart Fail.*, vol. 22, fasc. 12, pp. 2175–2186, dic. 2020, doi: 10.1002/ejhf.2018.
[48] A. Dwivedi *et al.*, «Heart Failure Guideline-Directed Medical Therapy Scoring Systems: A Scoping Review», *Circ. Heart Fail.*, vol. 19, fasc. 5, mag. 2026, doi: 10.1161/CIRCHEARTFAILURE.125.013881.
[49] M. Fiuzat *et al.*, «Optimal Background Pharmacological Therapy for Heart Failure Patients in Clinical Trials», *J. Am. Coll. Cardiol.*, vol. 79, fasc. 5, pp. 504–510, feb. 2022, doi: 10.1016/j.jacc.2021.11.033.
[50] A. Verma *et al.*, «DASH-HF Study: A Pragmatic Quality Improvement Randomized Implementation Trial for Patients With Heart Failure With Reduced Ejection Fraction», *Circ. Heart Fail.*, vol. 16, fasc. 9, set. 2023, doi: 10.1161/CIRCHEARTFAILURE.122.010278.
[51] A. Steverson *et al.*, «Validating the Association Between Composite Metrics of Guideline-Directed Medical Therapy and Clinical Outcomes for Patients With Heart Failure With Reduced Ejection Fraction», *J. Card. Fail.*, vol. 31, fasc. 3, pp. 588–591, mar. 2025, doi: 10.1016/j.cardfail.2024.08.054.
[52] G. Navarro, «A guided tour to approximate string matching», *ACM Comput. Surv.*, vol. 33, fasc. 1, pp. 31–88, 2001, doi: 10.1145/375360.375365.
[53] C. E. Ray *et al.*, «Alert fatigue measurement in clinical decision support: a systematic review», *J. Am. Med. Inform. Assoc. JAMIA*, vol. 33, fasc. 8, pp. 1523–1531, ago. 2026, doi: 10.1093/jamia/ocag064.
[54] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, e Y. Artzi, «BERTScore: Evaluating Text Generation with BERT», 2020, *arXiv*. doi: 10.48550/ARXIV.1904.09675.

[55] L. Zheng *et al.*, «Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena», 2023, *arXiv*. doi: 10.48550/ARXIV.2306.05685.

Appendice A: Form di annotazione di riferimento

Questa appendice riporta integralmente il form di annotazione di esempio fornito agli specialisti del settore e compilato per ognuna delle 20 lettere cliniche usate come dataset nel corso del presente lavoro di tesi e introdotto nel Capitolo 4 (Sottocapitolo 4.2).

Tabella A.1

Form di annotazione di esempio reso disponibile ai medici per la compilazione.

CLINICAL LETTER ANNOTATION FORM				
Annotator ID:	Maria Rossi	Letter ID:	example_cardiologia.docx	
A. DEMOGRAPHIC INFO				
Demographic characteristic		Value		Notes
sex		female		
age		58 years		
B. DRUG INTOLERANCE | List all known drug intolerances				
Drug / Substance				Notes
Dapagliflozin, Empagliflozin, Ramipril				
C. LAB VALUES | Exam name, value, unit and date				
Lab Exam	Value	Unit	Date	Notes
QRS	148	ms	24/11/2025	
LVEF	25	%	22/12/2025	
Hemoglobin	10.8	g/dL	22/12/2025	
D. COMORBIDITIES				
Comorbidities				Notes
Hypertension, mild anemia, atrial fibrillation				
E. DISORDERS AND CONDITIONS				
Disorder / Condition				Notes
dyspnea				
F. CLINICAL CRITERIA | Select the answer from the dropdown				
Variable	Definition / Clinical Hint		Answer	Notes

