UNIVERSITÀ
DI PAVIA

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

CORSO DI LAUREA MAGISTRALE IN BIOINGEGNERIA

TESI DI LAUREA

STUDIO DI FATTIBILITÀ PER L'ANALISI AUTOMATIZZATA DEL RISCHIO NELLE PROCEDURE OPERATIVE STANDARD PER LA TERAPIA CAR-T

Candidato: Umberto Mattioli

Relatore: Prof. Riccardo Bellazzi
Correlatrice: Laura Bergomi

A.A. 2025/2026

Indice

- Appendice D – Esempio di prompt utilizzato nella procedura di astrazione
- Appendice E – System prompt utilizzato nell'approccio LLM-as-a-judge

8. Bibliografia

Abstract

La terapia CAR-T è caratterizzata da un complesso processo clinico e conseguentemente anche organizzativo; infatti, è articolato in numerose fasi e deve garantire elevati standard di tracciabilità e di integrità del prodotto cellulare; dunque, è necessaria un'adeguata analisi del rischio.

In questo contesto l'analisi della documentazione procedurale rappresenta un passaggio fondamentale per individuare le possibili criticità del percorso, ma richiede un'attività manuale di lettura, organizzazione e sintesi delle informazioni molto significativa.

La presente tesi propone uno studio di fattibilità per lo sviluppo di una pipeline semi-automatica finalizzata a supportare l'analisi del rischio nel processo CAR-T. Il lavoro si sofferma sulle prime fasi del percorso e prevede l'associazione dei documenti procedurali alle relative fasi, l'estrazione strutturata delle informazioni di rischio e una successiva fase di astrazione e aggregazione delle informazioni semanticamente simili, con l'obiettivo di ridurre le ridondanze e ottenere una rappresentazione più sintetica e generale.

La qualità degli output è stata valutata mediante un confronto con un riferimento fornito da esperti del settore. A tale scopo sono state utilizzate sia metriche tradizionali per le informazioni testuali, sia approcci di confronto semantico basati su embedding e LLM-as-a-judge,

Nel complesso il lavoro mostra ed evidenzia come l'impiego di modelli linguistici possa rappresentare un supporto promettente per rendere più sistematica e tracciabile l'organizzazione delle informazioni di rischio.

La pipeline proposta non ha come obiettivo la sostituzione del lavoro e del giudizio degli esperti, ma intende costituire uno strumento di supporto, la cui applicabilità effettiva dovrà essere ulteriormente valutata attraverso dataset più ampi e da una validazione da parte di esperti nel dominio.

Abstract in English

CAR-T therapy is characterized by a complex clinical and organizational process; in fact, it is structured into numerous phases and must ensure high standards of cellular product traceability and integrity, therefore adequate risk analysis is required.

In this context, the analysis of the procedural documentation represents a fundamental step in order to identify critical issues in the process.

For this reason, a significant amount of work in terms of reading, organizing, and information synthesis is required.

This thesis presents a study on the feasibility of the development of a semi-automatic pipeline aimed at supporting risk analysis in the CAR-T process.

The work focuses on the early stages of the process and involves mapping procedural documents to their corresponding phases, structural extraction of risks, related information and a subsequent phase of abstraction, aggregation of semantically similar information with the aim of reducing redundancies and obtaining a more synthetic and general representation.

The output quality has been evaluated through comparison with a reference standard provided by domain experts.

For this purpose, both traditional metrics for textual information and semantic comparison approaches based on embedding and LLM-as-a-judge have been used.

Overall, the work shows and highlights how the use of Language Models can represent valid support in order to make the organization of risk information more systematic and traceable.

The proposed pipeline is not intended to replace the work and judgement of the experts, but rather to serve as a support tool whose actual applicability will need to be further evaluated through wider datasets and domain experts' validation.

1. Introduzione

1.1 Terapia CAR-T

La terapia CAR-T, acronimo di *Chimeric Antigen Receptor T-cell therapy*, rappresenta una delle innovazioni più importanti nell'ambito dell'immunoterapia cellulare. La differenza principale di questa terapia rispetto alle altre terapie farmacologiche tradizionali è che questo trattamento non si basa semplicemente sulla somministrazione di una molecola, ma sull'utilizzo di cellule vive del sistema immunitario del paziente stesso, che vengono prelevate, modificate geneticamente e poi reinfuse con l'obiettivo di riconoscere e attaccare in modo più mirato ed efficiente le cellule tumorali (1).

Il principio che sta alla base di questa terapia è quello di modificare i linfociti T (globuli bianchi che gestiscono la risposta immunitaria mirata) del paziente inserendo, tramite tecniche di ingegneria genetica, un recettore artificiale chiamato CAR. Questo recettore consente ai linfociti T di riconoscere specifici antigeni presenti sulla superficie delle cellule tumorali e di attivare una risposta citotossica contro di esse (1).

Il meccanismo generale della terapia CAR-T è schematizzato in Figura 1.1.

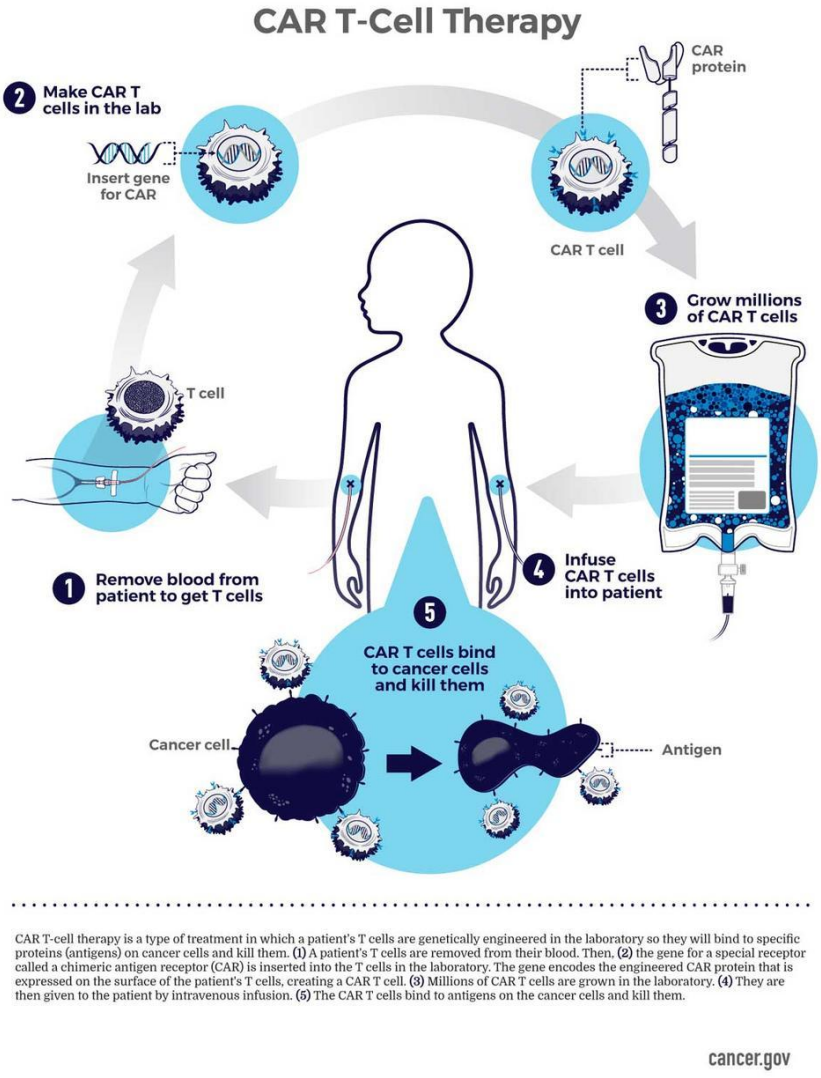


Figura 1.1 — Schema generale della terapia CAR-T.
Fonte: National Cancer Institute, "CAR T-Cell Therapy Infographic", cancer.gov.

Nella maggior parte dei casi, compreso il mio caso di studio, la terapia CAR-T è una terapia autologa (anche se talvolta può essere allogenica), cioè ottenuta a partire dalle cellule dello stesso paziente che riceverà il trattamento. Il prodotto terapeutico, dunque, è strettamente legato al singolo paziente: non si tratta quindi di un farmaco standard prodotto in serie, ma di un prodotto cellulare personalizzato. Esistono anche strategie allogeniche, basate su cellule provenienti da un donatore, ma nella pratica clinica la modalità autologa rappresenta ancora l'approccio principale (2).

Come schematizzato in Figura 1.1, il percorso terapeutico prevede il prelievo dei linfociti T dal paziente, la loro modifica a livello genetico e l'espansione in laboratorio fino alla successiva reinfusione. Una volta somministrate, le cellule CAR-T riconoscono l'antigene bersaglio espresso dalle cellule tumorali e ne determinano l'eliminazione.

Dal punto di vista clinico, le cellule CAR-T hanno trovato applicazione soprattutto nelle neoplasie ematologiche, in particolare in alcune forme di leucemia linfoblastica acuta, linfomi non-Hodgkin a cellule B e mieloma multiplo. In genere vengono considerate in pazienti con malattia recidivata o refrattaria, cioè quando le terapie convenzionali non hanno funzionato o la malattia si è ripresentata dopo precedenti trattamenti. In questi casi, la terapia CAR-T può rappresentare un'opzione terapeutica molto rilevante, soprattutto per pazienti che hanno già esaurito altre possibilità di cura (3).

L'applicazione ai tumori solidi, invece, è ancora più complessa e rappresenta un ambito di ricerca in continuo sviluppo. Le difficoltà principali riguardano la capacità delle cellule CAR-T di raggiungere efficacemente la massa tumorale, l'eterogeneità degli antigeni espressi dalle cellule neoplastiche e la presenza di un microambiente tumorale spesso immunosoppressivo, che può ridurre l'efficacia della risposta immunitaria. Per questo motivo, ad oggi, il razionale clinico della terapia CAR-T risulta consolidato nei tumori del sangue, mentre nei tumori solidi rimangono ancora diversi limiti biologici e terapeutici da superare (4).

Nonostante il grande potenziale clinico, la terapia CAR-T non è esente da criticità importanti. L'attivazione massiccia del sistema immunitario può infatti determinare eventi avversi anche importanti, tra cui la sindrome da rilascio di citochine, nota come CRS *(Cytokine Release Syndrome)*, e la neurotossicità associata alle cellule effettrici immunitarie, conosciuta come ICANS *(Immune Effector Cell-Associated Neurotoxicity Syndrome)*. Queste complicanze richiedono monitoraggio continuo, personale altamente formato e disponibilità rapida di terapie di supporto, perché possono evolvere rapidamente e diventare clinicamente rilevanti (3).

Per questi motivi, la terapia CAR-T deve essere considerata non solo come una terapia innovativa dal punto di vista biologico, ma anche come un processo clinico e organizzativo ad altissima complessità. Il trattamento coinvolge infatti diverse fasi, più figure professionali e una gestione molto rigorosa del prodotto cellulare, dalla raccolta fino alla reinfusione. Questa complessità rende necessario un percorso ben strutturato, nel quale ogni passaggio deve essere controllato, documentato e coordinato in modo preciso (3).

Il percorso terapeutico CAR-T non può essere ridotto al solo momento dell'infusione del farmaco. Al contrario, si tratta di un processo lungo e articolato, composto da diverse fasi strettamente collegate tra di loro. Ogni passaggio ha un ruolo specifico e può influenzare l'efficacia e la sicurezza del trattamento. Proprio per questo, la terapia CAR-T richiede una forte integrazione tra attività cliniche, trasfusionali, laboratoristiche, logistiche e organizzative (2).

La prima fase è rappresentata dalla selezione e valutazione del paziente. Prima di procedere con il trattamento, è necessario verificare che il paziente sia realmente idoneo alla terapia, sia dal punto di vista clinico sia dal punto di vista organizzativo. Vengono quindi valutati lo stato della malattia, le condizioni generali del paziente, la funzionalità degli organi e l'eventuale presenza di fattori che potrebbero aumentare il rischio di complicanze. In questa fase assume grande importanza anche il

consenso informato, poiché il paziente deve essere messo nelle condizioni di comprendere benefici attesi, possibili tossicità e criticità del percorso (3).

Dopo la conferma dell'idoneità, si procede con la leucaferesi, ovvero la raccolta dei linfociti T dal sangue periferico del paziente stesso. Questa fase è essenziale perché fornisce il materiale cellulare di partenza da cui verrà prodotto il farmaco CAR-T. La procedura deve rispettare requisiti clinici e tecnici precisi, poiché la qualità e la quantità delle cellule raccolte condizionano le fasi successive della produzione e quindi l'efficacia finale della terapia (2).

Una volta raccolte, le cellule devono essere identificate, confezionate e inviate al sito produttivo. Questa fase introduce una componente logistica molto delicata, perché il prodotto cellulare deve rimanere sempre correttamente associato al paziente da cui proviene. Per questo motivo diventano fondamentali la catena di identità (*Chain of Identity*) e la catena di custodia (*Chain of Custody)*. La prima garantisce che il prodotto sia sempre riconducibile al paziente corretto, mentre la seconda permette di tracciare tutti i passaggi, gli spostamenti e le responsabilità lungo il percorso (2).

Durante il periodo necessario alla produzione del farmaco, il paziente può ricevere una terapia ponte *(Bridge Therapy)*. Questa terapia ha lo scopo di controllare temporaneamente la malattia nell'attesa che il prodotto CAR-T sia disponibile. La scelta della terapia ponte dipende dalle condizioni cliniche del paziente, dall'aggressività della malattia e dai tempi previsti per la produzione del prodotto cellulare. Anche questa fase deve essere gestita con attenzione, perché il paziente deve arrivare all'infusione in condizioni compatibili con il trattamento (3).

Quando il prodotto CAR-T viene rilasciato dal sito produttivo, il paziente viene preparato all'infusione mediante chemioterapia linfodepletiva. Questa non ha lo scopo principale di eliminare direttamente il tumore, ma di creare un ambiente immunologico favorevole all'espansione e alla proliferazione delle cellule CAR-T dopo la reinfusione. Generalmente vengono utilizzati farmaci come fludarabina e ciclofosfamide, con l'obiettivo di ridurre alcune popolazioni cellulari e favorire la proliferazione dei linfociti T modificati (3).

La fase di infusione rappresenta il momento centrale del trattamento, ma non conclude il percorso. Prima della somministrazione, il prodotto deve essere ricevuto, controllato, conservato e, quando necessario, scongelato secondo procedure standardizzate e regolamentate su indicazione del produttore. L'infusione deve essere eseguita da personale qualificato, dopo il via libera dell'équipe medica e dopo un'ulteriore verifica delle condizioni del paziente. Durante e dopo la somministrazione vengono monitorati i parametri vitali e viene completata tutta la documentazione relativa al trattamento (2).

Dopo l'infusione, il paziente deve essere seguito con attenzione per riconoscere tempestivamente eventuali complicanze. Le principali criticità cliniche comprendono la sindrome da rilascio di citochine, la neurotossicità, le citopenie e le infezioni. In questa fase è quindi necessario garantire un monitoraggio frequente, la disponibilità di farmaci di supporto come il Tocilizumab in caso di CRS, protocolli di gestione delle emergenze e personale formato nel riconoscimento precoce delle tossicità (3).

Il percorso prosegue poi con i follow-up, sia a breve sia a lungo termine. Nel breve periodo, il controllo serve soprattutto a intercettare tossicità che possono insorgere, infezioni, citopenie e alterazioni neurologiche. Nel lungo periodo, invece, il follow-up permette di valutare la risposta della malattia, la persistenza dell'effetto terapeutico, il recupero immunologico e l'eventuale comparsa di complicanze tardive. Questo conferma che la terapia CAR-T non è una singola terapia isolata, ma un processo che si protrae nel tempo, estremamente complesso, e che quindi richiede continuità assistenziale (2).

Nel complesso, il percorso CAR-T è caratterizzato da una forte complessità clinica, logistica e organizzativa. La presenza di molti passaggi sequenziali, la necessità di mantenere la tracciabilità

del prodotto cellulare, il coinvolgimento di più strutture e il rischio di tossicità anche severe rendono indispensabile una gestione standardizzata ed estremamente regolamentata dell'intero processo. Proprio questa complessità rende particolarmente importante l'identificazione dei rischi presenti nelle diverse fasi del percorso e costituisce il presupposto per l'applicazione di metodologie strutturate di analisi del rischio (3).

La complessità del percorso CAR-T non deriva solo dal numero elevato di fasi che lo compongono, ma in particolar modo dal fatto che ogni fase è strettamente dipendente dalle altre. Il trattamento richiede infatti il coordinamento tra attività cliniche, trasfusionali, produttive, logistiche e documentali. Di conseguenza, un errore o un ritardo in un singolo passaggio può avere ripercussioni sull'intero percorso terapeutico, influenzando la sicurezza del paziente, la qualità del prodotto cellulare o la possibilità stessa di completare il trattamento (2,3).

Un primo ambito critico riguarda il rischio clinico. Il paziente candidato a terapia CAR-T presenta spesso una malattia avanzata, recidivata o refrattaria, e potrebbe avere un quadro clinico fragile. Per questo motivo è fondamentale valutare correttamente l'idoneità al trattamento, monitorare l'evoluzione clinica durante l'attesa del prodotto e riconoscere tempestivamente le tossicità dopo l'infusione. Eventi come la sindrome da rilascio di citochine, la neurotossicità, le infezioni e le citopenie richiedono infatti personale formato, protocolli condivisi e disponibilità rapida di terapie di supporto (3).

Un secondo ambito riguarda il rischio logistico e di tracciabilità. Poiché il prodotto CAR-T è generalmente autologo, esso è legato in maniera univoca al singolo paziente. Questo rende essenziali la corretta identificazione del materiale cellulare, la tracciabilità dei passaggi e il mantenimento della catena di identità e custodia. Errori di etichettatura, documentazione incompleta, ritardi nella spedizione o problemi nella conservazione del prodotto possono compromettere la sicurezza e l'affidabilità dell'intero processo (2).

Accanto ai rischi clinici e logistici, sono presenti anche rischi organizzativi e gestionali. Il percorso CAR-T coinvolge numerose figure professionali e diverse strutture, tra cui reparto clinico, servizio trasfusionale, laboratorio, farmacia, centro produttivo e personale amministrativo. La comunicazione tra questi attori deve essere continua e precisa, perché la pianificazione delle tempistiche è un elemento centrale: raccolta cellulare, produzione, terapia ponte, condizionamento linfodepletivo e infusione devono essere coordinati in modo coerente monitorando le condizioni del paziente e la disponibilità del prodotto (2,3).

Infine, un ulteriore elemento di criticità è rappresentato dalla gestione documentale e procedurale. Ogni passaggio del percorso deve essere registrato, verificato e svolto secondo procedure standardizzate. La documentazione non ha quindi solo un valore amministrativo, ma rappresenta uno strumento fondamentale per garantire qualità, sicurezza, tracciabilità e responsabilità lungo tutto il processo.

Nel complesso, il processo CAR-T può essere considerato un percorso ad alto rischio perché combina complessità clinica, personalizzazione del prodotto, vincoli logistici stringenti, necessità di tracciabilità e coinvolgimento di più professionisti e strutture.

Per questo motivo, la gestione del rischio non può essere limitata alla correzione degli errori dopo che si sono verificati, ma deve basarsi su un approccio proattivo, capace di identificare in maniera accurata ma soprattutto in anticipo i punti critici e deboli del processo e di definire misure preventive adeguate che riducano al minimo la probabilità di errore che rischierebbe di compromettere l'intero processo.