CLINICAL LETTER ANNOTATION FORM					
Worsening Heart Failure		Hospitalization OR unplanned outpatient visit for HF (and treatment with either intravenous diuretics, vasodilators or inotropes) within the last 6 months		VERO	
Loop Diuretic		Patient currently treated with a loop diuretic		VERO	
Peripheral Edema		Peripheral edema reported by patient or found during physical examination		VERO	
HFrEF 90 days		A diagnosis of HFrEF (LVEF<=40% and signs and symptoms of HF such as dyspnea NYHA II-IV) at least 3 months prior to enrollment.		VERO	
Comfort Care Only		Patient is currently receiving comfort care measures only or is enrolled in a hospice program.		FALSO	
Critically ill – HFrEF		Hospital admission for cardiogenic shock or any use of inotropes or vasopressors within the last 30 days?		FALSO	
HFrEF Etiology		Etiology of HFrEF		Ischemic	
Discharge Date (Hospital)		Date of hospital discharge			not reported
G. CURRENT THERAPY | add one row per drug					
Drug Name (patient letter)	Generic name	Brand name	Daily dose	Drug class	Notes
Lasix* 25mg (Furosemide 25mg)	Furosemide	Lasix	1 tablet	Loop diuretic	
Entresto* 97mg+103mg	Sacubitril/Valsartan	Entresto	2 tablet	ARNI	

CLINICAL LETTER ANNOTATION FORM					
(Sacubitril/valsartan)					
Cardioaspirin* 100mg (Acetylsalicylic acid 100mg)	Acetylsalicylic acid	Cardioaspirin	1 tablet	Other	
H. DISCHARGE THERAPY | add one row per drug					
Drug Name (patient letter)	Generic name	Brand name	Daily dose	Drug class	Notes
Lasix* 25mg (Furosemide 25mg)	Furosemide	Lasix	2 tablet	Loop diuretic	Dose changed
Entresto* 97mg+103mg (Sacubitril/valsartan)	Sacubitril/Valsartan	Entresto	2 tablet	ARNI	

Appendice B: prompt usati nella fase di estrazione

Questa appendice contiene i prompt utilizzati nella fase di estrazione delle informazioni dalle lettere cliniche. Tra questi figurano il prompt di sistema e le descrizioni di campo di ciascuna sezione dello schema di estrazione.

B.1 Prompt di sistema

Il prompt di sistema definisce il ruolo del modello e le regole generali di estrazione, tra cui lo schema di riferimento, la lingua o la gestione dei dati mancanti.

```
You are a clinical data extraction assistant.
The patient has been diagnosed with HFrEF (Heart Failure with reduced Ejection Fraction).
Given a clinical patient note, extract all relevant information and return it as a valid JSON object STRICTLY CONFORMING to the schema provided.
The output schema is:
$output_schema

Rules:
- For each field of the JSON output, return the corresponding value with the original span from the source document from which the value was extracted.
- If a field is not mentioned or cannot be inferred, set it to null.
- For enum fields, map clinical language to the closest valid enum value.
- All numeric values must respect the stated units and ranges.
- IMPORTANT: ALL extracted values must be returned in ENGLISH.
- Do not return information that is not included in the provided JSON schema.
- Return ONLY the JSON object, no preamble, no explanation, no markdown fences.
- Strictly follow the provided JSON schema.

Temporal rules:
- Always extract the most recent information available. Clinical notes may span multiple hospital visits;
  if a value (e.g. FEV1) is not explicitly updated in the current visit, it may have been carried over
  from a previous visit — treat it as valid but be aware it may not reflect the current status.
- Pay close attention to diagnosis sublists: items listed there may persist across visits without being
  re-stated explicitly. Extract them, but prefer any updated value if present.
```

B.2 Prompt per le informazioni demografiche

Il primo campo dello schema contiene le informazioni demografiche del paziente (età, sesso, etnia ecc.).

```
Demographic information explicitly reported in the clinical document.
Extract relevant patient demographic characteristics such as age, sex, ethnicity, race, or
```

other demographic descriptors when available.
Return each demographic item as a separate string.
Set to null if no demographic information is reported.

B.3 Prompt per le intolleranze

L'identificazione delle intolleranze farmacologiche del paziente nella sezione *drug_intolerances* costituisce un'informazione fondamentale in fase di definizione della terapia.