1.2 L'analisi del rischio

L'analisi del rischio rappresenta uno strumento fondamentale per individuare in modo sistematico le possibili criticità di un processo e per definire azioni finalizzate a ridurre la probabilità che tali criticità si trasformino in eventi avversi. In ambito sanitario assume particolare importanza, poiché molti processi assistenziali coinvolgono più professionisti, strutture e attività tra loro interdipendenti. In questi contesti, l'obiettivo non è soltanto intervenire dopo il verificarsi di un errore, ma identificare preventivamente i punti vulnerabili del processo e adottare misure in grado di ridurne la probabilità o limitarne le conseguenze (5).

In questo contesto si inseriscono metodologie strutturate come la FMEA, la FMECA e la HFMEA.

La FMEA, acronimo di *Failure Mode and Effects Analysis*, è una metodologia proattiva nata in ambito ingegneristico e successivamente migrata anche in ambito sanitario. Il suo obiettivo è analizzare un processo in modo sistematico, individuando i possibili modi di fallimento, le cause che possono determinarli e gli effetti che tali fallimenti potrebbero produrre. In altre parole, la FMEA ha come obiettivo di chiedersi, per ogni fase di un processo, quali sono le principali problematiche, perché potrebbero accadere e quali conseguenze potrebbero avere sul risultato finale (5).

Una variante della FMEA è la FMECA, acronimo di *Failure Mode, Effects and Criticality Analysis*. Rispetto alla FMEA tradizionale, la FMECA introduce in modo più esplicito il concetto di criticità, cioè la valutazione della rilevanza e dell'importanza del rischio associato a ciascun modo di fallimento. Questo consente non solo di identificare i possibili errori, ma anche di stabilire quali siano quelli prioritari e quindi quelli che richiedono un intervento più urgente. In questo senso, la FMECA aggiunge alla descrizione dei fallimenti una logica di classificazione e prioritizzazione, utile soprattutto nei processi ad alta complessità, dove non tutti i rischi hanno lo stesso impatto potenziale e quindi si stila una classifica in cui i rischi prioritari sono quelli in cui si interverrà in maniera più urgente (5).

In ambito sanitario, una delle evoluzioni più interessanti di questo approccio è rappresentata dalla HFMEA, acronimo di *Healthcare Failure Mode and Effect Analysis*. Questa metodologia è stata sviluppata dal *National Center for Patient Safety* del *Department of Veterans Affairs* con l'obiettivo di adattare la logica della FMEA ai processi clinici e assistenziali. La HFMEA mantiene l'impostazione proattiva della FMEA, ma la integra con strumenti pensati specificamente per il contesto sanitario, come l'analisi tramite team multidisciplinare, la rappresentazione del processo mediante flowchart, la matrice di rischio e l'albero decisionale (6,7).

La HFMEA si basa su una sequenza di passaggi strutturati:

1. definizione del processo da analizzare, individuando chiaramente il tema e i confini dell'analisi;
2. costituzione di un gruppo di esperti, composto da figure professionali con competenze diverse che sono direttamente coinvolte nel processo che si sta analizzando;
3. descrizione grafica del processo, attraverso un diagramma di flusso *(flowchart)* che permette di suddividere il percorso in fasi e sottoprocessi;
4. analisi dei pericoli. Per ciascun sottoprocesso vengono identificate le modalità con cui quella fase potrebbe non raggiungere il proprio obiettivo (*failure mode*). Per ogni failure mode vengono poi considerate le possibili cause e le potenziali conseguenze. A ciascun rischio viene assegnato un punteggio, detto Hazard Score, calcolato sulla base della gravità dell'evento e della probabilità che esso si verifichi. Questo punteggio permette di ottenere una prima classificazione dei rischi e di concentrare l'attenzione sugli elementi più rilevanti (7);
5. definizione delle azioni. Per i rischi considerati significativi, il team stabilisce se sia necessario eliminarli, controllarli o accettarli. Le azioni correttive e preventive devono essere concrete,

tracciabili e associate a responsabilità definite. L'obiettivo non è quindi soltanto descrivere e caratterizzare il rischio, ma trasformare l'analisi in uno strumento per migliorare la sicurezza del processo e ridurre la probabilità di eventi avversi (7).

Questa suddivisione è di fondamentale importanza perché consente di visualizzare in modo ordinato e organico il processo, evidenziando i possibili punti critici in cui è necessario un intervento (7).

Un elemento caratteristico della HFMEA è l'utilizzo dell'albero decisionale. Questo strumento consente di stabilire se un determinato *failure mode* debba essere ulteriormente analizzato o se possa essere escluso dalle fasi successive. La decisione tiene conto di diversi aspetti, tra cui la gravità del rischio, l'esistenza di controlli già presenti, la possibilità di rilevare l'errore prima che produca un danno e la presenza di eventuali punti singoli di fragilità del processo. In questo modo, l'analisi non si limita a elencare tutti i rischi possibili, ma permette di selezionare quelli che richiedono realmente un intervento (7).

Nel caso della terapia CAR-T, l'utilizzo di un approccio come la HFMEA risulta particolarmente coerente con la natura del percorso terapeutico. Come descritto nei paragrafi precedenti, la CAR-T è un processo formato da molte fasi interdipendenti, caratterizzato da elevata personalizzazione del prodotto, complessità logistica, necessità di tracciabilità e rischio di tossicità anche severe. Per questo motivo, una metodologia proattiva e strutturata, come HFMEA, permette di analizzare il percorso prima che l'errore si verifichi, individuando i punti più vulnerabili e deboli, adottando la definizione di misure preventive adeguate che permettano di arginare gli eventi avversi (2,3,7).

La sequenza delle principali fasi previste nella metodologia HFMEA è rappresentata schematicamente in Figura 1.2.

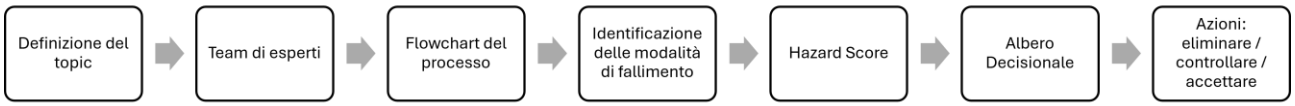


Figura 1.2 — Schema sintetico della metodologia HFMEA.

1.3 Motivazione e obiettivo della tesi

La terapia CAR-T rappresenta un esempio particolarmente significativo di processo sanitario ad alta complessità. La sua gestione richiede il coordinamento di attività cliniche, trasfusionali, produttive, logistiche, documentali e organizzative. Inoltre, il prodotto terapeutico è strettamente associato al singolo paziente e deve essere tracciato lungo tutto il percorso, dalla raccolta cellulare fino al follow-up. Questa combinazione di personalizzazione, complessità clinica e vincoli organizzativi rende il processo CAR-T particolarmente esposto a potenziali criticità (2,3).

In questo contesto, l'analisi del rischio assume un ruolo centrale. Tuttavia, la costruzione di una tabella di rischio completa e coerente richiede tipicamente un notevole sforzo di elaborazione manuale, risultando particolarmente onerosa in termini di tempo e risorse. È necessario consultare numerosi documenti procedurali, individuare le informazioni rilevanti, associare ogni rischio alla fase corretta del processo, distinguere cause e conseguenze, riconoscere eventuali ridondanze e trasformare descrizioni eterogenee in una struttura uniforme. Questo lavoro può diventare complesso soprattutto quando i documenti utilizzati provengono da fonti diverse, presentano livelli di dettaglio non omogenei, o descrivono lo stesso rischio con formulazioni differenti.

A partire da queste considerazioni, la presente tesi si propone di sviluppare e valutare una pipeline per il supporto alla costruzione di una tabella di rischio relativa al processo CAR-T. L'obiettivo non è sostituire il giudizio degli esperti clinici e di organizzazione sanitaria, che rimane essenziale in un'analisi di rischio, ma fornire una pipeline semi-automatica in grado di supportare le fasi di estrazione, organizzazione, astrazione e confronto delle informazioni contenute nei documenti procedurali.

In particolare, il lavoro si concentra sull'utilizzo di modelli linguistici di grandi dimensioni (*Large Language Models, LLMs*) per l'estrazione strutturata delle informazioni di rischio dai documenti disponibili. Le informazioni estratte vengono organizzate secondo le fasi del processo CAR-T e ricondotte a uno schema tabellare predefinito. Successivamente, le righe che descrivono rischi semanticamente simili vengono astratte e, quando opportuno, aggregate, con l'obiettivo di ridurre le ridondanze presenti e ottenere una rappresentazione più sintetica, meno spuria e standardizzata dei rischi.

Un ulteriore obiettivo della tesi è valutare la qualità degli output generati dalla pipeline attraverso il confronto con un riferimento di validazione stabilito da esperti. A questo scopo, vengono utilizzate metriche tradizionali per i campi strutturati e metodi di confronto semantico per i campi narrativi. In particolare, il confronto viene condotto sia tramite *embedding* sia tramite un approccio basato su *LLM-as-a-judge*, al fine di valutare non solo la corrispondenza tra gli elementi strutturati delle righe, ma anche la coerenza semantica tra rischio estratto e rischio di riferimento.

La tesi si configura quindi come uno studio di fattibilità volto a verificare se una pipeline semi-automatizzata basata su modelli linguistici possa contribuire alla costruzione e alla standardizzazione di una tabella di rischio per un processo sanitario complesso come quello CAR-T.

La pipeline proposta, quindi, può essere interpretata come uno strumento di supporto preliminare all'attività degli esperti. L'obiettivo, quindi, non è soltanto automatizzare una parte dell'estrazione delle informazioni, ma anche organizzare in modo più chiaro i dati di rischio, facilitarne il confronto con un riferimento e mettere in evidenza i punti critici che dovranno essere successivamente valutati in modo clinico e gestionale

2. Stato dell'Arte

2.1 Procedure Operative Standard (SOP) in ambito ospedaliero

Le Procedure Operative Standard, indicate comunemente con l'acronimo inglese SOP, *Standard Operating Procedures*, rappresentano uno degli strumenti utilizzati per standardizzare lo svolgimento delle attività all'interno di un'organizzazione. Le SOP possono essere definite come istruzioni scritte e dettagliate finalizzate a garantire uniformità nell'esecuzione di una specifica attività o funzione (8). In ambito clinico e sanitario, questi documenti hanno lo scopo di descrivere in modo standardizzato le modalità operative da seguire, riducendo la variabilità tra operatori e favorendo la tracciabilità, la qualità e la riproducibilità dei processi. Inoltre, all'interno dei sistemi di gestione per la qualità, la documentazione operativa rientra tra le informazioni documentate che devono essere mantenute, controllate e rese disponibili per supportare il corretto funzionamento dei processi (9).

Da un punto di vista pratico, una SOP può contenere non solo descrizioni testuali delle attività da svolgere, ma anche tabelle, diagrammi di flusso, checklist operative, moduli da compilare riferimenti ad allegati e documenti collegati, indicazioni sulle responsabilità del personale e sui controlli da effettuare, oltre a informazioni riguardanti la versione del documento o alle revisioni che sono state effettuate.

Per questo motivo, le SOP e, più in generale, i documenti procedurali possono rappresentare una fonte particolarmente rilevante per l'analisi del rischio. Descrivono infatti non solo le attività da svolgere, ma anche gli attori coinvolti, le responsabilità, i controlli previsti e i passaggi critici del processo.

Nei processi sanitari, nei quali più attività e figure professionali possono essere tra loro strettamente dipendenti, la standardizzazione delle modalità operative assume quindi particolare importanza. La disponibilità di procedure formalizzate permette di definire in modo più chiaro le responsabilità, i controlli e le modalità di esecuzione delle attività, fornendo allo stesso tempo una base documentale dalla quale possono essere individuate eventuali criticità del processo (9).

2.2 Elaborazione automatica del testo

La crescente disponibilità di documentazione sanitaria in formato digitale ha reso possibile l'applicazione di strumenti informatici per analizzare e organizzare automaticamente grandi quantità di informazioni. Una parte rilevante dei dati prodotti in ambito sanitario è infatti presente sotto forma di testo non strutturato, come referti, lettere cliniche, note mediche e documentazione relativa a procedure diagnostiche o terapeutiche. Anche se questi documenti contengono informazioni importanti, la loro forma testuale non strutturata rende più difficile un'elaborazione sistematica e richiede spesso un'attività manuale di lettura e riorganizzazione (10).

In questo contesto si inserisce l'elaborazione del linguaggio naturale (*Natural Language Processing* NLP), cioè l'insieme di tecniche informatiche utilizzate per elaborare e analizzare il linguaggio. Quando queste tecniche vengono applicate a documentazione sanitaria o clinica si parla in genere di *Clinical NLP*. L'obiettivo può essere molto diverso a seconda dell'applicazione: classificare documenti, individuare informazioni specifiche all'interno del testo, riconoscere entità o relazioni, sintetizzare contenuti oppure trasformare informazioni inizialmente non strutturate in dati organizzati e più facilmente elaborabili.

Nel corso degli anni, gli approcci utilizzati per l'elaborazione automatica del testo si sono progressivamente evoluti. Ai metodi basati principalmente su regole definite manualmente e a modelli statistici specifici per un determinato compito si sono affiancati modelli di apprendimento automatico e, successivamente, architetture neurali di crescente complessità. Una svolta determinante in questo percorso evolutivo è stata l'introduzione dell'architettura *Transformer*, basata sul meccanismo di *attention*, che ha permesso di modellare in maniera più efficace le relazioni tra gli elementi di una sequenza testuale (11).

In particolare, il meccanismo di self attention consente al modello di valutare simultaneamente le relazioni tra le diverse parole di una sequenza, attribuendo maggiore importanza agli elementi più rilevanti per il contesto. A differenza delle architetture ricorrenti precedentemente utilizzate, il Transformer permette di elaborare gli elementi della sequenza in parallelo, favorendone così l'efficienza nell'addestramento e una maggiore capacità di rappresentare dipendenze anche a dovuta distanza nel testo.

Nell'architettura Transformer possono essere distinte due componenti principali: l'*encoder* che elabora la sequenza di input e ne costruisce una rappresentazione contestuale e il *decoder,* che utilizza tale rappresentazione per generare una sequenza di output. A seconda del compito i modelli linguistici possono utilizzare entrambe le componenti oppure soltanto una di esse.

A partire dall'architettura Transformer si sono sviluppate diverse famiglie di modelli linguistici, caratterizzate da utilizzi differenti dei blocchi di *encoder* e *decoder.*

Una prima categoria è rappresentata dai modelli *encoder-only,* come ad esempio BERT, nei quali viene utilizzata esclusivamente la componente di codifica da parte del Transformer. Questi modelli elaborano il testo in ingresso considerando il contesto complessivo della sequenza e producono rappresentazioni contestuali delle parole, sono adatti a compiti di comprensione del testo, classificazione e al riconoscimento di entità (12).

Una seconda famiglia è costituita dai modelli *encoder-decoder, c*ome T5. In questo caso, l'encoder costruisce una rappresentazione del testo fornito in ingresso, mentre il decoder utilizza questa rappresentazione per la generazione di una nuova sequenza testuale. Questa struttura risulta adatta ai compiti in cui un testo deve essere trasformato in un altro testo, come traduzione, riassunto o generazione di una risposta a partire da un determinato input (13).

Parallelamente si sono sviluppati modelli basati prevalentemente sulla componente decoder del Transformer. Durante l'addestramento, questi modelli apprendono a predire il token successivo di

una sequenza sulla base dei token che lo precedono, assegnando una distribuzione di probabilità ai possibili token successivi. Con il termine token si indica una delle unità elementari in cui il testo viene suddiviso dal modello, che può corrispondere a una parola, a una parte di parola o a un simbolo. In fase di generazione, tale processo viene ripetuto in maniera autoregressiva: ogni token generato viene aggiunto alla sequenza e utilizzato come contesto per la predizione del token successivo. Questo approccio è alla base di numerosi modelli linguistici generativi moderni (14).

Su questa architettura si basano molti dei modelli linguistici sviluppati e utilizzati oggigiorno. In particolare, i modelli linguistici di grandi dimensioni (*Large Language* Models, LLMs) sono modelli addestrati su enormi quantità di dati testuali e caratterizzati da un elevato numero di parametri (raggiungendo l'ordine dei miliardi). A differenza dei modelli specializzati per uno specifico compito, come la classificazione del testo o il riconoscimento di entità, gli LLM possono essere guidati attraverso istruzioni testuali in fase di inferenza, definite nel *prompt*, per svolgere attività differenti senza richiedere necessariamente un nuovo addestramento specifico. Tale versatilità ne ha favorito la diffusione e l'utilizzo anche in ambito medico, come per esempio nell'elaborazione della documentazione sanitaria, per attività di sintesi, classificazione ed estrazione strutturata di informazioni (10).

Una delle applicazioni più rilevanti nell'ambito NLP è rappresentata dall'estrazione di informazioni (*Information Extraction)* specifiche da un testo non strutturato e la loro trasformazione in una rappresentazione strutturata. Questo tipo di elaborazione permette, a partire da documenti redatti in forma discorsiva, di ricondurre determinate informazioni a campi predefiniti, rendendole così utilizzabili da sistemi informatici o da procedure di analisi quantitativa.

L'estrazione automatica di informazioni da un testo è stata affrontata attraverso metodologie diverse già prima della diffusione degli LLM. I primi approcci si basavano prevalentemente su regole, dizionari e pattern definiti manualmente per individuare specifiche informazioni all'interno dei documenti. Successivamente si sono diffusi metodi di tipo statistico e di machine learning supervisionato, che venivano addestrati su dati annotati per riconoscere automaticamente entità, relazioni e altre informazioni di interesse. Con lo sviluppo dei modelli neurali e dei modelli linguistici pre-addestrati, tali compiti hanno potuto beneficiare di rappresentazioni contestuali del testo, fino alla più recente applicazione degli LLM mediante istruzioni definite nel *prompt* (15).

Questi approcci sono stati applicati anche in ambito sanitario prima della diffusione degli LLM. In particolare, nel settore radiologico sono stati sviluppati sistemi di NLP per identificare e estrarre automaticamente informazioni dai referti testuali, con l'obiettivo di trasformare il contenuto narrativo non strutturato in una rappresentazione più strutturata (16).

Un esempio recente in ambito sanitario è rappresentato dal lavoro di Wiest et al. (10), nel quale viene proposta una pipeline basata su LLM per estrarre informazioni cliniche predefinite da testi non strutturati, provenienti da un reparto di oncologia e convertirle in una forma strutturata. Il sistema è inoltre progettato per essere eseguito direttamente sull'infrastruttura ospedaliera, evitando il trasferimento dei dati clinici verso servizi esterni e preservando quindi la riservatezza dei documenti. Gli autori mostrano come l'utilizzo di modelli linguistici possa supportare il passaggio da testo libero a dati strutturati, mantenendo comunque la necessità di verificare la qualità delle informazioni estratte (10).

In questo contesto assumono particolare importanza gli embedding, cioè rappresentazioni numeriche vettoriali del contenuto testuale. Un embedding può rappresentare differenti unità linguistiche, come token, parole, frasi o interi documenti. I sentence embeddings costituiscono un caso specifico, nel quale un'intera frase o porzione di testo viene rappresentata mediante un unico vettore, cercando di preservarne il contenuto semantico. In questo modo, testi semanticamente simili tendono a essere rappresentati da vettori vicini nello spazio, rendendo possibile stimarne

quantitativamente la somiglianza. Un esempio di questo approccio è Sentence-BERT, progettato specificamente per ottenere rappresentazioni di frasi adatte al confronto semantico (17).

Negli anni successivi sono stati sviluppati modelli di embedding in grado di lavorare su più lingue e su testi di diversa lunghezza. Un esempio è BGE-M3, un modello multilingue progettato per supportare differenti modalità di rappresentazione e recupero dell'informazione, tra cui rappresentazioni dense, nelle quali il contenuto del testo viene sintetizzato in un vettore numerico compatto; sparse, maggiormente legate alla presenza e al peso dei termini nel testo e *multi-vector*, nelle quali vengono mantenute più rappresentazioni vettoriali dello stesso contenuto per consentire confronti più dettagliati. BGE-M3 è inoltre in grado di elaborare contenuti in numerose lingue (18).

La similarità vettoriale, tuttavia, non rappresenta l'unico modo per confrontare automaticamente due contenuti testuali. Quando il confronto richiede una maggiore interpretazione del significato, un approccio che si è sviluppato più recentemente consiste nell'utilizzare un modello linguistico come valutatore automatico. In questo caso, l'LLM non viene utilizzato principalmente per generare nuovo contenuto, ma per esaminare due o più output e valutarne la qualità, la coerenza o la corrispondenza rispetto a criteri definiti. Questo approccio viene generalmente indicato come *LLM-as-a-judge*: il modello riceve i contenuti da confrontare insieme a criteri di valutazione definiti nel *prompt* e assegna un giudizio o un punteggio sulla base della corrispondenza tra i testi e dei criteri di valutazione definiti (19).

L'utilizzo degli LLM come valutatori è stato inizialmente studiato soprattutto per giudicare la qualità di contenuti generati da altri modelli. Rispetto a metriche puramente lessicali o vettoriali, il vantaggio potenziale consiste nella possibilità di considerare il contesto e il significato complessivo della risposta. Inoltre, l'utilizzo di un modello come valutatore consente di automatizzare e scalare il confronto a un numero elevato di output, riducendo la necessità di una valutazione manuale caso per caso. Allo stesso tempo, questa metodologia presenta alcune criticità, poiché la valutazione può dipendere dal modello utilizzato, dal prompt, dalla scala di giudizio e da possibili bias del valutatore automatico. Inoltre, trattandosi di un modello generativo, il giudizio può presentare una certa variabilità tra esecuzioni, limitandone la completa riproducibilità deterministica (19).

L'applicazione di questo approccio in ambito sanitario è più recente. Croxford et al. (20) hanno valutato l'utilizzo di LLM come giudici automatici per la qualità di sintesi cliniche generate a partire da cartelle sanitarie elettroniche. I giudizi prodotti dai modelli sono stati confrontati con quelli di valutatori clinici umani attraverso una griglia di valutazione definita. Il miglior modello analizzato dagli autori ha mostrato un elevato accordo con la valutazione degli esperti, con un Intraclass Correlation Coefficient pari a 0,818. I risultati suggeriscono quindi che l'LLM-as-a-judge possa rappresentare uno strumento scalabile per supportare la valutazione automatica di contenuti clinici complessi, pur non eliminando la necessità di una validazione umana (20).

L'utilizzo di modelli linguistici in ambito sanitario introduce però anche aspetti tecnici e organizzativi che devono essere considerati. Una prima distinzione riguarda l'utilizzo di modelli accessibili tramite servizi esterni, generalmente attraverso infrastrutture cloud e API, rispetto a modelli che possono essere installati ed eseguiti localmente. I servizi cloud possono consentire l'accesso a modelli di grandi dimensioni senza richiedere un'infrastruttura computazionale dedicata all'interno della struttura sanitaria; allo stesso tempo, quando i dati trattati contengono informazioni cliniche o documentazione riservata, il trasferimento dei contenuti verso servizi esterni introduce problematiche relative alla riservatezza e alla protezione dei dati. L'esecuzione locale consente invece di mantenere i dati all'interno dell'infrastruttura dell'organizzazione, riducendo la necessità di trasferire la documentazione verso sistemi esterni. Questa soluzione è stata adottata anche in alcune pipeline proposte per l'estrazione di informazioni cliniche. Ad esempio, Wiest et al. (10) descrivono un sistema eseguito direttamente sull'infrastruttura ospedaliera, progettato proprio per evitare il trasferimento esterno dei dati sanitari. Tuttavia, l'utilizzo locale richiede la disponibilità di risorse computazionali adeguate: modelli con un numero elevato di parametri possono infatti

necessitare di GPU *(Graphic Processing Unit)* e quantità rilevanti di memoria, rendendo necessario trovare un compromesso tra capacità del modello, velocità di elaborazione e risorse disponibili (10).

Un ulteriore limite è rappresentato dalla possibilità che i modelli generino informazioni non presenti o non correttamente supportate dal testo di partenza, fenomeno generalmente indicato come *allucinazioni*. In ambito sanitario questo aspetto assume particolare rilevanza, perché un'informazione apparentemente plausibile ma non presente nella documentazione originale può modificare in maniera importante il contenuto del documento. Oltre alle *allucinazioni*, possono verificarsi omissioni, difficoltà nella gestione di documenti molto lunghi e prestazioni non uniformi in funzione della lingua o del tipo di testo analizzato. Per questo motivo, l'automazione basata su modelli linguistici deve essere considerata principalmente come uno strumento di supporto e deve prevedere modalità di verifica e controllo degli output (10,20).

Queste caratteristiche rendono NLP, LLM, embedding e sistemi di valutazione automatica strumenti potenzialmente utili anche nell'elaborazione della documentazione procedurale. La loro applicazione all'analisi del rischio richiede però di considerare non soltanto la capacità di estrarre e confrontare informazioni, ma anche la qualità, la tracciabilità e l'affidabilità degli output prodotti. Su questi aspetti si concentrano i lavori più direttamente collegati all'analisi automatizzata del rischio, descritti nei paragrafi successivi.

2.3 Metodologie di analisi del rischio in sanità e lavori correlati

L'applicazione di metodologie strutturate di analisi del rischio ha assunto un ruolo sempre più importante anche in ambito sanitario, dove la sicurezza del processo dipende spesso dall'interazione tra attività, professionisti e strutture differenti. In questo contesto, l'obiettivo dell'analisi non è solamente individuare e mitigare gli errori già avvenuti, ma cercare di identificare in maniera preventiva i possibili punti di debolezza di un processo e stabilire quali di essi richiedano maggiore attenzione (5).

Come descritto precedentemente nel Capitolo 1, metodologie come FMEA, FMECA e HFMEA permettono di scomporre un processo nelle sue diverse fasi, identificare i possibili modi di fallimento e analizzarne cause e conseguenze. In ambito sanitario questo tipo di valutazione viene generalmente svolto attraverso il coinvolgimento di un gruppo multidisciplinare, nel quale professionisti con competenze differenti contribuiscono alla descrizione del processo e all'individuazione delle criticità sulla base della propria esperienza. La HFMEA, sviluppata specificamente per il settore sanitario dal Department of Veterans Affairs, mantiene infatti tra i suoi elementi centrali la definizione del processo, la sua rappresentazione attraverso un diagramma di flusso e la successiva analisi dei potenziali *failure modes* da parte del team (6,7).

Un elemento particolarmente importante di questi approcci è quindi la conoscenza dettagliata del processo oggetto di analisi. L'identificazione di un rischio non può essere separata dal contesto nel quale esso può verificarsi: è necessario conoscere le attività svolte, le figure coinvolte, le responsabilità, i controlli previsti e le relazioni tra le diverse fasi. Per questo motivo, la mappatura del processo e la disponibilità di documentazione procedurale rappresentano elementi importantissimi nella costruzione dell'analisi del rischio, soprattutto quando il percorso considerato è composto da numerosi passaggi interdipendenti (7).

L'applicazione di queste metodologie risulta particolarmente rilevante e sensata anche nel percorso CAR-T. Un esempio è rappresentato dal lavoro di Talarmin et al. (21), nel quale è stata condotta un'analisi del rischio del circuito farmaceutico CAR-T presso l'Hôpital Saint-Louis di Parigi. Gli autori hanno inizialmente costruito una mappatura del processo, ricostruendo le attività svolte dall'unità di terapia cellulare e dalla farmacia ospedaliera, e hanno successivamente applicato una FMECA per individuare i potenziali rischi e valutarne la criticità. Per i rischi considerati maggiormente rilevanti sono state quindi definite azioni correttive o preventive. Lo studio ha mostrato come le criticità del percorso CAR-T non siano esclusivamente di tipo clinico, ma possano riguardare anche aspetti organizzativi, logistici e procedurali. Tra i rischi individuati, gli autori evidenziano in particolare problemi legati alla tracciabilità e alla variabilità delle procedure tra differenti prodotti CAR-T. L'introduzione di specifiche azioni correttive e preventive ha permesso di ridurre la criticità di diversi rischi identificati (21).

L'analisi rimane tuttavia basata principalmente sulla ricostruzione del processo e sulla valutazione da parte degli esperti, senza prevedere strumenti automatici per l'estrazione delle informazioni dalla documentazione procedurale (21).

Nel complesso, questi approcci mostrano come l'analisi del rischio in sanità richieda una conoscenza estremamente approfondita del processo e una capacità di integrare informazioni provenienti da attività e professionalità differenti. La documentazione e la mappatura del processo costituiscono quindi una base importante per individuare e organizzare le possibili criticità, mentre il giudizio degli esperti rimane centrale per valutarne la rilevanza e definire le azioni da intraprendere.

Nei processi particolarmente articolati, tuttavia, la quantità di documentazione disponibile e il numero di informazioni da analizzare possono rendere complessa la costruzione e l'aggiornamento di una rappresentazione strutturata dei rischi. Questo aspetto ha favorito l'interesse verso strumenti

informatici in grado di supportare alcune fasi dell'analisi, dall'estrazione delle informazioni fino alla loro organizzazione e valutazione.

Un'applicazione di strumenti automatici collegata maggiormente alla sicurezza sanitaria è stata proposta da Denecke e Paula (22), che hanno valutato l'utilizzo di un LLM per analizzare automaticamente *incident report* provenienti dalla banca dati svizzera CIRRNET®. Il modello Gemma-2 è stato utilizzato su oltre 10.000 segnalazioni con l'obiettivo di estrarre eventi, cause e fattori contribuenti e raggrupparli per temi. Su un campione di 100 report valutati manualmente, gli autori hanno riportato un'accuratezza del 92% nell'estrazione degli eventi, dell'84% per le cause e del 72% per i fattori contribuenti (22). Questo studio mostra come i modelli linguistici possano essere utilizzati anche per identificare automaticamente elementi direttamente collegati alla sicurezza, come eventi, cause e fattori contribuenti. Allo stesso tempo, gli autori evidenziano una riduzione delle prestazioni nei fattori contribuenti, per i quali il modello tende talvolta a interpretare o generare informazioni non sufficientemente supportate dal testo. Il lavoro presenta quindi una forte analogia con l'estrazione delle cause di rischio, ma utilizza come fonte *incident report* relativi a eventi già segnalati, mentre un'analisi proattiva del rischio può partire anche dalla documentazione che descrive il processo prima che un evento avverso si verifichi (22).

Un lavoro ancora più vicino all'integrazione tra modelli linguistici e analisi proattiva del rischio è quello di Nair et al. (23), che hanno valutato l'utilizzo degli LLM come supporto a una *Failure Mode and Effects Analysis* in radioterapia. In questo studio quattro modelli linguistici, ovvero ChatGPT-4, Gemini 2.5 Pro, phi4-reasoning-14b e openAI/oss-120b, sono stati utilizzati per generare potenziali failure modes associati al workflow del *Radiation Planning Assistant*. I rischi proposti dai modelli sono stati successivamente rivalutati da un gruppo multidisciplinare di esperti, che ne ha verificato la rilevanza e riassegnato i punteggi di severità, probabilità di occorrenza e rilevabilità secondo il framework TG-100 (23).

Complessivamente, i modelli hanno generato 190 potenziali failure modes, ridotti a 79 elementi unici e interpretabili dopo la revisione degli esperti. Il confronto tra valutazione automatica e valutazione umana ha inoltre mostrato differenze nella stima della criticità dei rischi, confermando la necessità di mantenere una supervisione esperta. Gli autori concludono infatti che gli LLM possono ampliare la capacità di individuare potenziali *failure modes*, ma devono essere considerati uno strumento complementare e non sostitutivo del giudizio umano (23).

Rispetto al lavoro di Nair et al. (23), il presente studio utilizza un'impostazione diversa. L'obiettivo non è chiedere direttamente al modello di generare nuovi *failure modes* a partire dalla descrizione di un workflow, ma verificare se informazioni di rischio già presenti, in maniera esplicita o implicita, nella documentazione procedurale possano essere estratte, organizzate e successivamente sintetizzate in una struttura tabellare uniforme e confrontate con un riferimento. L'individuazione dei rischi viene quindi mantenuta maggiormente vincolata alle informazioni disponibili nei documenti di partenza, introducendo inoltre una fase successiva di astrazione e tracciabilità delle righe utilizzate. Un'ulteriore problematica riguarda la valutazione automatica degli output prodotti dai modelli linguistici. Croxford et al. (20) hanno studiato l'utilizzo di un approccio LLM-as-a-judge per valutare sintesi cliniche generate automaticamente a partire da cartelle sanitarie elettroniche. I giudizi prodotti dai modelli sono stati confrontati con quelli di sette valutatori clinici, utilizzando una griglia di valutazione strutturata. Il modello con le migliori prestazioni ha raggiunto un Intraclass Correlation Coefficient (ICC) pari a 0,818 rispetto alla valutazione umana, mostrando la possibilità di utilizzare gli LLM anche come strumenti di valutazione automatica di contenuti clinici complessi (20). Anche se non si riferisce all'analisi del rischio, questo studio è particolarmente rilevante perché mostra un possibile utilizzo degli LLM non nella generazione dell'informazione, ma nella valutazione della corrispondenza e della qualità di un output secondo criteri definiti. Questo principio risulta particolarmente interessante nei casi in cui la similarità tra due testi non possa essere valutata esclusivamente sulla base della coincidenza delle parole o della vicinanza vettoriale.

Nel complesso, i lavori presenti in letteratura mostrano quindi diversi punti di contatto con il problema affrontato nella presente tesi. Da un lato, l'analisi strutturata del rischio è già stata applicata al percorso CAR-T attraverso metodologie tradizionali basate sulla mappatura del processo e sulla valutazione degli esperti. Dall'altro, i modelli linguistici sono stati utilizzati per estrarre informazioni strutturate da documentazione sanitaria, analizzare cause e fattori contribuenti negli incident report e, più recentemente, supportare direttamente l'identificazione dei failure modes nell'ambito di una FMEA. Parallelamente, approcci basati su LLM-as-a-judge hanno mostrato la possibilità di utilizzare gli stessi modelli anche per la valutazione automatica di output testuali complessi (10,20–23).

Questi lavori affrontano tuttavia prevalentemente singole componenti del problema. A partire da questi approcci, la presente tesi combina in un'unica pipeline diverse metodologie già utilizzate separatamente in letteratura, applicandole alla documentazione procedurale del processo CAR-T e valuta la fattibilità di una pipeline semi-automatizzata che parte dalla documentazione procedurale, estrae le informazioni di rischio secondo uno schema strutturato, riduce le ridondanze mediante una successiva fase di astrazione e confronta infine gli output ottenuti con un riferimento fornito attraverso metriche tradizionali e metodologie semantiche. L'obiettivo rimane quello di supportare il lavoro degli esperti e non di sostituirne la valutazione, mantenendo la possibilità di ricostruire il collegamento tra i risultati prodotti e le informazioni di origine.

3. Materiali e metodi

Il lavoro è stato organizzato seguendo una sequenza di passaggi successivi. Inizialmente sono stati analizzati i documenti procedurali disponibili e la mappatura del processo CAR-T fornita dal Policlinico San Matteo, utilizzata per individuare le fasi considerate nello studio e associare a ciascuna di esse la documentazione pertinente. Successivamente è stata analizzata la struttura della tabella di rischio fornita come riferimento ed è stato definito lo schema utilizzato per organizzare in modo uniforme le informazioni estratte. A partire da questi elementi è stata sviluppata la pipeline computazionale, articolata nella fase di estrazione strutturata delle informazioni di rischio, nella successiva astrazione e unione dei rischi semanticamente simili e nella costruzione di una mappa di unione per mantenerne la tracciabilità. Infine, i macro-rischi ottenuti sono stati confrontati con il riferimento fornito dall'ospedale mediante metriche tradizionali e due differenti strategie di valutazione semantica, basate rispettivamente su embedding e LLM-as-a-judge. Il flusso complessivo del lavoro è rappresentato schematicamente in Figura 3.1.

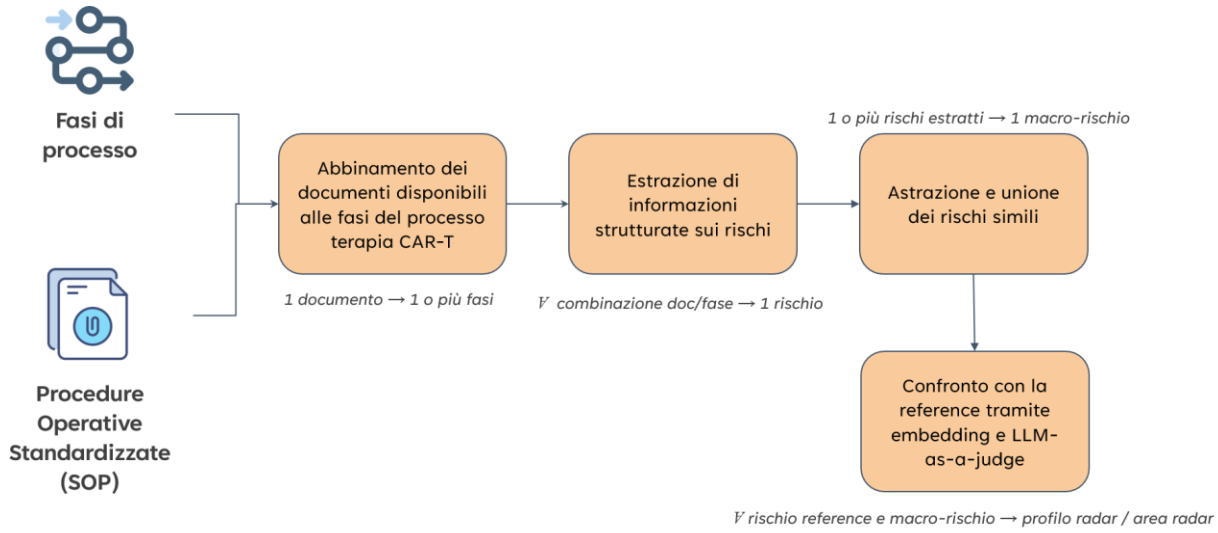


Figura 3.1: Schema generale della pipeline sviluppata per l'estrazione, l'astrazione e il confronto delle informazioni di rischio nel processo CAR-T. Il flusso parte dalle fasi del processo e dai documenti procedurali disponibili prosegue con l'abbinamento documento-fase e l'estrazione strutturata delle informazioni, fino all'astrazione delle righe simili e al confronto con il riferimento tramite embedding e LLM-as-a-judge.

3.1 Documenti procedurali e definizione del processo CAR-T

Il lavoro è stato sviluppato a partire da un insieme di documenti procedurali relativi al percorso di terapia CAR-T, forniti dal Policlinico San Matteo di Pavia. Tali documenti hanno rappresentato il materiale di partenza della pipeline, poiché contenevano le informazioni operative, cliniche, organizzative, logistiche e documentali necessarie per ricostruire i rischi associati alle diverse fasi del processo.

La documentazione disponibile è organizzata in diverse tipologie documentali, identificate tramite sigle interne. In particolare, sono state considerate quattro macrocategorie principali: “ALL”, “IO”, “MSI” e “PP”. La sigla “ALL” indica gli allegati, cioè documenti di supporto alle procedure principali, come moduli, schede o materiali da compilare. La sigla “IO” indica le istruzioni operative, ovvero documenti che descrivono nel dettaglio le modalità di esecuzione di determinate attività. La sigla “MSI” indica invece manuali e specifiche istruzioni, cioè documenti tecnici o operativi contenenti indicazioni puntuali necessarie per lo svolgimento corretto di alcune attività. Infine, la sigla “PP” indica il piano di processo/prodotto, utilizzato per descrivere e organizzare in modo strutturato il processo considerato.