Drug intolerances explicitly mentioned in the clinical letter.
Extract all medications reported as not tolerated by the patient.
Return each drug intolerance as a separate string using concise medical terminology.
Do not infer information that is not explicitly stated.
Avoid duplicates and set to null if no drug intolerances are reported.

B.4 Prompt per gli esami di laboratorio

La sezione *lab_values* si occupa dell'estrazione e strutturazione dei parametri fisiologici. La pipeline utilizza un prompt principale di sezione affiancato da regole di estrazione specifiche per ciascun attributo del modello Pydantic.

B.4.1 Prompt di sezione

Il seguente prompt definisce il contesto di estrazione per l'intero blocco dei valori di laboratorio:

Extract all medical measurements, vital signs, and laboratory results reported in the clinical document.
Specifically, you must include laboratory tests (such as creatinine, BNP) and physiological parameters like Heart Rate (HR), Left Ventricular Ejection Fraction (LVEF), Systolic Blood Pressure (SBP), and Diastolic Blood Pressure (DBP).
For each measurement, extract the test or parameter name, its numerical value, measurement unit, and associated date/time when available.
Return one LabEntry object per result.
CRITICAL RULE: If multiple measurements or values are reported for the same test or parameter across different times or dates, extract ONLY the latest (most recent) measurement available in the text.
Do not infer missing values or timestamps.
Set to null if no matching data are reported.

B.4.2 Specifica delle istruzioni sui singoli campi

Per garantire l'omogeneità dei dati e il corretto casting nei tipi di dato definiti nello schema JSON, per ciascun attributo vengono applicate le seguenti istruzioni di parsing:

- Campo *lab_exam* (esame di laboratorio):

```
Name of the laboratory test or examination exactly as reported in the clinical document (e.g. creatinine, potassium).
Always populate this field when a laboratory value is extracted.
```

- Campo *lab_value* (valore):

```
Laboratory result including numerical value only.
Return the value exactly as reported in the source text.
Set to null if no numerical value is reported.
```

- Campo *lab_unit* (unità di misura):

```
Measurement unit associated with the laboratory result.
Return the unit exactly as reported in the source text.
Set to null if no unit is reported.
```

- Campo *lab_date* (data):

```
Date associated with the laboratory result.
Use dd.mm.yyyy format when possible.
Set to null if no timestamp is reported.
```

- Campo *notes* (note):

```
Extra information about the laboratory result. Set to null if not reported.
```

B.5 Prompt per le comorbidità

La sezione *comorbidities* contiene le diagnosi concomitanti afferenti all'HFrEF.

```
List of Heart Failure comorbidities explicitly mentioned in the clinical letter.
Heart Failure (HF) is the target disease of this study and MUST NEVER be returned in this field.
Exclude any diagnosis, condition, syndrome, phenotype, subtype, stage, severity descriptor, or clinical expression that represents, contains, modifies, qualifies, or refers to Heart Failure.
CRITICAL: If a candidate extraction contains the words 'heart failure', 'cardiac failure', 'compensation', 'decompensation', or an HF abbreviation (HF, HFrEF, HFmrEF, HFpEF, HFimpEF), discard it.
Return each comorbidity as a separate string.
```

Do not duplicate the same underlying disease using different clinical phenotypes, stages, subtypes, or temporal evolutions.
If a diagnosis represents an improved, recovered, chronic, acute, decompensated, compensated, or otherwise qualified form of the same disease, return only the diagnosis exactly as reported in the letter and do not add the parent or previous condition unless it is also explicitly listed as a separate active diagnosis.
Do not infer conditions that are not explicitly stated.
Set to null if no comorbidities are reported.

B.6 Prompt per disturbi e condizioni mediche

La sezione *disorders_and_conditions* si riferisce a diagnosi, sintomi o condizioni patologiche non classificabili come comorbilità.