Un documento centrale per l’organizzazione del lavoro è stato “ALL PP 093.1.3 – Mappatura del processo terapia CAR”. Questo documento è stato utilizzato come riferimento per individuare la sequenza delle fasi del processo CAR-T e per orientare l’associazione dei diversi documenti procedurali alle fasi pertinenti. In questo modo, la documentazione non è stata analizzata come un insieme generico di testi, ma è stata ricondotta alla struttura del processo, permettendo di collegare le informazioni estratte alla fase del percorso in cui risultavano più rilevanti. Questa scelta ha permesso di evitare un’estrazione generica e poco contestualizzata, organizzando invece i rischi in base al punto del processo in cui possono manifestarsi. In questo modo, la fase non rappresenta soltanto un’etichetta descrittiva, ma costituisce il riferimento principale per collegare documenti, attività, rischi, cause, conseguenze e opportunità. La suddivisione in fasi è stata mantenuta nelle successive fasi di astrazione e confronto, così da preservare il contesto delle informazioni durante l’intera pipeline.

Prima dell’elaborazione automatica, sono stati esclusi i documenti non informativi rispetto all’obiettivo del lavoro, come materiali prevalentemente grafici che non fornivano contenuto testuale utile per l’estrazione strutturata delle informazioni di rischio.

La documentazione procedurale è stata utilizzata come fonte primaria per l’identificazione delle informazioni di rischio. A partire dai documenti selezionati, la pipeline ha estratto informazioni relative all’attività descritta, alle strutture e agli attori coinvolti, allo scenario di rischio, alle opportunità, alle cause e alle conseguenze. Queste informazioni sono state successivamente organizzate in una tabella strutturata, coerente con lo schema definito per l’analisi del rischio, come riportato nella Figura 3.2, il codice relativo alla definizione dello schema Pydantic è riportato in Appendice A.

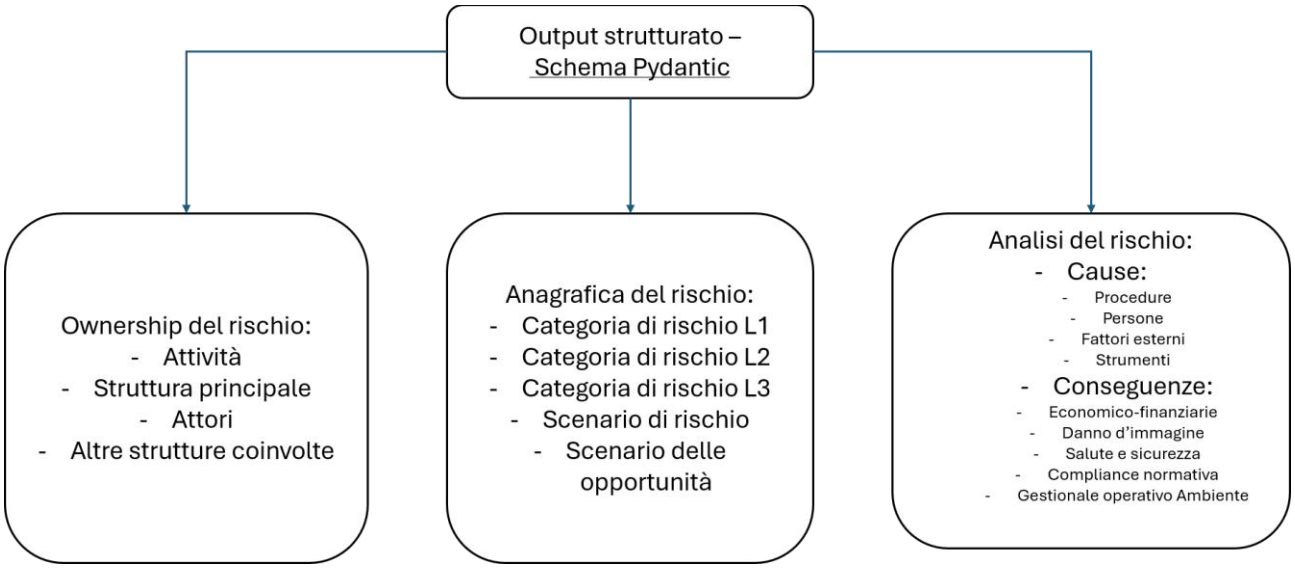


Figura 3.2 - Rappresentazione schematica della struttura Pydantic utilizzata per l'estrazione delle informazioni di rischio

Il processo complessivo riportato nella mappatura comprende ventiquattro fasi. Tuttavia, nella presente tesi il lavoro è stato limitato alle prime dodici. È stata fatta questa scelta per mantenere il perimetro dello studio più controllabile e per concentrare lo sviluppo e la valutazione della pipeline su una porzione significativa del percorso, comprendente le fasi iniziali e intermedie della terapia. Le prime dodici fasi includono infatti attività fondamentali come l'indicazione alla terapia, la valutazione dell'idoneità del paziente, la richiesta del farmaco, la gestione del kit di spedizione, la procedura aferetica, il trasporto e la preparazione del prodotto cellulare. La mappatura delle prime dodici fasi del processo CAR-T è stata riportata nella Figura 3.3.

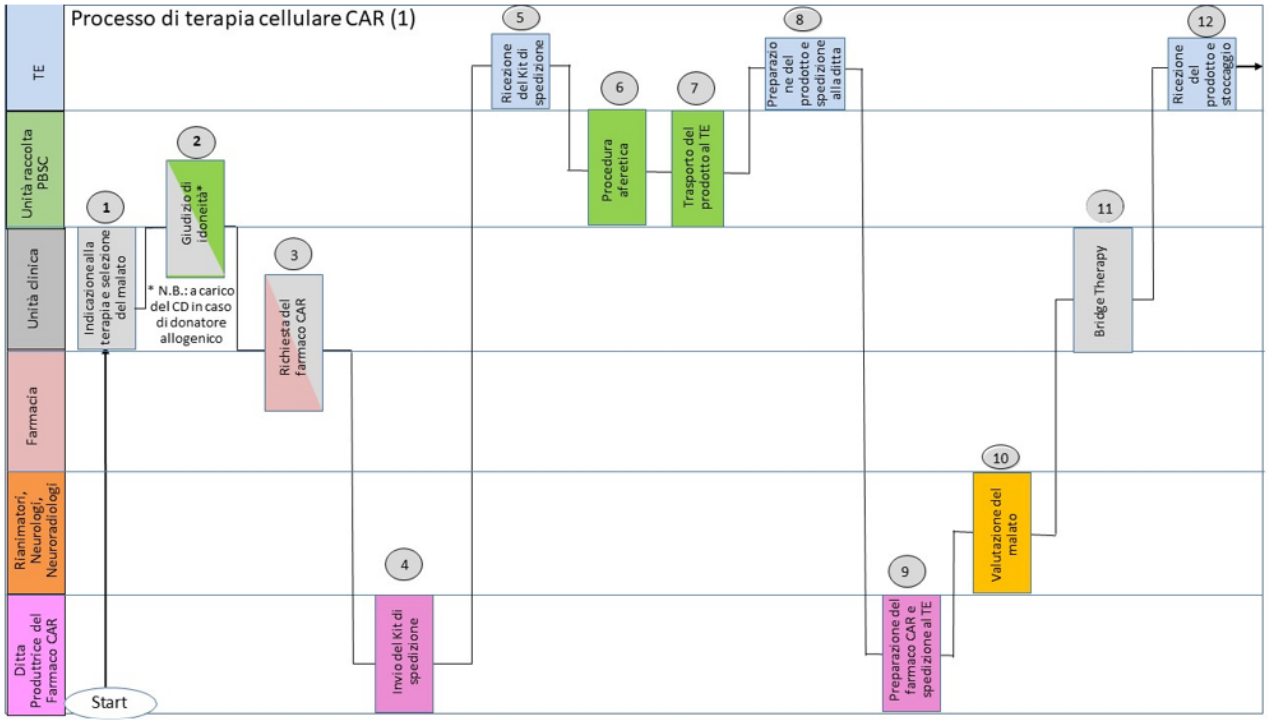


Figura 3.3 — Mappatura delle prime 12 fasi del processo CAR-T e delle principali strutture coinvolte. Lo schema, organizzato per corsie, evidenzia la sequenza delle prime dodici fasi del

percorso e le unità organizzative principalmente coinvolte in ciascun passaggio.

Dopo aver definito le fasi del processo CAR-T e selezionato i documenti procedurali da utilizzare, è stato necessario associare ciascun documento alla fase o alle fasi del percorso a cui risultava maggiormente pertinente. Questo passaggio ha rappresentato una fase metodologica importante, poiché ha permesso di delimitare il contesto di analisi prima dell'estrazione automatica delle informazioni di rischio. I documenti procedurali forniti sono quindi stati associati a una o più fasi del processo CAR-T, in modo che le informazioni estratte potessero essere interpretate in relazione al momento operativo in cui il rischio può manifestarsi.

L'abbinamento tra documenti e fasi è stato necessario perché la documentazione disponibile presentava livelli diversi di specificità. Alcuni documenti descrivevano attività molto circoscritte e specifiche, chiaramente riconducibili a una singola fase del processo, mentre altri contenevano informazioni più generali o trasversali, applicabili a più passaggi del percorso. Per questo motivo, non è stata assunta una corrispondenza univoca tra documento e fase, ma è stata prevista la possibilità che uno stesso documento potesse essere associato a una o anche a più fasi del processo.

L'associazione documento-fase è stata affrontata mediante due strategie. Il primo metodo è stato di tipo manuale ed è stato basato sulla lettura del titolo, del contenuto e della funzione operativa dei documenti. In questa fase, ciascun documento è stato analizzato per individuare le attività descritte, le strutture coinvolte e il punto del percorso CAR-T in cui tali attività risultavano collocabili. L'obiettivo era costruire una mappatura ragionata tra documentazione procedurale e fasi del processo, utilizzando la mappatura ufficiale del percorso CAR-T come riferimento.

La seconda strategia è stata di tipo automatico ed è stata sviluppata tramite codice Python, utilizzando un modello linguistico locale di dimensioni contenute, phi4 (con 14B di parametri). La scelta di utilizzare un modello locale è stata adottata per preservare la riservatezza dei documenti, essendo essi confidenziali. In questo caso, l'obiettivo era valutare la possibilità di sostituire l'abbinamento manuale con una metodologia totalmente automatizzata. Il codice ha utilizzato le informazioni testuali disponibili nei documenti e le ha confrontate con le denominazioni e il contenuto descrittivo delle fasi del processo, al fine di individuare le associazioni più plausibili e produrre una lista di abbinamenti.

L'abbinamento automatico non è stato utilizzato come criterio definitivo per lo sviluppo della pipeline, poiché ha mostrato una copertura più conservativa rispetto all'interpretazione manuale. Per questo motivo, il lavoro è proseguito utilizzando come riferimento principale l'abbinamento manuale, ritenuto più completo e più aderente alla struttura del percorso CAR-T. La procedura automatica è stata quindi considerata come una prova esplorativa e come possibile supporto metodologico, ma non come sostituto della valutazione manuale.

Il risultato di questa fase è stato una mappa di associazione tra documenti e fasi del processo. Questa mappa ha definito, per ogni fase considerata, quali documenti dovessero essere forniti come input alla successiva fase di estrazione strutturata delle informazioni di rischio.

3.2 Analisi dei dati di riferimento

Per poter organizzare in modo uniforme le informazioni estratte dai documenti e rendere successivamente possibile il confronto con la tabella di riferimento fornita dall'ospedale, è stato definito uno schema comune composto da 22 colonne. La struttura è stata costruita a partire dalle informazioni presenti nel riferimento fornito e comprende campi relativi alla tracciabilità, al contesto operativo del rischio, alla sua classificazione, allo scenario, alle cause e alle conseguenze. Lo stesso schema è stato mantenuto nelle diverse fasi della pipeline, in modo che la tabella estratta, la tabella successivamente astratta e il riferimento standardizzato fossero organizzati secondo una struttura coerente. I campi che compongono lo schema sono riportati nella Tabella 3.1.

Macro-area	Campo	Tipo di dato
TRACCIABILITÀ	Codice Documento	Testo
	Step ID	Numero
OWNERSHIP DEL RISCHIO	Fase	Testo
	Attività	Testo
	Struttura principale	Scelta da elenco predefnito
	Attori	Testo
	Altre strutture coinvolte	Scelta da elenco predefnito
ANAGRAFICA DEL RISCHIO	Categoria Rischio Livello 1	Scelta da elenco predefnito
	Categoria Rischio Livello 2	Scelta da elenco predefnito
	Categoria Rischio Livello 3	Scelta da elenco predefnito
	Scenario di rischio	Testo
	Scenario delle opportunità	Testo
ANALISI DEL RISCHIO	Principali cause del rischio – PROCEDURE	Testo
	Principali cause del rischio – PERSONE	Testo
	Principali cause del rischio – FATTORI ESTERNI	Testo
	Principali cause del rischio – STRUMENTI	Testo
	Principali conseguenze del rischio – ECONOMICO-FINANZIARIE	Testo
	Principali conseguenze del rischio – DANNO D'IMMAGINE	Testo
	Principali conseguenze del rischio – SALUTE E SICUREZZA	Testo
	Principali conseguenze del rischio – COMPLIANCE NORMATIVA	Testo
	Principali conseguenze del rischio – GESTIONALE OPERATIVO	Testo
	Principali conseguenze del rischio – AMBIENTE	Testo

Tabella 3.1 Schema delle variabili della tabella di rischio.
I campi sono organizzati nelle tre macro-aree ownership del rischio, anagrafica del rischio e analisi del rischio. Lo schema ha costituito la struttura comune utilizzata per l'intera pipeline

I campi di tracciabilità comprendono *codice_doc*, *step_id* e *fase*. Il campo *codice_doc* consente di mantenere un riferimento alla provenienza della riga o al documento da cui l'informazione è stata estratta. Il campo *step_id* identifica numericamente la fase del processo, mentre il campo *fase* ne riporta la descrizione testuale. Questi campi sono stati fondamentali perché l'intera pipeline è stata organizzata mantenendo la separazione per fase: una riga prodotta per una determinata fase è stata infatti astratta e confrontata solo con righe appartenenti alla stessa fase del processo.

La prima macro-area, definita ownership del rischio, comprende le informazioni che permettono di identificare il contesto operativo e organizzativo del rischio. Rientrano in questa area i campi *attivita*, *struttura_principale*, *attori* e *altre_strutture_coinvolte*. Il campo attività descrive il passaggio operativo a cui il rischio è associato. I campi relativi a strutture e attori permettono invece di individuare l'unità organizzativa principalmente coinvolta, le figure professionali interessate e le eventuali altre strutture che partecipano al processo. Questa macro-area consente quindi di collegare ogni rischio non soltanto a una fase del percorso, ma anche alle responsabilità e ai soggetti coinvolti nella sua gestione.

La seconda macro-area è rappresentata dall'anagrafica del rischio. Questa comprende i campi dedicati alla classificazione del rischio e alla descrizione dello scenario. In particolare, la classificazione è articolata in tre livelli: *categoria_rischio_livello_1, categoria_rischio_livello_2* e *categoria_rischio_livello_3.* I primi due livelli sono stati utilizzati per descrivere il rischio in modo progressivamente più specifico, passando da una categoria generale a una sottocategoria più dettagliata. Il terzo livello, invece, non è sempre presente: è stato compilato solo per alcune categorie di rischio di livello 2, quando era disponibile o necessario un ulteriore grado di dettaglio. Questa impostazione ha permesso di mantenere una classificazione flessibile, evitando di forzare una sottoclassificazione anche nei casi in cui non fosse prevista dalla struttura del riferimento.

Categoria di rischio L1	**Categoria di rischio L2**	**Categoria di rischio L3**
CLINICO-SANITARI	Anestesiologico	
	Assistenziale	
	Atti di autolesione e tentativi di suicidio	
	Caduta	
	Chirurgico	
	Correlato alla procedura	
	Diagnostico	
	Gestione / redazione documenti	
	Identifcazione del paziente	
	Infezione correlata all'assistenza	
	Ostetrico e neonatale, inclusi i trigger	
	Prevenzione	
	Sperimentazioni cliniche	
	Terapeutico	
	Trasfusionale	
ESTERNI	Contesto socio-economico nazionale e regionale	
	Eventi naturali o accidentale	
	Evoluzione del contesto normativo	
	Gestione Terze Parti	

	Illeciti esterni	
	Sicurezza informatica	
FINANZIARI	Contabilità e reporting fnanziario	
	Erariale patrimoniale	
	Fiscale	
	Liquidità e credito	
	Tassi d'interesse	
STRATEGICI	Comunicazione e relazioni istituzionali	
	Governance	
	Immagine/Reputazione	
	Investimenti e patrimonio	
	Pianifcazione strategica	
	Sistema di Controllo Interno	
OPERATIVI	Attività, processi e procedure	
	Asset infrastrutturali e tecnologia	
	Business Continuity	
	Comunicazione e relazioni	
	Continuità e coordinamento percorsi assistenziali	
	Edifci e spazi comuni	
	Gestione apparecchiature sanitarie	
	Gestione farmaci e dispositivi medici	
	Illeciti interni	
	Persone e cultura	
	Informativa e reporting	Informativa strategica / di programmazione
		Misurazione delle performance
		Informativa interna ed esterna
	Salute, sicurezza e ambientale	Atti di violenza su operatori
		Rischio biologico
		Rischio chimico
		Rischio fsico
		Rischio da atmosfera esplosiva
		Rischio incendio
		Rischio da infortuni generici
		Rischio da movimentazione di carichi e pazienti
		Rischio radiazioni
		Rischio sociale
		Rischio stress lavoro correlato
		Rischio da videoterminale
COMPLIANCE	Anticorruzione	
	Codice etico	
	Contrattualistica e controversie legali	
	Normativa (regionale, nazionale, comunitaria)	

	Regolamenti interni	
	Sicurezza delle informazioni e tutela privacy	

Tabella 3.2 – Livelli di rischio gerarchici delle categorie di rischio utilizzata nella fase di estrazione, articolata nei livelli L1, L2 e, dove previsto, L3.

All'interno dell'anagrafica del rischio rientrano anche i campi *scenario_di_rischio* e *scenario_delle_opportunita*. Il primo descrive la situazione indesiderata che potrebbe verificarsi durante una determinata fase del processo, mentre il secondo riporta eventuali opportunità di miglioramento, prevenzione o controllo associate a quel rischio. Questi due campi rappresentano una componente centrale della tabella, perché descrivono il significato sostanziale del rischio e risultano quindi particolarmente rilevanti nelle successive fasi di astrazione e confronto semantico.

La terza macro-area è costituita dall'analisi del rischio, che comprende le cause e le conseguenze associate a ciascun elemento. Le principali cause sono state suddivise in quattro categorie al fine di distinguere tra cause legate a procedure o istruzioni operative, cause riconducibili a errore umano, formazione o responsabilità degli operatori, cause dipendenti da fattori esterni e cause associate a dispositivi, strumenti, materiali o sistemi informatici. Le conseguenze sono state invece suddivise in sei sottocategorie al fine di descrivere gli effetti potenziali del rischio su più dimensioni: economica, reputazionale, clinica, normativa, gestionale-operativa e ambientale. In questo modo, la tabella non si limita a riportare il rischio in sé, ma consente anche di analizzarne le possibili ricadute sul paziente, sull'organizzazione e sul processo.

Nel complesso, lo schema della tabella è stato definito per garantire uniformità, tracciabilità e confrontabilità tra le diverse componenti della pipeline. La stessa struttura è stata utilizzata per le informazioni estratte dai documenti procedurali, per i macro-rischi prodotti nella fase di astrazione e per il riferimento standardizzato. Questa scelta ha permesso di mantenere una corrispondenza diretta tra i campi delle diverse tabelle e di predisporre i dati per la successiva fase di confronto automatico.

Lo schema completo delle variabili considerate è riportato in Tabella 3.1.

Prima di procedere con il confronto automatico, è stata efettuata una fase di riorganizzazione del riferimento. Poiché erano presenti alcuni campi non rilevanti ai fni dell'analisi, mentre altre informazioni utili erano riportate in forma aggregata, è stato necessario riorganizzare la tabella. Un primo intervento ha riguardato la selezione dei campi da utilizzare. Alcune colonne presenti nella tabella originaria non erano necessarie ai fni del confronto automatico e sono state quindi escluse. Un secondo intervento ha riguardato invece la separazione di alcune informazioni che nella tabella fornita erano riportate insieme. In particolare, le informazioni relative agli attori e alla struttura principale comparivano in forma aggregata. Per rendere possibile il confronto automatico, le informazioni relative agli attori e alla struttura principale sono state quindi separate e ricondotte alle colonne specifiche struttura_principale e attori. Un ulteriore intervento ha riguardato le conseguenze del rischio. Nella tabella originaria, queste erano riportate in un'unica colonna complessiva; nella versione standardizzata sono state invece suddivise secondo le sei categorie previste dallo schema della pipeline: conseguenze economico-finanziarie, danno d'immagine, salute e sicurezza, compliance normativa, gestionale-operativo e ambiente.

Infine, sono stati uniformati i nomi delle colonne, l'ordine dei campi e la disposizione delle informazioni, così da rendere la tabella risultante dalla pipeline compatibile con il riferimento.

3.3 Pipeline computazionale

La pipeline computazionale è stata articolata in tre passaggi principali:

1. estrazione strutturata delle informazioni di rischio;
2. astrazione e unione dei rischi simili;
3. costruzione di una mappa di unione per mantenere la tracciabilità tra righe di input e macro-rischi prodotti.

Al termine di queste fasi, la tabella risultante è stata confrontata con il riferimento descritto nel paragrafo 3.2, utilizzando metriche tradizionali e approcci semantici basati su embedding e LLM-as-a-judge. Le tre fasi della pipeline computazionale sono descritte nei paragrafi successivi, mentre la metodologia utilizzata per il confronto con il riferimento è riportata nel paragrafo 3.4.

I principali passaggi della pipeline computazionale sono riassunti schematicamente in Figura 3.4.

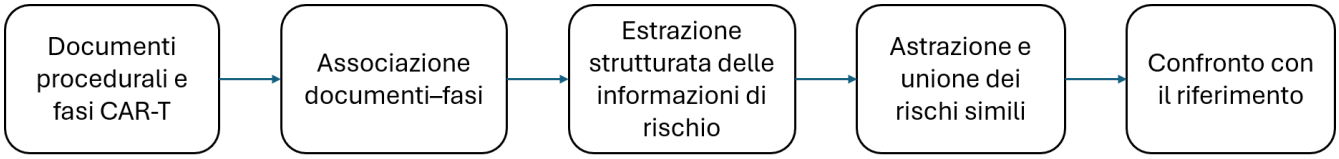


Figura 3.4 - Schema dei principali passaggi della pipeline computazionale.

3.3.1 Estrazione strutturata delle informazioni di rischio

Dopo aver definito lo schema della tabella di rischio e aver associato i documenti procedurali alle rispettive fasi del processo CAR-T, è stata sviluppata la fase di estrazione strutturata delle informazioni. L'obiettivo di questa fase è trasformare il contenuto testuale dei documenti in righe tabellari organizzate secondo lo schema descritto nel paragrafo precedente nella Figura 3.2. In questo modo, informazioni inizialmente presenti in forma discorsiva e distribuita all'interno di documenti differenti sono state ricondotte a una struttura comune e confrontabile.

L'estrazione è stata effettuata tramite modelli linguistici installati ed eseguiti localmente utilizzando Ollama, un ambiente software che consente di scaricare, gestire ed eseguire modelli linguistici direttamente sul computer in locale, senza necessità di inviare i dati a servizi esterni. La scelta di lavorare con modelli locali è stata legata principalmente alla natura della documentazione analizzata. I documenti forniti contenevano infatti procedure interne, informazioni organizzative, dettagli operativi e responsabilità relative al percorso CAR-T e dovevano quindi essere considerati documenti riservati. Per questo motivo è stato preferito un approccio locale, evitando l'invio dei documenti a servizi esterni e mantenendo l'elaborazione all'interno di un ambiente controllato.

La fase di estrazione è stata sviluppata in Python. I documenti sono stati forniti al modello sulla base dell'abbinamento documento-fase descritto nel paragrafo 3.1. Per ogni fase del processo, il modello riceve in input l'identificativo della fase, la descrizione della fase stessa, i documenti procedurali associati e lo schema strutturato da compilare, attraverso il system prompt. Un esempio di prompt utilizzato nell'approccio single shot è riportato in Appendice B. Questa impostazione permette di mantenere l'estrazione contestualizzata rispetto alla fase del processo CAR-T, evitando che il modello analizzi la documentazione in modo generico o non coerente con il momento operativo considerato.

Per vincolare la struttura dell'output è stata utilizzata la libreria Python Pydantic, uno strumento che consente di definire modelli di dati specificando campi, tipi e vincoli che devono essere rispettati. A partire da questi modelli è possibile generare uno schema strutturato dell'output atteso e verificarne automaticamente la conformità. Nel presente lavoro, Pydantic è stato utilizzato per definire i campi che il modello linguistico doveva restituire, in coerenza con le tre macro-aree della tabella descritte nel paragrafo 3.2. In questo modo, la risposta del modello non viene lasciata completamente libera, ma viene organizzata in un file JSON coerente con la struttura della tabella di rischio finale. Questo ha permesso di ottenere output ordinati e direttamente utilizzabili nelle successive fasi di elaborazione automatica, oltre a rendere l'output leggibile e facilmente elaborabile. Pydantic ha permesso di controllare che i campi richiesti fossero presenti e coerenti con lo schema previsto. Alcuni campi sono stati definiti come scelte vincolate, ad esempio struttura_principale e le categorie dei livelli di rischio. Altri campi, come scenario, cause e conseguenze, sono stati lasciati a risposta libera, ma guidati da descrizioni operative dettagliate. Questa impostazione ha ridotto la variabilità dell'output.

Sono stati testati tre modelli linguistici locali: phi4 (14B di parametri), qwen3.5 (9B) e gpt-oss (20B). Phi4 è un modello della famiglia Phi sviluppata da Microsoft, progettata per ottenere buone capacità di comprensione e ragionamento mantenendo dimensioni relativamente contenute. Qwen3.5 appartiene invece alla famiglia di modelli Qwen sviluppata da Alibaba ed è un modello open-weight progettato per compiti di comprensione, generazione e ragionamento. Gpt-oss-20b è infine un modello open-weight sviluppato da OpenAI, orientato in particolare a compiti di ragionamento e al rispetto di istruzioni strutturate. Nel presente lavoro i tre modelli sono stati eseguiti localmente tramite Ollama.

La scelta di confrontare più modelli è stata introdotta per osservare eventuali differenze nella qualità degli output prodotti, sia dal punto di vista della coerenza semantica con i documenti sia dal punto di vista della stabilità strutturale della risposta. In particolare, l'attenzione è stata posta sulla capacità

dei modelli di rispettare lo schema imposto, mantenere la corretta associazione con la fase analizzata, produrre righe leggibili e restituire informazioni effettivamente pertinenti al contenuto documentale.

Oltre al confronto tra modelli, sono state sperimentate due diverse strategie di estrazione: un approccio *single shot* e un approccio suddiviso in fasi o "*divided"*. Questa scelta è stata introdotta perché l'estrazione dell'intera tabella di rischio rappresenta un compito complesso, che richiede al modello di gestire contemporaneamente informazioni appartenenti a categorie differenti. Il confronto tra le due strategie ha permesso quindi di valutare se fosse più efficace richiedere al modello la generazione immediata di una riga completa oppure suddividere il compito in passaggi più piccoli.

Nel primo approccio, definito *single shot*, il modello viene interrogato una sola volta per ciascuna combinazione documento-fase. In un'unica chiamata vengono richieste contemporaneamente tutte le informazioni della tabella, suddivise nelle tre macro-aree. Questo approccio permette di mantenere una visione unitaria dell'output, perché tutte le informazioni sono generate nello stesso passaggio, ma richiede al modello di gestire contemporaneamente campi differenti per struttura e contenuto.

Nel secondo approccio, definito *divided*, l'estrazione viene invece suddivisa in più passaggi. Il modello viene interrogato separatamente per produrre prima le informazioni relative all'ownership, poi quelle relative all'anagrafica del rischio e infine quelle relative all'analisi del rischio. In questo caso sono quindi generati più JSON parziali, ciascuno riferito a una specifica macro-area. Questa strategia è stata pensata per ridurre il carico computazionale della singola chiamata al modello linguistico e per verificare se una scomposizione del compito potesse rendere l'output più preciso.

Per ciascun modello sono state quindi testate entrambe le strategie. In questo modo, la fase metodologica non si è limitata all'applicazione di un unico modello, ma ha previsto un confronto esplorativo tra più combinazioni modello-strategia. Tale confronto è stato utilizzato per valutare quale configurazione fosse più adatta alla pipeline, considerando sia la qualità del contenuto estratto sia la capacità tecnica di produrre output strutturati e utilizzabili nelle fasi successive.

Il prompt utilizzato per l'estrazione contiene istruzioni specifiche volte a limitare la generazione di informazioni non supportate dai documenti. Il modello è infatti istruito a estrarre e riorganizzare le informazioni presenti nel materiale fornito, evitando di introdurre rischi, cause o conseguenze non deducibili dal testo. Allo stesso tempo, non è richiesta una semplice copia letterale dei documenti, ma una riformulazione strutturata e coerente con lo schema della tabella di rischio. Questo passaggio è necessario perché i documenti procedurali richiedono un lavoro di interpretazione e ricollocazione delle informazioni.

Per migliorare la contestualizzazione degli output, nel prompt sono stati inseriti anche elementi di supporto, tra cui i livelli di rischio e la loro struttura gerarchica, riportata in Tabella 3.2 e informazioni relative all'organigramma del processo CAR-T. I livelli di rischio sono stati utilizzati per guidare la classificazione del rischio nei livelli previsti dalla tabella, mentre l'organigramma ha supportato l'identificazione delle strutture e degli attori coinvolti. Questi elementi non sono stati utilizzati per generare informazioni autonome, ma come contesto interpretativo per orientare il modello nei casi in cui il documento risultasse ambiguo o richiedesse una lettura organizzativa più ampia.

La procedura di estrazione è stata applicata alle prime dodici fasi del processo CAR-T, utilizzando come riferimento l'abbinamento manuale documento-fase descritto nel paragrafo 3.1. Per ogni combinazione tra fase e documento, il modello ha prodotto una o più righe di rischio, successivamente raccolte in un file tabellare. Ogni riga è stata collegata allo step_id, alla fase corrispondente e al documento di origine, in modo da mantenere la tracciabilità dell'intero processo.

La fase di estrazione ha prodotto una tabella iniziale ricca ma potenzialmente ridondante. Documenti diversi possono infatti descrivere aspetti simili dello stesso rischio, oppure uno stesso rischio può

essere richiamato con formulazioni differenti in più parti della documentazione. Questa ridondanza non è stata considerata un errore della procedura, ma una conseguenza attesa dell'analisi di documenti complessi e parzialmente sovrapposti. In questa fase, l'obiettivo principale è recuperare il maggior numero possibile di informazioni potenzialmente rilevanti, rimandando alla fase successiva il compito di accorpare le righe simili e ottenere una rappresentazione più sintetica.

3.3.2 Astrazione e unione di righe simili

Dopo la fase di estrazione strutturata, la tabella prodotta dalla pipeline contiene un insieme di righe di rischio associate alle diverse fasi del processo CAR-T. Questo output rappresenta una prima organizzazione delle informazioni presenti nei documenti procedurali, ma risultava ancora caratterizzato da ridondanza, eccessiva specificità e alcune incoerenze di classificazione. Infatti, documenti diversi possono descrivere aspetti simili dello stesso rischio, oppure uno stesso rischio può essere richiamato più volte con formulazioni differenti all'interno della documentazione analizzata.

L'obiettivo della fase di astrazione e unione è duplice: da un lato, ridurre la ridondanza della tabella iniziale, trasformando più righe estratte in un numero minore di macro-rischi, dall'altro, rivalutare le informazioni prodotte nella fase precedente, correggendo o armonizzando eventuali classificazioni incoerenti, scenari troppo specifici, formulazioni ridondanti o elementi non perfettamente allineati alla fase del processo. Con il termine macro-rischio si intende una riga astratta che riassume e integra più righe estratte quando queste descrivono scenari di rischio semanticamente sovrapponibili o fortemente correlati.

L'input della fase di astrazione è rappresentato dalla tabella ottenuta dalla fase precedente. In particolare, per la fase di astrazione è stata utilizzata come input la tabella ottenuta mediante il modello phi4:14b nella configurazione divided. Tale configurazione è stata scelta come output di partenza della fase successiva in quanto risultava la più adatta a fornire una rappresentazione strutturata e completa delle informazioni necessarie per le successive elaborazioni della pipeline. A differenza della fase di estrazione, in questo passaggio non sono stati forniti al modello i documenti procedurali originali o le SOP complete, ma solo le righe tabellari già prodotte, rielaborate e organizzate secondo lo schema di rischio a 22 colonne.

Questa differenza ha avuto un impatto importante anche sulla scelta del modello. Nella fase di estrazione è stato preferito l'utilizzo di modelli locali, poiché l'input è costituito direttamente dalla documentazione procedurale interna. Nella fase di astrazione, invece, l'input non contiene più le SOP nella loro forma originale, ma informazioni già rielaborate dalla pipeline e trasformate in righe strutturate. Per questo motivo, è stato possibile utilizzare un modello non installato localmente, mantenendo comunque una maggiore cautela rispetto al trattamento dei documenti originali.

Per la fase di astrazione è stato utilizzato il modello gemini-3.5-flash. La scelta di questo modello è legata alla necessità di svolgere un compito più complesso rispetto alla semplice estrazione di informazioni. L'astrazione richiedeva infatti di riconoscere righe tra loro simili, raggrupparle, sintetizzare informazioni parzialmente sovrapposte, correggere classificazioni instabili e produrre una formulazione più generale e coerente del rischio. Rispetto ai modelli locali utilizzati nella fase di estrazione, l'utilizzo di un modello più avanzato e accessibile tramite API permette di impiegare una risorsa più potente, più veloce e più adatta a compiti di ragionamento, sintesi e normalizzazione testuale. Il system prompt utilizzato per guidare il modello nella fase di astrazione è riportato in Appendice D.

Dal punto di vista metodologico, il codice di astrazione è stato impostato secondo una strategia di raggruppamento preliminare e sintesi. Il modello viene istruito a leggere tutte le righe input di una determinata fase, individuare gruppi di righe semanticamente affini e generare un macro-rischio per ciascun gruppo. Questa impostazione è introdotta per fare in modo che ogni riga astratta derivi direttamente da un insieme riconoscibile di righe iniziali. L'output finale rappresenta quindi il contenuto comune dei rischi associati ad una fase del processo CAR-T, integrando eventuali elementi complementari senza introdurre informazioni aggiuntive.

Anche nella fase di astrazione è stata mantenuta la struttura tabellare a 22 colonne. Questa scelta ha permesso di produrre un output direttamente utilizzabile nella fase successiva di confronto con il riferimento, senza richiedere ulteriori trasformazioni della struttura dati.

Dal punto di vista tecnico, anche l'output dell'astrazione è stato richiesto in formato JSON strutturato e validato tramite schemi Pydantic. Questo ha consentito di controllare che ogni macro-rischio rispetti lo schema previsto, che le categorie di rischio siano coerenti e che non vengano introdotti campi aggiuntivi non compatibili con la tabella finale.

Nel processo di astrazione, al modello è stato richiesto di sintetizzare le righe semanticamente affini mantenendo le informazioni necessarie a distinguere i diversi rischi. Ogni macro-rischio doveva quindi rappresentare in modo coerente il contenuto delle righe di origine, integrando eventuali elementi complementari senza introdurre informazioni non supportate dagli input.

3.3.3 Mappa di unione e tracciabilità dell'astrazione

Oltre alla tabella finale dei macro-rischi, nella fase di astrazione è richiesto al modello di produrre anche una mappa di unione. Questo output è stato introdotto per mantenere traccia del collegamento tra le righe estratte iniziali e le righe astratte generate dal modello. Poiché l'astrazione prevede l'accorpamento di più righe simili in un unico macro-rischio, è necessario conservare un'informazione esplicita su quali righe input siano state utilizzate per costruire ciascuna riga finale.

La mappa di unione è stata quindi pensata come uno strumento pratico di controllo. Per ogni macro-rischio generato, il modello doveva indicare le righe di partenza confluite in quella riga astratta e riportare un breve criterio di unione, cioè una motivazione della logica con cui quelle righe sono state accorpate. In questo modo, non viene salvato soltanto il risultato finale dell'astrazione, ma anche il collegamento con le informazioni originarie da cui quel risultato deriva.

La mappa di unione è stata salvata in un file separato rispetto alla tabella finale dei macro-rischi poiché serve per controllare e interpretare il processo di astrazione. In particolare, la mappa permette di verificare se il modello ha accorpato righe coerenti tra loro. La mappa è utile anche per controllare che nessuna riga input venga persa durante l'astrazione. Infatti, quando si riduce una tabella più estesa a una tabella più sintetica, esiste il rischio che alcune informazioni vengano escluse o non vengano riportate nel risultato finale. Per questo motivo, il codice prevede un controllo sulla copertura delle righe input, verificando che le righe della fase analizzata siano state assegnate ad almeno un macro-rischio.

Nel caso in cui una riga non sia stata inserita nella mappa, il codice prevede automaticamente un tentativo di correzione tramite Gemini. Se la copertura risulta ancora incompleta, una procedura assegna la riga mancante al gruppo più simile. Questo passaggio non ha lo scopo di creare nuovi macro-rischi inutili, ma di evitare che informazioni potenzialmente rilevanti spariscano durante la sintesi. In questo modo, la riduzione della tabella non avviene tramite eliminazione incontrollata, ma attraverso un processo tracciabile.

3.4 Metodologia di confronto e metriche di valutazione

L'ultima fase del lavoro ha riguardato il confronto automatico tra la tabella prodotta dalla pipeline e il riferimento fornito dall'ospedale. L'obiettivo è valutare in modo sistematico quanto i macro-rischi generati attraverso il processo di astrazione siano coerenti con i rischi presenti nella tabella di riferimento.

Per ogni riga del riferimento, il codice confronta tutte le righe astratte appartenenti alla stessa fase del processo CAR-T. In questo modo, ogni rischio di riferimento viene confrontato esclusivamente con i macro-rischi prodotti per la fase corrispondente, evitando confronti tra momenti differenti del percorso. Ad esempio, una riga del riferimento relativa alla procedura aferetica può essere confrontata soltanto con i macro-rischi astratti appartenenti alla medesima fase.

Per ogni coppia formata da una riga del riferimento e una riga astratta appartenente alla stessa fase, il codice calcola un insieme di punteggi riferiti alle diverse componenti della tabella. Il confronto non è quindi basato su un'unica colonna o su un singolo valore di similarità, ma tiene conto di più elementi: attività, strutture coinvolte, attori, classificazione tassonomica, scenario di rischio, scenario delle opportunità, cause e conseguenze. In questo modo, ogni coppia è descritta attraverso un profilo multidimensionale, permettendo di valutare separatamente le diverse componenti del rischio.

Questa impostazione risulta particolarmente utile perché due righe possono mostrare una buona corrispondenza rispetto ad alcuni aspetti e una corrispondenza più debole rispetto ad altri. Una riga astratta, ad esempio, può descrivere correttamente lo scenario di rischio ma presentare una classificazione del rischio diversa da quella riportata nel riferimento; allo stesso modo, può individuare correttamente attori e strutture coinvolte ma riportare cause o conseguenze più generali. La valutazione separata delle diverse dimensioni consente quindi di mantenere visibile questa informazione durante tutto il processo di confronto.