Extract only clinically recognized diagnoses, diseases, syndromes, pathological conditions, or symptoms explicitly reported in the letter that do NOT qualify as comorbidities.
AVOID findings already classified as comorbidities.
Do not extract acute, transient, or resolved events that are clearly described as no longer clinically relevant, unless they are explicitly reported as part of the patient's current comorbidity profile.
Heart Failure (HF) is the target disease of this study and MUST NEVER be returned in this field.
Exclude any diagnosis, condition, syndrome, phenotype, subtype, stage, severity descriptor, or clinical expression that represents, contains, modifies, qualifies, or refers to Heart Failure (HF).
CRITICAL: If a candidate extraction contains the words 'heart failure', 'cardiac failure', 'compensation', 'decompensation', or an HF abbreviation (HF, HFrEF, HFmrEF, HFpEF, HFimpEF), DISCARD it.
Do NOT extract quantitative or functional cardiac measurements, indices, or parameters used to describe disease severity or cardiac function, including but not limited to: left ventricular ejection fraction (LVEF / EF / FE), ventricular volumes, wall motion indices, strain measures, or similar cardiological functional metrics.
Return each disorder or condition as a separate string using concise medical terminology.
Do not infer information that is not explicitly stated.
Avoid duplicates and set to null if no disorders or conditions are reported.

B.7 Prompt per l'estrazione di variabili binarie

Lo schema JSON utilizzato in fase di estrazione prevede una serie di sezioni binarie, volti a valutare diverse condizioni del paziente utili alla gestione della terapia farmacologica:

- Sezione *worsening_heart_failure* (insufficienza cardiaca peggiorata):

Whether the patient is currently experiencing worsening heart failure.
Return True if the patient has been hospitalized or had an unplanned outpatient visit for heart failure (and treatment with either intravenous diuretics, vasodilators or inotropes) within the last 6 months.
Return True if explicitly reported.
Return False if explicitly denied.

Set to False if the information is not available.

- Sezione *loop_diuretic* (diuretici dell'ansa):

Whether the patient is treated with a loop diuretic BEFORE the time of the visit.
Consider ONLY drugs the patient was ALREADY taking at the time of the visit.
Do NOT consider the medication regimen prescribed after the visit.
Return True if loop diuretic therapy is explicitly reported.
Return False if explicitly denied.
Set to False if the information is not available.

- Sezione *peripheral_edema* (edema periferico):

Presence of peripheral edema reported by the patient or identified during physical examination.
Return True only if edema is explicitly described.
Return False if explicitly absent.
Set to False if not reported.

- Sezione *HFrEF_90_days* (diagnosi di HFrEF negli ultimi 90 giorni):

Whether the patient had a diagnosis of heart failure with reduced ejection fraction (HFrEF) at least 3 months prior to enrollment or evaluation.
HFrEF is defined as left ventricular ejection fraction (LVEF) less than or equal to 40% associated with signs and symptoms of heart failure such as dyspnea or NYHA class II-IV.
Return True only if this condition is explicitly documented.
Return False if explicitly denied.
Set to null if the information is unavailable.

- Sezione *comfort_care_only* (sole cure palliative):

Whether the patient is currently receiving comfort care only.
Return True if the clinical document explicitly states that the patient is under comfort care, palliative care, hospice care, or has a do-not-resuscitate (DNR) order indicating a focus on comfort measures rather than curative treatment.
Return False if explicitly denied or if the patient is enrolled in a hospice program.
Set to False if the information is not available.

- Sezione *HFrEF_etiology* (eziologia dell'HFrEF):

Etiology or underlying cause of heart failure with reduced ejection fraction (HFrEF) as explicitly reported in the clinical document.
Select one from the predefined list: ISCHEMIC / NON_ISCHEMIC
Set to null if the etiology is not reported.

B.8 Prompt per la data di dimissione

La sezione discharge_date è dedicata all'estrazione della data di dimissione, la cui presenza nel documento identifica il documento come lettera di dimissione ospedaliera.

```
Date of the most recent hospital discharge related to heart failure.
Use dd.mm.yyyy format when possible.
Set to null if no discharge date is reported or if the patient has not been hospitalized for heart failure.
```

B.9 Prompt per le terapie farmacologiche

Le sezioni relative alle terapie farmacologiche contengono le informazioni relative alle prescrizioni famacologiche antecedenti (current_therapy) e posteriori alla visita (discharge_therapy). È stato utilizzato un prompt principale di sezione affiancato da regole di estrazione specifiche per ciascun attributo dei farmaci estratti.

B.9.1 Prompt di sezione per la terapia in ingresso

La sezione *current_therapy* è dedicata alle prescrizioni farmacologiche antecedenti alla visita alla quale è relativa la lettera clinica.

```
Current pharmacological therapies reported in the clinical letter.
List all drugs the patient was already taking at the time of the visit.
Extract medications that the patient is currently taking or receiving at the time of the clinical assessment, visit, or hospital admission.
Include chronic home medications and ongoing treatments explicitly described as current therapy.
Exclude discharge medications unless they are explicitly stated to be part of the current regimen at the time of assessment.
For each therapy include drug name, dosage information, dose modifications, uptitration status, relevant dates, and additional notes when available.
Return one TherapyEntry object per medication.
Set to null if no current therapy is reported.
```

B.9.2 Prompt di sezione per la terapia in uscita

La sezione *current_therapy* è dedicata alla posologia aggiornata al termine della visita. Il calcolo degli score di aderenza si basa sui dati riportati in questa parte dello schema.

```
Extract the COMPLETE discharge medication regimen the patient must follow after the current visit.
CRITICAL: You must include BOTH newly prescribed medications AND pre-existing/chronic medications that the patient is instructed to continue taking.
```

Scan the entire text (including 'Medical History', 'Anamnesis', 'Home Therapy', and 'Discharge') to build the full list.
For each therapy include drug name, dosage information, dose modifications, uptitration status, start dates if available, and additional notes when available.
For each therapy include drug name, dosage information, dose modifications, uptitration status, relevant dates, and additional notes when available.
Return one TherapyEntry object per medication.
Set to null ONLY if the text explicitly states no medications are prescribed at discharge. "

B.9.3 Specifica delle istruzioni sui singoli campi

Per garantire omogeneità nel formato delle informazioni estratte relative alla terapia all'ingresso (*current_therapy*) e a quella alla dimissione (*discharge_therapy*), sono stati definiti i seguenti prompt, che descrivono in modo specifico ogni campo associato all'estrazione di un farmaco:

- Campo *drug_name* (nome del farmaco):

Drug name as reported in the source, exactly as written.
Always populate this field.

- Campo *generic_drug_name* (nome generico del farmaco):

Drug generic name or ATC code, only if clearly identifiable as such.
Return the output translated in ENGLISH.
Exclude the dose from the output.
Set to null if not reported or not distinguishable.

- Campo *brand_drug_name* (nome commerciale del farmaco):

Drug brand name, only if clearly identifiable as such.
Set to null if not reported or not distinguishable.

- Campo *active_ingredient* (principio attivo):

Extract the dose strength and its mass/activity unit of measurement (e.g., mg, mcg, µg, UI) for the medication, as explicitly reported in the text.
Formatting rules:
- For single active ingredients: Output MUST be strictly in the format <number> <unit> (e.g., 20 mg, 50 mcg, 1000 UI).
- For combination drugs (e.g., 97mg+103mg or 97mg/103mg): Join the numbers with a slash and follow with a single space and unit, formatted as <number>/<number> <unit> (e.g., 97/103 mg). If units differ, maintain their respective units (e.g., 50 mcg/10 mg).
Ignore pharmaceutical forms, tablet/pill counts, administration timing, and frequencies.

- Campo *dose_units_number* (numero di unità di dose):