I campi caratterizzati da valori predefiniti della tabella sono stati valutati mediante metriche tradizionali, mentre per i campi caratterizzati da contenuto testuale narrativo sono state utilizzate due strategie alternative di valutazione semantica: una basata su embedding e una basata su LLM-as-a-judge.

Per il campo relativo alle strutture coinvolte è stato utilizzato l'indice di Jaccard (24). Prima del confronto, la struttura_principale e le altre_strutture_coinvolte sono state riunite in un unico insieme per ciascuna riga. Questa scelta è stata effettuata perché, ai fini del confronto, risulta più importante verificare la presenza complessiva delle strutture interessate dal rischio piuttosto che penalizzare rigidamente una differenza nella loro classificazione come struttura principale o secondaria. Di conseguenza, se una determinata struttura risulta principale nel riferimento ma compare tra le altre strutture coinvolte nella riga risultante dalla pipeline, la corrispondenza non viene automaticamente considerata errata.

L'indice di Jaccard tra due insiemi A e B può essere espresso come:

$$J(A,B) = \frac{|\ A \cap B\ |}{|\ A \cup B\ |}$$

Il valore ottenuto è compreso tra 0 e 1: un valore pari a 1 indica che i due insiemi coincidono, mentre un valore pari a 0 indica l'assenza di elementi condivisi.

Ad esempio, se nella riga astratta sono presenti le strutture TE, Unità raccolta PBSC e Farmacia, mentre nella riga di riferimento sono presenti TE, Unità raccolta PBSC e Ditta produttrice del farmaco CAR, le due righe condividono due strutture: TE e Unità raccolta PBSC. L'unione complessiva contiene invece quattro strutture: TE, Unità raccolta PBSC, Farmacia e Ditta produttrice del farmaco CAR. Lo score di Jaccard sarà quindi:

$$J = \frac{2}{4} = 0{,}5$$

In questo caso, il valore 0,5 indica una corrispondenza parziale: le due righe condividono alcune strutture, ma non coincidono completamente. Se invece le stesse strutture fossero presenti in entrambe le righe, anche con diversa collocazione tra struttura principale e altre strutture coinvolte, lo score potrebbe risultare pari a 1. Questo permette di valutare la presenza effettiva delle strutture coinvolte nel rischio, senza dare troppo peso alla loro posizione formale nella tabella.

Per la classificazione tassonomica del rischio è stato invece utilizzato uno score gerarchico basato sui tre livelli L1, L2 e L3. La scelta di uno score crescente in base ai livelli classificati correttamente, deriva dal fatto che i livelli di rischio non rappresentano un campo testuale libero, ma una classificazione organizzata gerarchicamente. Una discrepanza nel primo livello rappresenta quindi un errore più rilevante rispetto a una differenza limitata al livello più specifico. Lo score è stato assegnato secondo una logica progressiva: se il livello 1 risulta differente o assente, il punteggio è pari a 0; se coincide solamente il livello 1, il punteggio è 0,33; se coincidono il livello 1 e il livello 2, ma non il livello 3, il punteggio è 0,66; infine, se tutti i livelli risultano coerenti, oppure se il livello 3 è correttamente assente quando non previsto, il punteggio è pari a 1. Questa impostazione permette quindi di distinguere una classificazione appartenente a una macrocategoria completamente diversa da una discrepanza limitata al livello tassonomico più specifico.

Per gli attori è stato adottato un approccio misto. In primo luogo, è stata calcolata la sovrapposizione lessicale tramite indice di Jaccard, analogamente a quanto effettuato per le strutture. Tuttavia, questo campo presenta una maggiore variabilità linguistica, poiché lo stesso ruolo può essere indicato attraverso formulazioni differenti. Per questo motivo, la componente deterministica è stata integrata con una misura di similarità semantica ottenuta tramite embedding. Il punteggio finale relativo agli attori è stato calcolato attribuendo un peso del 70% alla similarità embedding e del 30% all'indice di Jaccard. In questo modo è stato possibile tenere conto sia della coincidenza esplicita dei ruoli sia della presenza di descrizioni differenti ma semanticamente vicine.

Tutti i punteggi prodotti dalle metriche tradizionali sono stati espressi su una scala compresa tra 0 e 1, così da poter essere successivamente inseriti nello stesso profilo multidimensionale.

Per le componenti testuali più descrittive è stata invece implementata una prima versione del confronto basata su embedding semantici (17). Questo approccio è stato applicato ai campi relativi ad attività, scenario di rischio, scenario delle opportunità, cause del rischio e conseguenze del rischio. Tali campi possono infatti descrivere concetti analoghi attraverso formulazioni anche molto diverse, per cui un confronto basato esclusivamente sulla coincidenza delle parole sarebbe risultato eccessivamente rigido.

Nel presente lavoro è stato utilizzato il modello BAAI/bge-m3, un modello di embedding multilingue, scelta particolarmente adatta poiché le tabelle di rischio e i dati elaborati nella pipeline sono in lingua italiana (18). Il modello è stato utilizzato mediante la libreria sentence-transformers, attraverso la quale ciascun testo veniva trasformato in una rappresentazione vettoriale.

Per ogni coppia riferimento–riga astratta appartenente alla stessa fase, sono calcolati gli embedding dei campi testuali da confrontare. La similarità tra le rispettive rappresentazioni vettoriali è poi quantificata attraverso la similarità coseno:

$$\cos(A, B) = \frac{A \cdot B}{\| A \| \| B \|}$$

dove Ae B rappresentano i vettori associati ai due testi confrontati (25). Il numeratore rappresenta il prodotto scalare tra i vettori, mentre il denominatore normalizza il risultato rispetto alla loro lunghezza. Il punteggio utilizzato nella pipeline è stato riportato su una scala compresa tra 0 e 1, dove valori maggiori indicano una maggiore similarità semantica.

Per le cause del rischio, il confronto tramite embedding è stato effettuato separatamente sulle quattro sottocategorie previste dalla tabella: procedure, persone, fattori esterni e strumenti. Analogamente, per le conseguenze sono state confrontate separatamente le sei componenti: economico-finanziarie, danno d'immagine, salute e sicurezza, compliance normativa, gestionale-operativo e ambiente. In questo modo, il confronto manteneva il livello di dettaglio previsto dallo schema della tabella invece di considerare cause e conseguenze come due singoli blocchi testuali.

Parallelamente è stata implementata una seconda versione del confronto semantico basata su LLM-as-a-judge (19). In questo caso, i campi testuali non sono confrontati attraverso la distanza tra rappresentazioni vettoriali, ma sono forniti a un modello linguistico incaricato di esprimere direttamente un giudizio sulla loro corrispondenza concettuale.

Il modello utilizzato come valutatore è stato gpt-oss-20b. Per ogni coppia riferimento–riga astratta, il modello riceve i contenuti dei campi da confrontare insieme a criteri espliciti di valutazione. Per l'attività è valutato il grado di corrispondenza tra le attività descritte nelle due righe; per lo scenario di rischio il modello deve stabilire se il macro-rischio astratto include o rappresenta il rischio specifico riportato nel riferimento; per lo scenario delle opportunità è valutata la coerenza tra le eventuali azioni migliorative o mitigative (26). Cause e conseguenze sono invece considerate rispetto al grado di copertura delle informazioni presenti nelle due righe.

L'LLM assegna uno score intero compreso tra 0 e 5, dove 0 indica assenza di corrispondenza e 5 una corrispondenza molto forte o pienamente equivalente. I valori intermedi rappresentano differenti gradi di corrispondenza parziale. Per cause e conseguenze, il giudizio è prodotto separatamente per le rispettive sottocategorie previste dalla tabella.

Poiché le altre metriche della pipeline sono espresse su una scala compresa tra 0 e 1, gli score generati dall'LLM sono stati normalizzati dividendo ogni valore per 5:

$$S_{\text{norm}} = \frac{S_{\text{judge}}}{5}$$

In questo modo, i punteggi prodotti dal modello sono confrontabili direttamente con quelli ottenuti all'interno dello stesso sistema di valutazione multidimensionale.

Oltre allo score numerico, il modello produce anche una breve motivazione del giudizio. Tale motivazione non entra direttamente nel calcolo della valutazione multidimensionale, ma viene mantenuta tra gli output del confronto come strumento di supporto all'interpretazione. Essa permette infatti di comprendere perché una coppia sia stata considerata più o meno coerente e risulta utile sia per la revisione qualitativa dei confronti sia per l'eventuale affinamento del prompt utilizzato. Il system prompt utilizzato per guidare il modello nella valutazione delle coppie riferimento–riga astratta è riportato integralmente in Appendice E.

La struttura generale delle metriche utilizzate nei due approcci può quindi essere riassunta come mostrato in Tabella 3.3.

Componente valutata	Metodi utilizzati per il confronto
Struttura principale + altre strutture coinvolte	Jaccard

Attori	70% embedding + 30% Jaccard
Livelli di rischio L1–L2–L3	Score gerarchico
Attività	Embedding, LLM-as-a-judge
Scenario di rischio	Embedding, LLM-as-a-judge
Scenario delle opportunità	Embedding, LLM-as-a-judge
Cause del rischio	Embedding, LLM-as-a-judge
Conseguenze del rischio	Embedding, LLM-as-a-judge

Tabella 3.3 — Metodi utilizzati per il confronto delle diverse componenti della tabella di rischio

Dopo il calcolo delle metriche tradizionali e semantiche, ogni coppia riferimento–riga astratta è stata rappresentata attraverso un profilo multidimensionale mediante radar plot. L'obiettivo è evitare di ridurre immediatamente il confronto a un singolo valore complessivo, mantenendo invece visibili le diverse componenti che contribuiscono alla qualità del match.

Il radar plot è composto da otto dimensioni: attività, attori, strutture, livelli di rischio, scenario di rischio, scenario delle opportunità, cause e conseguenze. Ogni asse assume un valore compreso tra 0 e 1, dove valori maggiori indicavano una maggiore corrispondenza tra la riga astratta e quella di riferimento per la specifica componente.

La dimensione relativa alle strutture corrisponde al punteggio Jaccard ottenuto considerando congiuntamente struttura principale e altre strutture coinvolte. La dimensione dei livelli di rischio corrisponde invece allo score gerarchico sui tre livelli di classificazione del rischio. Per gli attori viene utilizzato il punteggio combinato costituito dal 70% di similarità embedding e dal 30% di Jaccard.

Le dimensioni attività, scenario di rischio e scenario delle opportunità assumono invece valori differenti a seconda della versione del confronto: nel primo approccio derivano dalla similarità embedding, mentre nel secondo sono ottenute dagli score normalizzati prodotti dal LLM-as-a-judge.

Per le cause e per le conseguenze è stato necessario ricondurre le rispettive sottocategorie a due singole dimensioni del radar. La dimensione cause è stata quindi ottenuta aggregando i punteggi delle quattro categorie relative a procedure, persone, fattori esterni e strumenti, mentre la dimensione conseguenze è stata ottenuta aggregando i punteggi delle sei categorie relative a conseguenze economico-finanziarie, danno d'immagine, salute e sicurezza, compliance normativa, gestionale-operativo e ambiente. In entrambi i casi è stata utilizzata la media dei punteggi delle sottocategorie. L'ordine degli otto assi è stato mantenuto costante in tutti i confronti, così da rendere i radar ottenuti direttamente confrontabili.

A partire dagli otto punteggi è stata successivamente calcolata l'area del poligono definito dal radar plot. Poiché tutti gli assi sono normalizzati tra 0 e 1 e disposti a intervalli angolari costanti, l'area consente di riassumere in un unico valore l'estensione complessiva del profilo multidimensionale. L'area ottenuta è stata infine normalizzata rispetto all'area massima teoricamente ottenibile da un radar con tutti gli otto valori pari a 1, ottenendo così un valore compreso tra 0 e 1.

L'area radar normalizzata è stata utilizzata come criterio per individuare il best match. Per ciascuna riga del riferimento, il codice confronta tutte le righe astratte appartenenti alla stessa fase, calcolava il radar plot e la relativa area normalizzata per ciascuna coppia e seleziona infine come miglior abbinamento la riga astratta caratterizzata dall'area più elevata. La logica del confronto rimane

quindi orientata dal riferimento verso la tabella astratta: per ogni rischio presente nel riferimento viene ricercato il macro-rischio più coerente tra quelli disponibili nella fase corrispondente.

Durante la selezione, la riga astratta scelta non viene rimossa dall'insieme dei candidati. Una stessa riga astratta può quindi essere selezionata come best match per più righe del riferimento. Questa scelta risulta coerente con la natura stessa dell'astrazione, poiché un singolo macro-rischio può rappresentare in forma sintetica più rischi specifici presenti nel riferimento. Allo stesso tempo, alcune righe astratte possono non essere mai selezionate come miglior abbinamento.

È importante sottolineare che il best match rappresentava esclusivamente il migliore abbinamento disponibile tra i candidati della stessa fase e non necessariamente un abbinamento corretto in senso assoluto. Analogamente, l'area radar non è stata interpretata come una misura clinica assoluta della correttezza della riga prodotta dalla pipeline, ma come un criterio quantitativo utilizzato per ordinare e confrontare automaticamente i possibili abbinamenti. La valutazione definitiva dei rischi rimane pertanto legata alla revisione da parte di esperti di dominio.

Oltre alla selezione dei best match, il codice conserva anche i punteggi relativi a tutte le coppie confrontabili. Sono stati quindi prodotti sia file contenenti l'insieme completo dei confronti tra riferimento e righe astratte della stessa fase, sia file contenenti esclusivamente il miglior abbinamento individuato per ciascuna riga del riferimento. Sono stati inoltre generati output affiancati per facilitare l'analisi qualitativa delle coppie, metriche aggregate per fase e a livello globale e file dedicati all'identificazione delle righe astratte riutilizzate come best match o mai selezionate. Tali output hanno costituito la base per le analisi quantitative e qualitative presentate nel Capitolo 4.

4. Risultati

In questo capitolo sono presentati i risultati ottenuti nelle diverse fasi della pipeline sviluppata. L'analisi proposta nei paragrafi successivi segue l'ordine metodologico descritto nel capitolo precedente, partendo dalla valutazione preliminare dell'abbinamento tra documenti procedurali e fasi del processo CAR-T. Successivamente vengono riportati i risultati dell'estrazione strutturata delle informazioni di rischio, quelli relativi alla fase di astrazione e unione delle righe simili e, infine, i risultati del confronto automatico con il riferimento standardizzato.

I risultati vengono presentati sia in termini quantitativi sia attraverso esempi qualitativi. Questa scelta è stata adottata perché l'obiettivo della pipeline non è soltanto produrre output formalmente strutturati, ma ottenere informazioni coerenti, sintetiche e confrontabili con un riferimento manuale. Per questo motivo, oltre alle metriche globali e per fase, vengono riportati esempi specifici di estrazione, astrazione, match forti, match deboli e differenze tra l'approccio basato su embedding e quello basato su LLM-as-a-judge.

4.1 Abbinamento documenti–fasi

Nel presente lavoro, l'estrazione finale è stata condotta utilizzando l'abbinamento documenti–fasi definito manualmente. Questa scelta è stata adottata per garantire una maggiore copertura delle fasi analizzate e un controllo più accurato sulla pertinenza dei documenti associati a ciascun passaggio del processo. Tuttavia, è stata valutata anche una procedura automatica di abbinamento tramite modello linguistico, con l'obiettivo di esplorare la possibilità di rendere questa fase maggiormente automatizzata in sviluppi futuri della pipeline.

Il confronto tra abbinamento manuale e abbinamento automatico è riportato in Tabella 4.1.

Strategia di abbinamento	Associazioni documento–fase	Fasi coperte sulle prime 12 fasi del processo
Abbinamento manuale	69	12/12
Abbinamento automatico	47	10/12

Tabella 4.1 — Confronto tra abbinamento manuale e abbinamento automatico dei documenti alle fasi del processo CAR-T.

L'abbinamento manuale ha prodotto 69 associazioni documento–fase e ha garantito la copertura di tutte le dodici fasi considerate. L'abbinamento automatico, invece, ha individuato 47 associazioni e ha coperto 10 fasi su 12.

Questo risultato mostra che il modello linguistico è stato in grado di riconoscere una parte rilevante delle relazioni tra documenti e fasi, ma con un comportamento più conservativo rispetto alla valutazione manuale. In particolare, l'approccio automatico tendeva a individuare soprattutto le associazioni più esplicite, tralasciando alcuni collegamenti indiretti o trasversali che, sulla base della lettura dei documenti, risultavano comunque rilevanti per l'analisi del rischio.

Per questo motivo, l'abbinamento automatico non è stato utilizzato come input definitivo per la fase di estrazione. La pipeline è stata quindi eseguita utilizzando la mappatura manuale, considerata più completa e più adatta a garantire la copertura informativa delle fasi analizzate.

Nel complesso, questa prova ha mostrato che l'abbinamento automatico documenti–fasi può rappresentare una componente promettente per una futura pipeline completamente automatizzata, ma nella configurazione valutata richiede ancora supervisione manuale per evitare la perdita di documenti potenzialmente rilevanti per l'estrazione di informazioni.

4.2 Estrazione strutturata

Dopo aver definito l'abbinamento tra documenti procedurali e fasi del processo, è stata valutata la fase di estrazione strutturata delle informazioni di rischio.

Sono stati confrontati tre modelli linguistici locali (phi4:14b, qwen3.5:9b e gpt-oss:20b) e per ciascuno sono state testate due strategie di estrazione (single shot, divided).

Una prima valutazione ha riguardato la stabilità tecnica degli output prodotti. In particolare, è stato conteggiato il numero di fallimenti osservati durante l'esecuzione della pipeline. Per fallimento tecnico si intende un output non direttamente utilizzabile nelle fasi successive, ad esempio per problemi di formato, struttura JSON non valida, errori di parsing o mancato rispetto dello schema richiesto tramite Pydantic. Il confronto tra le diverse configurazioni è riportato in Tabella 4.2. Per ciascuna configurazione l'estrazione è stata eseguita sulle 69 associazioni documento – fase relative alle prime dodici fasi del processo.

Modello e strategia	Fallimenti tecnici	Percentuale di fallimento
phi4:14b – single shot	2/69	2,90%
phi4:14b – divided	0/69	0,00%
qwen3.5:9b – single shot	8/69	11,59%
qwen3.5:9b – divided	15/69	21,74%
gpt-oss:20b – single shot	18/69	26,09%
gpt-oss:20b – divided	20/69	28,99%

Tabella 4.2 — Fallimenti tecnici osservati nelle diverse configurazioni di estrazione.

I risultati mostrano che phi4:14b è stato il modello più stabile dal punto di vista tecnico. Nella configurazione divided non sono stati osservati fallimenti, mentre nella configurazione single shot sono stati rilevati soltanto 2 fallimenti su 69 associazioni documento - fase, pari al 2,90%. Questo indica una buona capacità del modello di rispettare lo schema richiesto e di produrre output compatibili con la successiva elaborazione automatica.

Il modello qwen3.5:9b ha mostrato una stabilità inferiore. Nella configurazione single shot sono stati osservati 8 fallimenti su 69, pari all'11,59%, mentre nella configurazione divided i fallimenti sono aumentati a 15 su 69, pari al 21,74 %. Questo suggerisce che, per tale modello, la suddivisione del compito in più chiamate non ha prodotto un miglioramento della robustezza tecnica.

Il modello gpt-oss:20b ha presentato le maggiori criticità. La configurazione single shot ha prodotto 18 fallimenti su 69, pari al 26,09 %, mentre la configurazione divided ha prodotto 20 fallimenti su 69, pari al 28,99%. Questi valori indicano una minore regolarità nella generazione di output strutturati e una maggiore difficoltà di integrazione automatica nella pipeline.

Nel complesso, la valutazione tecnica ha evidenziato una maggiore affidabilità di phi4:14b, in particolare nella configurazione divided. Tuttavia, la scelta della configurazione finale non è stata basata esclusivamente sul numero di fallimenti tecnici. È stata infatti considerata anche la qualità degli output prodotti, valutando la coerenza con la fase analizzata, la specificità delle informazioni estratte e la corretta separazione tra attività, scenario di rischio, cause e conseguenze.

Per questo motivo, alla valutazione quantitativa dei fallimenti è stata affiancata un'analisi qualitativa di porzioni di tabella generate a partire dalla stessa combinazione documento–fase. Tale confronto ha permesso di osservare in maniera più concreta le differenze tra i modelli e tra le strategie di estrazione, evidenziando non soltanto la stabilità formale degli output, ma anche la loro effettiva utilizzabilità nelle successive fasi di astrazione e confronto.

A titolo di esempio, la Tabella 4.3 riporta il campo *scenario_di_rischio* estratto dalle diverse configurazioni a partire dalla stessa combinazione documento–fase. L'esempio si riferisce al documento ALL IO 093.1.1.1, associato alla fase 7 del processo, relativa al trasporto del prodotto di raccolta al TE. È stato scelto di riportare soltanto lo scenario di rischio perché questo campo è particolarmente rappresentativo della capacità del modello di comprendere il contenuto operativo del documento e di trasformarlo in una descrizione coerente del rischio.

Modello e configurazione	Scenario di rischio estratto
phi4:14b – single shot	Durante il trasporto del prodotto ottenuto da raccolta leucaferetica verso il TE, utilizzando un contenitore dedicato ai prodotti cellulari freschi, potrebbero verificarsi problemi quali danneggiamento del contenitore, gestione termica impropria o perdita accidentale del campione.
phi4:14b – divided	Il trasporto del prodotto ottenuto da raccolta leucaferetica potrebbe non avvenire in condizioni adeguate di temperatura, compromettendo la qualità e l'efficacia delle cellule CAR-T. Questo potrebbe verificarsi a causa di malfunzionamenti del sistema refrigerato, errori nella gestione della catena del freddo o ritardi nel trasporto.
qwen3.5:9b – single shot	Dati incompleti e parziali sulla gestione del patrimonio e degli investimenti, che possono portare a errori nella pianificazione strategica e a una cattiva gestione delle risorse aziendali.
qwen3.5:9b – divided	Violazione della Legge 190/12.
gpt-oss:20b – single shot	Scelta di un contenitore in grado di mantenere la cassetta alla temperatura richiesta durante il trasporto, evitando temperature superiori a 25°C.
gpt-oss:20b – divided	Trasporto.

Tabella 4.3 — Esempio di scenario di rischio estratto da diverse configurazioni modello–strategia per una stessa combinazione documento–fase.

Le due versioni del modello phi4:14b risultano le più coerenti con il contenuto della fase, poiché descrivono un rischio effettivamente legato al trasporto del prodotto cellulare, alla gestione della temperatura, all'integrità del contenitore e alla qualità del materiale raccolto. In particolare, la configurazione single shot produce una descrizione più completa e prolissa del rischio, mentre la configurazione divided restituisce uno scenario maggiormente focalizzato sulla catena del freddo e sulle condizioni di trasporto.

Al contrario, qwen3.5:9b produce scenari non coerenti con la fase analizzata. Nella configurazione single shot lo scenario risulta orientato verso temi patrimoniali e strategici, mentre nella configurazione divided viene riportato un riferimento alla violazione della Legge 190/12, non pertinente rispetto al trasporto del prodotto cellulare. gpt-oss:20b riconosce parzialmente il tema

generale del trasporto nella configurazione single shot, ma lo formula più come una misura preventiva che come uno scenario di rischio. Nella configurazione divided, invece, restituisce uno scenario eccessivamente sintetico e poco informativo.

Questo confronto qualitativo, insieme ad altri esempi analoghi osservati durante l'analisi, conferma che la scelta di phi4:14b non è stata basata soltanto sulla minore percentuale di fallimenti tecnici, ma anche sulla maggiore aderenza semantica degli output alla fase analizzata. La configurazione divided è stata poi selezionata per la prosecuzione della pipeline perché, sull'intero insieme delle esecuzioni, ha garantito il miglior compromesso tra qualità del contenuto, regolarità della struttura e stabilità tecnica, non producendo fallimenti nei 69 tentativi considerati.

Sulla base della valutazione complessiva, la configurazione phi4:14b in modalità divided è stata selezionata per la prosecuzione della pipeline. Questa configurazione ha infatti mostrato il miglior compromesso tra stabilità tecnica, regolarità della struttura e qualità del contenuto estratto, non producendo fallimenti nei 69 tentativi considerati.

4.3 Astrazione

Dopo la selezione della configurazione di estrazione più adatta, rappresentata da phi4:14b in strategia divided, la tabella estratta è stata utilizzata come input per la fase di astrazione e unione delle righe simili.

La tabella di partenza conteneva 69 righe estratte. Al termine della fase di astrazione, queste righe sono state ricondotte a 28 macro-rischi. La riduzione complessiva è stata quindi pari a 41 righe, corrispondente a una diminuzione di circa il 59,4% rispetto alla tabella estratta iniziale.

La Tabella 4.4 riporta, per ciascuna fase, il numero di righe presenti prima e dopo il processo di astrazione.

Fase	Righe estratte	Macro-rischi generati
1	3	2
2	6	2
3	1	1
4	5	2
5	5	2
6	14	5
7	12	5
8	7	3
9	3	1
10	6	2
11	6	2
12	1	1
Totale	69	28

Tabella 4.4 — Numero di righe prima e dopo il processo di astrazione per ciascuna fase del processo CAR-T.

I risultati mostrano che la riduzione non è stata uniforme tra le diverse fasi. Le fasi con il maggior numero di righe estratte sono state anche quelle in cui il processo di astrazione ha prodotto una maggiore riduzione. In particolare, la fase 6 è passata da 14 righe estratte a 5 macro-rischi, mentre la fase 7 è passata da 12 righe estratte a 5 macro-rischi. Questo risultato era atteso, poiché queste fasi erano caratterizzate da una maggiore quantità di informazioni procedurali e da una maggiore presenza di rischi descritti in modo parzialmente sovrapposto.

Anche la fase 8, relativa alla preparazione del prodotto e alla spedizione alla ditta, ha mostrato una riduzione rilevante, passando da 7 righe estratte a 3 macro-rischi. In questo caso, l'astrazione ha permesso di accorpare righe riferite a rischi simili, soprattutto legati alla tracciabilità, alla documentazione, all'etichettatura, alla conformità alle procedure e alla corretta gestione del prodotto cellulare.

Al contrario, alcune fasi presentavano già un numero molto basso di righe estratte. È il caso della fase 3 e della fase 12, per le quali era presente una sola riga iniziale. In questi casi, il processo di astrazione non ha prodotto una riduzione numerica, ma ha comunque mantenuto la riga all'interno della tabella finale, garantendo la copertura della fase.

Nel complesso, il processo di astrazione ha permesso di trasformare una tabella iniziale più estesa ma ridondante in una tabella più concisa costituita da macro-rischi più generali. Questa riduzione non è stata ottenuta eliminando semplicemente righe considerate duplicate, ma attraverso una sintesi guidata delle informazioni estratte. Le righe finali mantengono infatti lo stesso, ma rappresentano rischi più generali, meno frammentati e privi di ridondanze.

Oltre alla riduzione numerica, il processo di astrazione ha consentito anche una maggiore omogeneità degli output. Le righe astratte risultano infatti più sintetiche, meno dipendenti dalla formulazione specifica dei singoli documenti e più orientate alla descrizione del rischio come elemento del processo. Questo ha permesso di ottenere una tabella finale più vicina alla logica del riferimento, nella quale un singolo rischio può rappresentare più eventi specifici o più formulazioni operative riconducibili allo stesso problema.

Per interpretare meglio il risultato quantitativo dell'astrazione, sono stati analizzati alcuni esempi concreti di unione. L'obiettivo era verificare se il modello fosse in grado non solo di ridurre il numero di righe, ma anche di produrre macro-rischi più coerenti, sintetici e rappresentativi rispetto agli scenari iniziali.

Gli esempi riportati in questo paragrafo mostrano porzioni della tabella estratta prima dell'astrazione e il corrispondente macro-rischio generato. Per rendere il confronto leggibile sono riportati solo i campi più rappresentativi: scenario di rischio e livelli di rischio. Questi campi sono particolarmente utili perché permettono di valutare sia la qualità della sintesi semantica sia la capacità del modello di ricondurre il rischio a una categoria più coerente.

La possibilità di valutare le righe in input che hanno partecipato alla generazione dell'output dell'astrazione è stata garantita dalla mappa di unione, output secondario della fase di astrazione descritto nel paragrafo 3.3.3, che ha permesso di mantenere la tracciabilità tra righe iniziali e macro-rischi prodotti.

Un primo esempio riguarda la fase 7, relativa al trasporto del prodotto al TE. In questa fase, la tabella estratta conteneva più righe riferite a rischi simili, legati al trasporto del prodotto leucaferetico, al mantenimento di condizioni adeguate e alla possibile compromissione della qualità del materiale cellulare. Tuttavia, le righe iniziali presentavano classificazioni tassonomiche diverse e non sempre coerenti con la natura effettiva del rischio, come mostrato nella Tabella 4.5.

Scenario di rischio iniziale	**Rischio**
Trasporto del prodotto leucaferetico in condizioni di temperatura non adeguate, con possibile compromissione della qualità delle cellule.	Strategici / Comunicazione e relazioni istituzionali

Prodotto della raccolta leucaferetica non trasportato correttamente al TE, con possibili danni al prodotto cellulare.	Finanziari / Erariale patrimoniale
Rischio di danneggiamento o contaminazione del prodotto durante il trasferimento.	Esterni / Contesto socio-economico nazionale e regionale

Macro-rischio generato	**Rischio**
Deviazioni di temperatura, ritardi logistici o danni accidentali alle sacche durante il trasporto del prodotto cellulare al TE, con possibile compromissione di qualità, integrità e utilizzabilità del prodotto.	Operativi / Gestione farmaci e dispositivi medici

Tabella 4.5 — Esempio di astrazione nella fase 7: unione di scenari relativi al trasporto del prodotto al TE.

L'esempio mostra che il modello ha riconosciuto il nucleo comune delle righe iniziali, riconducendolo a un unico macro-rischio relativo al trasporto del prodotto cellulare. Le formulazioni iniziali descrivevano aspetti leggermente diversi dello stesso problema: temperatura non adeguata, trasporto non corretto, danneggiamento o contaminazione del prodotto. Il macro-rischio finale integra questi elementi in una descrizione più completa, centrata sulla possibilità che deviazioni di temperatura, ritardi logistici o danni accidentali compromettano qualità, integrità e utilizzabilità del prodotto.

Un aspetto rilevante riguarda anche i livelli di rischio. Le righe iniziali erano state classificate in categorie molto diverse, come strategici, finanziari ed esterni. Queste classificazioni risultano poco aderenti al contenuto operativo dello scenario, perché il rischio non riguarda principalmente la comunicazione istituzionale, il patrimonio o il contesto esterno, ma la gestione corretta del prodotto cellulare durante il trasporto. I livelli di rischio finali, ricondotti alla categoria operativi / gestione farmaci e dispositivi medici, appare quindi più coerente con la natura del rischio.

Un secondo esempio riguarda la fase 8, relativa alla preparazione del prodotto e alla spedizione alla ditta. Anche in questo caso, la tabella estratta conteneva più righe riferite a criticità legate alla gestione del prodotto cellulare, alla conformità alle procedure e all'idoneità dei contenitori utilizzati per il trasporto, come mostrato nella Tabella 4.6.

Scenario di rischio iniziale	**Rischio**
Errore nella preparazione del trasporto dei prodotti cellulari, per non corretta conformità alle SOP o attrezzature non adeguate.	Strategici / Governance
Mancata conformità alle SOP durante la spedizione alla ditta, con possibile danneggiamento o degradazione del prodotto.	Strategici / Sistema di Controllo Interno
Interruzione delle attività e trasferimento dei pazienti in emergenza, con possibili violazioni dei regolamenti interni.	Compliance / Regolamenti interni

Macro-rischio generato	Rischio
Danneggiamento strutturale della sacca criopreservata o degradazione termica del prodotto cellulare durante preparazione e confezionamento nel dry shipper, dovuto a non conformità alle SOP o uso di contenitori non idonei.	Operativi / Attività, processi e procedure

Tabella 4.6 — Esempio di astrazione nella fase 8: unione di scenari relativi alla preparazione del prodotto e alla spedizione alla ditta.

In questo caso, il modello ha sintetizzato gli input in un macro-rischio centrato sul danneggiamento della sacca criopreservata o sulla degradazione termica del prodotto cellulare durante la preparazione e il confezionamento nel *dry shipper*. Anche in questo caso il processo di astrazione non si limita a eliminare duplicati, ma integra elementi complementari presenti nelle righe iniziali: non conformità alle SOP, attrezzature non adeguate, possibile danneggiamento del prodotto e criticità legate al confezionamento.

I livelli di rischio finali risultano più coerenti rispetto alle classificazioni iniziali scelte dalla fase di estrazione. Le righe di partenza erano state classificate come rischi strategici o di compliance, ma il contenuto dello scenario era principalmente legato all'aspetto operativo della procedura, alla conformità tecnica e alla corretta gestione del prodotto cellulare. Per questo motivo, la riclassificazione in operativi / attività, processi e procedure rappresenta una classificazione più coerente con il rischio descritto.

Nel complesso, questi esempi evidenziano che l'astrazione ha svolto un doppio compito. Da un lato, ha ridotto la ridondanza della tabella iniziale, mettendo insieme righe riferite allo stesso nucleo di rischio. Dall'altro, ha reso più coerente la classificazione tassonomica, correggendo casi in cui la fase di estrazione aveva attribuito categorie non pienamente allineate alla natura operativa dello scenario, ma privilegiando aspetti di tipo finanziario, strategico o di compliance. Questi risultati mostrano quindi che l'astrazione non ha rappresentato un semplice passaggio di compressione delle righe estratte, ma anche una fase di riorganizzazione semantica guidata dell'informazione.

Un ulteriore aspetto valutato ha riguardato l'astrazione delle cause del rischio. Questo passaggio è risultato particolarmente rilevante perché, nella fase di estrazione, alcune colonne relative alle cause non contenevano sempre vere cause in senso stretto, ma informazioni eterogenee, talvolta più simili a dettagli procedurali, conseguenze, eventi clinici o descrizioni generiche del processo.

In particolare, alcune righe estratte riportavano nelle cause elementi molto granulari della procedura, come volumi di sangue processato, dispositivi utilizzati o modalità di campionamento. In altri casi venivano riportate informazioni che descrivevano più propriamente conseguenze o criticità cliniche, come mancato attecchimento, infezioni o reazioni avverse, piuttosto che fattori causali del rischio. Questa variabilità ha reso necessaria una fase di astrazione non solo finalizzata all'unione delle righe simili, ma anche alla riorganizzazione semantica delle informazioni contenute nelle colonne delle cause.

La possibilità di analizzare in modo controllato questo passaggio è stata resa disponibile dalla mappa di unione. La mappa ha permesso di ricostruire quali righe iniziali fossero confluite in ciascun macro-rischio e quindi di verificare come il modello avesse sintetizzato le cause riportate nelle righe di partenza. In questo modo, è stato possibile controllare non soltanto il risultato finale, ma anche il collegamento tra input estratti e output astratto.

Un esempio riguarda la fase 6, relativa alla procedura aferetica. In questa fase, le righe iniziali contenevano cause molto eterogenee, alcune effettivamente pertinenti alla procedura, altre invece più difficili da interpretare come vere cause del rischio, come riportato nella Tabella 4.7.

Cause del rischio iniziali
Procedure: uso di anticoagulante a base di citrato; processamento di 12–15 litri di sangue intero; raccolta obiettivo di circa 5–10×10⁹ cellule mononucleate; campionamento post-raccolta solo da dispositivo dedicato. **Persone:** controllo delle operazioni da parte di due operatori; competenza e formazione degli operatori nella raccolta leucoaferetica. **Fattori esterni:** uso di dispositivi specifici per campionamento post-raccolta; utilizzo di contenitori validati e sigillature ermetiche termoadesive. **Strumenti:** dispositivo dedicato al campionamento post-raccolta; sigillatura con due saldature ermetiche termoadesive.
Procedure: il numero richiesto di cellule staminali potrebbe non essere raggiunto durante un ciclo di raccolta, rendendo necessaria una seconda procedura. **Persone:** mancato attecchimento delle cellule staminali trasferite. **Fattori esterni:** rischio di infezioni batteriche, virali e fungine. **Strumenti:** gli strumenti utilizzati durante la raccolta e l'infusione delle cellule staminali potrebbero presentare problemi tecnici che richiedono ripetizioni della procedura.
Procedure: uso improprio del criopreservante DMSO; ritardi nell'aferesi o nella criopreservazione; interruzione della catena del freddo. **Persone:** mancata formazione degli operatori; eccessiva fiducia nella tecnologia. **Fattori esterni:** fluttuazioni di temperatura; interruzioni della catena del freddo. **Strumenti:** malfunzionamento dei contenitori di trasporto; usura o difetti dei dispositivi.
Procedure: lesioni o non conformità della sacca criopreservata; danni accidentali da azoto liquido durante lo scongelamento. **Persone:** esposizione a potenziali contagi o reazioni avverse alle terapie cellulari; eventuali errori di gestione del rischio clinico durante le procedure. **Fattori esterni:** situazioni di emergenza che richiedono il trasferimento dei malati; nuove mutagenesi inserzionali nei tumori secondari. **Strumenti:** problemi di integrità o funzionalità dei sistemi informatici utilizzati per la gestione delle procedure; difetti nei dispositivi medici, come le sacche criopreservate.

Cause del rischio dopo astrazione
Procedure: uso improprio di anticoagulanti o mezzi criopreservativi; errori nel processo di sigillatura termoadesiva delle sacche; mancato rispetto dei tempi critici per la criopreservazione. **Persone:** mancata formazione specifica degli operatori sulle procedure di aferesi; errori umani nella manipolazione e nel campionamento post-raccolta. **Fattori esterni:** malfunzionamento dei sistemi di controllo della temperatura durante lo stoccaggio temporaneo. **Strumenti:** difetti o usura nei dispositivi di raccolta e sigillatura; malfunzionamento dei contenitori di trasporto dedicati.

Tabella 4.7 — Esempio di astrazione delle cause del rischio nella fase 6. Le righe iniziali presentavano cause eterogenee, talvolta costituite da dettagli procedurali o elementi non direttamente causali. L'output astratto riorganizza le informazioni in cause più coerenti e pertinenti rispetto alla procedura aferetica.

L'esempio mostra come la fase di astrazione abbia permesso di riorganizzare e riordinare informazioni inizialmente frammentate e non sempre collocate correttamente. Alcuni elementi riportati nelle righe iniziali, come il mancato attecchimento, il rischio di infezioni o l'esposizione a reazioni avverse, descrivono più propriamente possibili conseguenze cliniche o eventi associati al percorso, piuttosto che cause operative del rischio. Altri elementi, come il volume di sangue processato o il numero obiettivo di cellule raccolte, rappresentano dettagli procedurali importanti del processo, ma non sono necessariamente definibili come cause di rischio.

L'output astratto risulta più coerente perché riformula le informazioni mantenendo gli elementi effettivamente interpretabili come cause del rischio. Le cause procedurali vengono ricondotte all'uso improprio di anticoagulanti o criopreservativi, agli errori di sigillatura e al mancato rispetto dei tempi critici. Le cause legate alle persone vengono sintetizzate nella mancata formazione e negli errori umani nella manipolazione e nel campionamento. Le cause esterne vengono ricondotte a criticità nel controllo della temperatura, mentre le cause strumentali vengono collegate a difetti, usura o malfunzionamenti di dispositivi e contenitori.