Calculate ONLY the total count of single physical dose units (e.g., tablets, drops, vials, puffs) to be taken across the ENTIRE 24-HOUR DAY, expressed as a single decimal or integer number.
Rules:
MULTIPLE TIMESTAMPS FOR ONE DOSE: If a dose is linked to multiple hours separated by hyphens, commas, or slashes (e.g., '1 cpr at 8-20', '1/2 cpr hour 8,14,20'), multiply the unit dose by the number of listed hours (e.g., 1 tablet at 8 and 20 = 1 * 2 = 2).
- SUM MULTIPLE EVENTS: If multiple individual doses, timestamps, or times of day are listed (e.g., '1 tablet at 8:00 and 1 tablet at 20:00', '1 tab morning and 1 tab night'), SUM all individual physical units together (1 + 1 = 2).
- FREQUENCY MULTIPLICATION: If an explicit daily frequency is stated (e.g., '1/2 tablet twice a day', '1 puff 3 times daily'), multiply the unit dose by the frequency (0.5 * 2 = 1).
- TIMESTAMPS VS COUNTS: Do NOT use the hour value as a dose (e.g., in 'at 14:00', 14 is the hour, NOT 14 tablets). However, DO count the physical dose associated with that time.
- CONVERSIONS: Convert words, fractions, and partial units to decimals (e.g., 1/2 -> 0.5, half -> 0.5, one -> 1, two -> 2).
- IGNORE DOSAGE STRENGTHS: Ignore mass/volume units (e.g., mg, mcg, ml) and focus ONLY on physical unit counts.
- SINGLE EVENT: If only a single administration is described for the day with no other times mentioned (e.g., 1/2 tablet at 14), the daily total is that single dose (0.5).
Examples:
- '1 cpr at 8-20' -> 2
- '1/2 cpr at 8-14-20' -> 1.5
- '1/2 tablet twice a day' -> 1
- '1 tablet morning, 1/2 tablet at 14:00, 1 tablet at night' -> 2.5
- 'Bisoprolol 5mg 1/2 tablet at 8:00' -> 0.5
Output MUST be ONLY the single calculated number or null. No text, explanation, or unit labels.

- Campo *drug_class* (classe farmacologica):

Analyze the drug name following these steps:
1. Identify the active ingredient.
2. Assign the corresponding class from the predefined list: ACE_I / ARB / ARNI / BETA_BLOCKER / MRA / SGLT2_INHIBITOR / LOOP_DIURETIC.
Examples:
- ACE_I: enalapril, ramipril, lisinopril
- ARB: losartan, valsartan, candesartan
- ARNI: sacubitril/valsartan
- BETA_BLOCKER: bisoprolol, carvedilol, metoprolol
- MRA: spironolactone, eplerenone, canrenone (and its potassium salt, canrenoate)
- SGLT2_INHIBITOR: dapagliflozin, empagliflozin" "- LOOP_DIURETIC: furosemide, torasemide, bumetanide
3. If the active ingredient does not match any list, set to null.

- Campo *notes* (note):

Extra information about the specified drug.
Specify any relevant details such as dose modifications, uptitration status, dates of therapy

initiation or changes, or other clinically relevant information. Set to null if not reported.

Appendice C: Tabelle integrali degli score di aderenza

In questa appendice sono raccolti gli score di aderenza calcolati a partire dai dati estratti per la terapia di dimissione dalle 20 lettere cliniche dai 5 modelli utilizzati nella fase di IE. In ogni tabella sono annotati gli i punteggi ottenuti dal calcolo di uno dei tre score selezionati nel corso del presente lavoro di tesi.

Tabella C.1

Punteggi calcolati per lo score mHFC per ogni lettera clinica (sulle righe) e per ogni modello (sulle colonne).

ID Lettera	DeepSeek	Gemma	GPT	Qwen	Claude	Riferimento
L01	8.0	8.0	8.0	8.0	8.0	8.0
L02	7.0	7.0	7.0	6.0	7.0	7.0
L03	9.0	9.0	9.0	8.0	9.0	9.0
L04	6.0	6.0	6.0	6.0	6.0	6.0
L05	8.0	8.0	8.0	8.0	8.0	8.0
L06	4.0	4.0	4.0	4.0	4.0	4.0
L07	6.0	6.0	6.0	4.0	6.0	6.0
L08	8.0	8.0	8.0	8.0	8.0	8.0
L09	7.0	7.0	7.0	7.0	7.0	7.0
L10	9.0	9.0	9.0	8.0	9.0	9.0
L11	8.0	8.0	8.0	8.0	8.0	8.0
L12	8.0	8.0	8.0	8.0	8.0	8.0
L13	7.0	7.0	7.0	5.0	7.0	7.0
L14	8.0	8.0	8.0	8.0	8.0	8.0
L15	6.0	6.0	6.0	6.0	6.0	6.0

ID Lettera	DeepSeek	Gemma	GPT	Qwen	Claude	Riferimento
L16	7.0	7.0	7.0	6.0	7.0	7.0
L17	8.0	8.0	6.0	7.0	8.0	8.0
L18	4.0	4.0	4.0	4.0	4.0	4.0
L19	5.0	5.0	5.0	5.0	5.0	5.0
L20	6.0	6.0	6.0	6.0	6.0	6.0

Tabella C.2

Punteggi calcolati per lo score OPS per ogni lettera clinica (sulle righe) e per ogni modello (sulle colonne).

ID Lettera	DeepSeek	Gemma	GPT	Qwen	Claude	Riferimento
L01	8.0	8.0	8.0	6.0	8.0	8.0
L02	5.0	5.0	5.0	5.0	5.0	5.0
L03	8.0	8.0	8.0	6.0	8.0	8.0
L04	5.0	5.0	5.0	5.0	5.0	5.0
L05	8.0	8.0	8.0	8.0	8.0	8.0
L06	3.0	3.0	3.0	3.0	3.0	3.0
L07	4.0	4.0	4.0	4.0	4.0	4.0
L08	5.0	5.0	5.0	5.0	5.0	5.0
L09	5.0	5.0	5.0	5.0	5.0	5.0
L10	6.0	6.0	6.0	6.0	6.0	6.0
L11	6.0	6.0	6.0	6.0	6.0	6.0

ID Lettera	DeepSeek	Gemma	GPT	Qwen	Claude	Riferimento
L12	6.0	6.0	6.0	6.0	6.0	6.0
L13	5.0	5.0	5.0	4.0	5.0	5.0
L14	6.0	6.0	6.0	6.0	6.0	6.0
L15	4.0	4.0	4.0	4.0	4.0	4.0
L16	5.0	5.0	5.0	5.0	5.0	5.0
L17	5.0	5.0	4.0	5.0	5.0	5.0
L18	2.0	2.0	2.0	2.0	2.0	2.0
L19	4.0	4.0	4.0	4.0	4.0	4.0
L20	7.0	7.0	7.0	7.0	7.0	7.0

Tabella C.3

Punteggi calcolati per lo score *modified* C4P per ogni lettera clinica (sulle righe) e per ogni modello (sulle colonne).

ID Lettera	DeepSeek	Gemma	GPT	Qwen	Claude	Riferimento
L01	14.8	14.8	14.8	13.5	14.8	14.8
L02	12.5	12.5	12.5	12.0	12.5	12.5
L03	15.0	15.0	15.0	12.8	15.0	15.0
L04	10.2	10.2	10.2	10.2	9.2	10.2
L05	14.8	14.8	14.8	14.8	14.8	14.8
L06	7.0	7.0	7.0	7.0	7.0	7.0

ID Lettera	DeepSeek	Gemma	GPT	Qwen	Claude	Riferimento
L07	10.0	10.0	10.0	9.0	10.0	10.0
L08	13.0	13.0	13.0	13.0	13.0	13.0
L09	12.5	12.5	12.5	12.5	12.5	12.5
L10	13.5	13.5	13.5	10.2	13.5	13.5
L11	13.5	13.5	13.5	13.5	13.5	13.5
L12	13.0	13.0	13.0	13.0	13.0	13.0
L13	11.8	11.8	11.8	9.8	11.8	11.8
L14	13.5	13.5	13.5	13.5	13.5	13.5
L15	9.0	9.0	9.0	9.0	9.0	9.0
L16	11.5	11.5	11.5	11.0	11.5	11.5
L17	10.0	10.0	7.0	11.1	10.0	11
L18	5.5	5.5	5.5	5.5	5.5	5.5
L19	8.5	8.5	8.5	8.5	8.5	8.5
L20	12.8	12.8	12.8	12.8	12.8	12.8