Questo esempio evidenzia ancora una volta che l'astrazione non ha avuto soltanto una funzione di compressione della tabella, ma anche di "pulizia" delle informazioni. Il modello ha infatti contribuito a distinguere tra dettagli procedurali, cause effettive, conseguenze e informazioni non pienamente pertinenti, producendo una rappresentazione finale più ordinata e più adatta al confronto con il riferimento.

4.4 Confronto con il riferimento

L'output della fase di astrazione è stato confrontato con il riferimento.

Tra le prime dodici fasi analizzate, non tutte erano effettivamente confrontabili con il riferimento. In particolare, le fasi 4, 9, 10 e 11 riguardavano attività attribuite principalmente alla ditta produttrice del farmaco CAR e, per questo motivo, non presentavano righe corrispondenti nel riferimento. Tali fasi sono state comunque mantenute nell'analisi, in accordo con il giudizio degli esperti, valutando i possibili rischi dal punto di vista dell'organizzazione ospedaliera e non da quello interno della ditta produttrice. Di conseguenza, il confronto quantitativo con il riferimento è stato condotto sulle fasi rimanenti. Nel complesso, nelle fasi confrontabili erano presenti 21 macro-rischi astratti e 30 righe di riferimento. Poiché ogni riga del riferimento è stata confrontata con tutte le righe astratte appartenenti alla stessa fase, sono stati generati complessivamente 86 confronti. Per ciascuna riga del riferimento è stato poi individuato il *best match* considerando l'area radar normalizzata più elevata tra quelle disponibili nella stessa fase.

Il confronto è stato eseguito utilizzando sia metriche tradizionali sia metriche semantiche. Le metriche tradizionali, come l'indice di Jaccard e il punteggio gerarchico utilizzato per i livelli di rischio, sono rimaste invariate nelle due versioni del confronto. Per la valutazione dei campi testuali più complessi sono state invece utilizzate due strategie alternative: una basata su embedding semantici e una basata su LLM-as-a-judge. In entrambi i casi, la logica di matching è rimasta la stessa. I risultati sono stati analizzati sia considerando i migliori abbinamenti individuati per ciascuna riga del riferimento, sia considerando l'insieme completo degli 86 confronti disponibili.

4.4.1 Valutazione globale

Una prima valutazione ha riguardato i risultati globali ottenuti dalle due versioni del confronto. Entrambi gli approcci hanno individuato un *best match* per tutte le righe del riferimento. Questo significa che, per ogni rischio presente nel riferimento, era disponibile almeno una riga astratta appartenente alla stessa fase sulla quale effettuare il confronto.

I principali risultati globali sono riportati in Tabella 4.8.

Metrica	**Embedding**	**LLM-as-a-judge**
Fasi confrontabili	8	8
Righe del riferimento con best match	30/30	30/30
Confronti totali generati	86	86
Macro-rischi astratti nelle fasi confrontabili	21	21
Macro-rischi astratti scelti almeno una volta	13	12
Macro-rischi astratti riutilizzati	9	9
Macro-rischi astratti mai selezionati	8	9
Area radar media dei best match	0,127	0,212
Area radar mediana dei best match	0,132	0,211
Area radar massima dei best match	0,228	0,548

Tabella 4.8 — Risultati globali del confronto automatico tra tabella astratta e riferimento.

Entrambi gli approcci hanno quindi assegnato un *best match* a tutte le 30 righe del riferimento. La copertura delle righe del riferimento è risultata completa in entrambe le versioni del confronto. Una prima differenza riguarda invece il numero di macro-rischi astratti selezionati almeno una volta. L'approccio embedding ha selezionato 13 macro-rischi distinti sui 21 disponibili nelle fasi confrontabili, mentre l'approccio LLM-as-a-judge ne ha selezionati 12. I due insiemi risultano in larga parte sovrapponibili, ma non completamente coincidenti: l'approccio embedding ha selezionato in modo esclusivo due macro-rischi, mentre il judge ne ha selezionato uno non individuato dall'altro approccio.

In entrambi i casi sono inoltre presenti macro-rischi astratti riutilizzati, cioè selezionati come *best match* per più righe del riferimento. Questo comportamento risulta coerente con la logica dell'astrazione, poiché una singola riga astratta può descrivere un macro-rischio più generale e rappresentare quindi più rischi specifici presenti nel riferimento. In entrambe le versioni del confronto sono stati riutilizzati 9 macro-rischi astratti; sette di questi risultano comuni ai due approcci, mentre i restanti differiscono in funzione del metodo di valutazione utilizzato.

Allo stesso tempo, alcune righe astratte non sono mai state selezionate come *best match*. Considerando esclusivamente le fasi confrontabili, l'approccio embedding ha lasciato 8 macro-rischi non selezionati, mentre il judge ne ha lasciati 9. Sette macro-rischi non selezionati risultano comuni ai due approcci, mentre i restanti differiscono in funzione del metodo di valutazione. Questo risultato non indica necessariamente che tali macro-rischi siano errati, ma solamente che non sono risultati

il miglior abbinamento per nessuna delle righe presenti nel riferimento. In alcuni casi possono infatti rappresentare rischi non presenti nel riferimento, mentre in altri possono essere meno coerenti rispetto ad altri macro-rischi disponibili nella stessa fase.

Distribuzione delle aree radar dei best match

Dopo aver selezionato, per ciascuna riga del riferimento, il macro-rischio astratto con area radar normalizzata più elevata, è stata analizzata la distribuzione delle aree ottenute sui 30 best match. La distribuzione dei valori per i due approcci è riportata in Figura 4.1. Questa analisi permette di osservare il comportamento finale dei due approcci considerando esclusivamente gli abbinamenti effettivamente selezionati dalla pipeline.

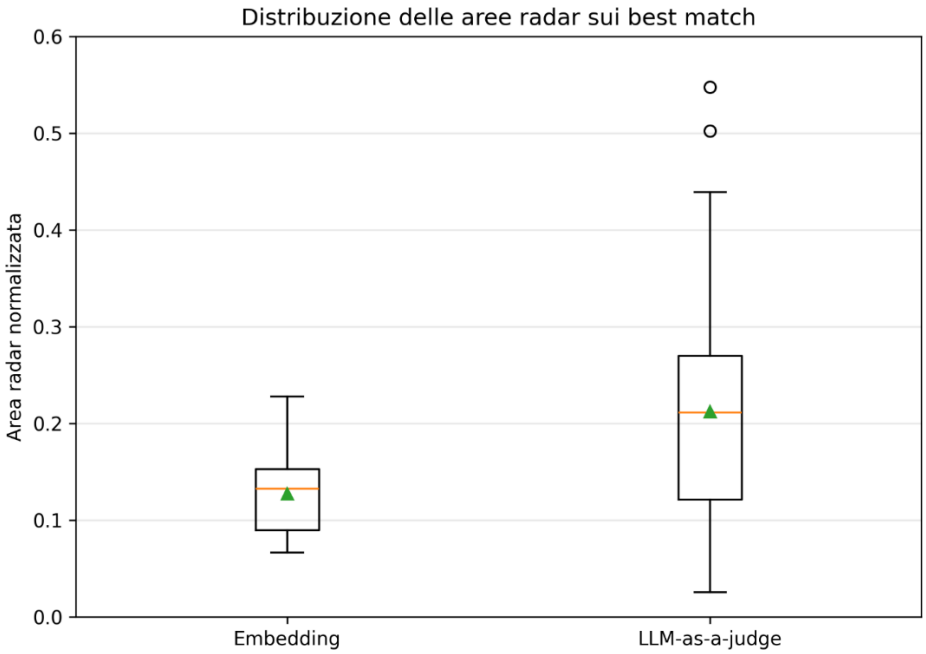


Figura 4.1 — Distribuzione delle aree radar normalizzate sui best match. Il boxplot confronta le aree radar ottenute per i 30 best match selezionati tramite embedding e tramite LLM-as-a-judge.

Statistica	Embedding	LLM-as-a-judge
Numero di best match	30	30
Media area radar	0,127	0,212
Mediana area radar	0,132	0,211
Deviazione standard	0,044	0,135
Minimo	0,066	0,026
Primo quartile	0,089	0,121
Terzo quartile	0,153	0,270
Massimo	0,228	0,548

Tabella 4.9 — Statistiche descrittive delle aree radar normalizzate sui best match.

I risultati mostrati in Tabella 4.9 evidenziano che l'approccio basato su embedding produce valori di area radar più concentrati. La media è pari a 0,127, la mediana a 0,132 e la deviazione standard a 0,044. Anche il valore massimo raggiunto è relativamente contenuto, pari a 0,228. La distribuzione risulta quindi concentrata in un intervallo relativamente ristretto.

L'approccio basato su LLM-as-a-judge presenta invece una distribuzione più ampia. La media dell'area radar è pari a 0,212, la mediana a 0,211 e la deviazione standard a 0,135. Il valore massimo raggiunge 0,548, risultando nettamente superiore rispetto al massimo ottenuto con embedding. Il judge assegna quindi valori maggiormente differenziati tra i diversi best match.

La maggiore dispersione dei valori ottenuti con il judge è visibile anche nel boxplot: mentre l'embedding produce una distribuzione più compatta, l'LLM-as-a-judge presenta un intervallo più ampio tra i valori ottenuti. Questo comportamento indica che i due approcci producono distribuzioni differenti nella valutazione degli abbinamenti selezionati.

Distribuzione delle aree radar su tutti i confronti

Oltre all'analisi dei soli best match, è stata valutata la distribuzione delle aree radar su tutti i confronti generati dalla pipeline. La distribuzione dei valori per i due approcci è riportata in Figura 4.2, Questa analisi permette di osservare il comportamento dei due approcci sull'intero insieme dei possibili abbinamenti tra righe del riferimento e righe astratte appartenenti alla stessa fase, e non soltanto sulle coppie selezionate come migliori.

Nel complesso, per ciascun approccio sono stati generati 86 confronti. Per ogni coppia è stata calcolata l'area radar normalizzata, ottenendo così una distribuzione complessiva dei punteggi assegnati da embedding e LLM-as-a-judge.

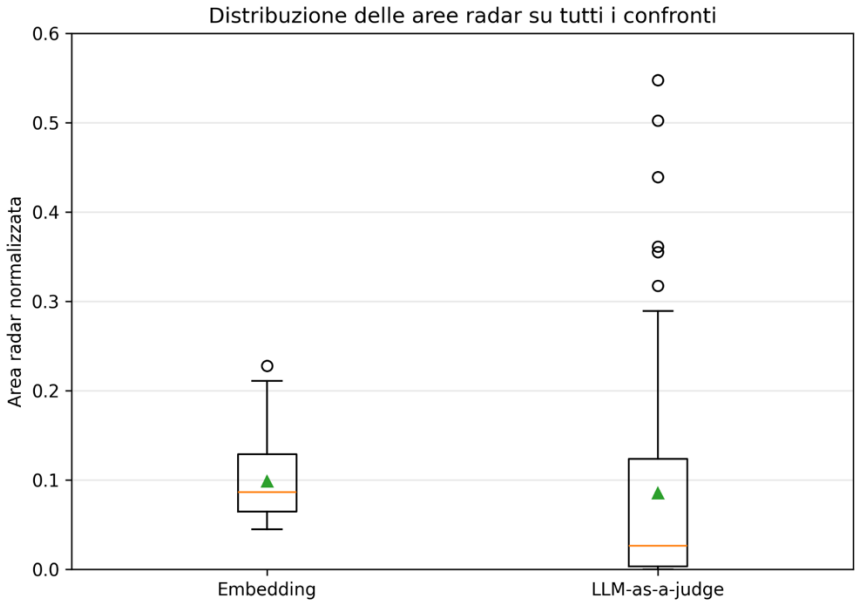


Figura 4.2 — Distribuzione delle aree radar normalizzate su tutti i confronti. Il boxplot confronta le aree radar calcolate sugli 86 confronti disponibili per ciascun approccio.

Statistica	Embedding	LLM-as-a-judge
Numero di confronti	86	86
Media area radar	0,099	0,085
Mediana area radar	0,086	0,026
Deviazione standard	0,042	0,124
Minimo	0,045	0,000
Primo quartile	0,065	0,003
Terzo quartile	0,129	0,124
Massimo	0,228	0,548

Tabella 4.10 — Statistiche descrittive delle aree radar normalizzate su tutti i confronti disponibili.

I risultati mostrati in Tabella 4.10 mostrano che considerando tutti gli 86 confronti, la distribuzione ottenuta con embedding risulta nuovamente più concentrata. La media è pari a 0,099, la mediana a 0,086 e la deviazione standard a 0,042. I valori sono distribuiti in un intervallo relativamente ristretto, compreso tra 0,045 e 0,228.

L'LLM-as-a-judge mostra invece un comportamento differente. Considerando tutti i confronti, l'area radar presenta una media pari a 0,085 e una mediana molto più bassa, pari a 0,026. La deviazione standard è invece più elevata, pari a 0,124, con valori compresi tra 0 e 0,548. La distribuzione ottenuta dal judge risulta quindi molto più ampia e caratterizzata dalla presenza contemporanea di numerosi valori bassi e di alcuni valori sensibilmente più elevati.

Mettendo a confronto le distribuzioni dei best match con quelle ottenute su tutti gli abbinamenti, emerge una differenza rilevante. Sui best match, il judge presenta un'area media superiore rispetto all'embedding (0,212 contro 0,127); considerando invece tutti gli 86 confronti, la sua area media risulta leggermente inferiore (0,085 contro 0,099) e la mediana scende fino a 0,026. L'embedding mantiene invece valori più omogenei nei due insiemi di confronto.

Nel complesso, questi risultati mostrano quindi due comportamenti differenti: l'embedding produce valori più concentrati e relativamente uniformi, mentre l'LLM-as-a-judge genera una distribuzione più dispersa, assegnando valori molto bassi a numerosi confronti e valori più elevati ad alcune coppie selezionate come best match. Le differenze tra i due approcci vengono approfondite nel paragrafo successivo attraverso l'analisi delle singole dimensioni che compongono il profilo radar.

4.4.2 Confronto tra embedding e LLM-as-a-judge

Dopo aver analizzato separatamente le distribuzioni delle aree radar sui best match e su tutti i confronti disponibili, è stato approfondito il confronto tra i due approcci di valutazione semantica: embedding e LLM-as-a-judge.

Per comprendere meglio il comportamento dei due metodi, sono stati confrontati i valori medi e le relative deviazioni standard delle singole dimensioni. I risultati sono riportati in Tabella 4.11.

Dimensione radar	**Embedding, media ± DS**	**LLM-as-a-judge, media ± DS**
Attività	0,564 ± 0,098	0,673 ± 0,322
Attori	0,310 ± 0,040	0,307 ± 0,042
Strutture	0,248 ± 0,115	0,262 ± 0,107
Livelli di rischio	0,388 ± 0,402	0,321 ± 0,406
Scenario di rischio	0,528 ± 0,077	0,593 ± 0,270
Scenario delle opportunità	0,128 ± 0,202	0,353 ± 0,239
Cause	0,372 ± 0,109	0,520 ± 0,237
Conseguenze	0,363 ± 0,128	0,478 ± 0,227

Tabella 4.11 — Valori medi e deviazione standard (DS) delle dimensioni radar sui best match ottenuti con embedding e LLM-as-a-judge.

I risultati mostrano differenze soprattutto nelle dimensioni valutate attraverso i due approcci semantici. L'LLM-as-a-judge presenta valori medi superiori per attività, scenario di rischio, scenario delle opportunità, cause e conseguenze. In particolare, per l'attività il valore medio passa da 0,564 nell'approccio embedding a 0,673 nel judge, mentre per lo scenario di rischio passa da 0,528 a 0,593.

La differenza più marcata è osservabile nello scenario delle opportunità, per il quale l'embedding presenta un valore medio pari a 0,128, mentre l'LLM-as-a-judge raggiunge 0,353. Differenze rilevanti sono presenti anche nelle cause, con valori medi rispettivamente pari a 0,372 e 0,520, e nelle conseguenze, che passano da 0,363 con embedding a 0,478 con il judge.

Oltre ai valori medi, emergono differenze anche nella dispersione dei punteggi. Per le componenti semantiche, l'LLM-as-a-judge presenta generalmente deviazioni standard più elevate rispetto all'embedding. Nell'attività, ad esempio, la deviazione standard passa da 0,098 a 0,322, mentre nello scenario di rischio passa da 0,077 a 0,270. Un comportamento analogo è osservabile per cause e conseguenze. L'embedding restituisce invece valori più concentrati intorno alla media, in accordo con quanto osservato nelle distribuzioni delle aree radar riportate nel paragrafo precedente.

Per attori, strutture e livelli di rischio, invece, le differenze tra i due approcci risultano molto più contenute. Questo comportamento è coerente con la metodologia utilizzata, poiché queste dimensioni non vengono valutate esclusivamente attraverso embedding o judge. Le strutture sono confrontate mediante indice di Jaccard, i livelli di rischio attraverso uno score gerarchico e gli attori

tramite una combinazione composta per il 70% dalla similarità embedding e per il 30% dall'indice di Jaccard.

In particolare, la dimensione relativa agli attori presenta valori praticamente sovrapponibili, pari a 0,310 ± 0,040 per l'embedding e 0,307 ± 0,042 per il judge. Anche per le strutture la differenza è limitata, con una media pari a 0,248 nell'approccio embedding e 0,262 nell'approccio LLM-as-a-judge. Queste dimensioni risultano quindi sostanzialmente stabili tra le due versioni del confronto.

I livelli di rischio presentano invece un valore medio leggermente superiore nell'approccio embedding, pari a 0,388, rispetto a 0,321 ottenuto con LLM-as-a-judge. Tuttavia, questi ultimi vengono calcolati nello stesso modo in entrambi i codici attraverso lo score gerarchico L1–L2–L3. La differenza non deriva quindi da una diversa modalità di valutazione dei livelli di rischio, ma dal fatto che i due approcci possono selezionare come best match righe astratte differenti, che presentano a loro volta classificazioni sui livelli di rischio diverse.

Nel complesso, l'analisi delle singole dimensioni conferma quindi quanto osservato attraverso le distribuzioni delle aree radar. Le componenti condivise dai due approcci rimangono relativamente simili, mentre le principali differenze emergono nei campi testuali valutati semanticamente. L'LLM-as-a-judge produce in queste dimensioni valori medi generalmente più elevati e una maggiore variabilità, mentre l'embedding restituisce punteggi più concentrati e uniformi.

La valutazione delle singole dimensioni non permette tuttavia, da sola, di stabilire se un abbinamento sia effettivamente corretto o meno. Per questo motivo, oltre alle analisi quantitative globali, sono stati considerati alcuni esempi concreti di confronto, utili per osservare come i due approcci si comportino in presenza di match forti, casi ambigui e abbinamenti deboli.

Per interpretare in modo più concreto le differenze tra embedding e LLM-as-a-judge, sono stati analizzati alcuni esempi specifici di confronto tra righe del riferimento e i macro-rischi astratti. L'obiettivo non era soltanto osservare quale riga venisse selezionata come best match, ma anche valutare la forma del radar plot e le singole dimensioni che contribuivano all'area finale.

L'analisi dei radar plot è utile perché permette di capire se un match è bilanciato su più componenti oppure se il valore dell'area dipende solo da alcune dimensioni. Inoltre, consente di evidenziare alcune criticità dei metodi di confronto: l'embedding può assegnare valori relativamente alti a righe lessicalmente o semanticamente vicine ma non perfettamente coerenti dal punto di vista operativo, mentre il judge può riconoscere meglio la corrispondenza concettuale, pur restando vincolato alle dimensioni strutturate del radar, come livelli di rischio, attori e strutture.

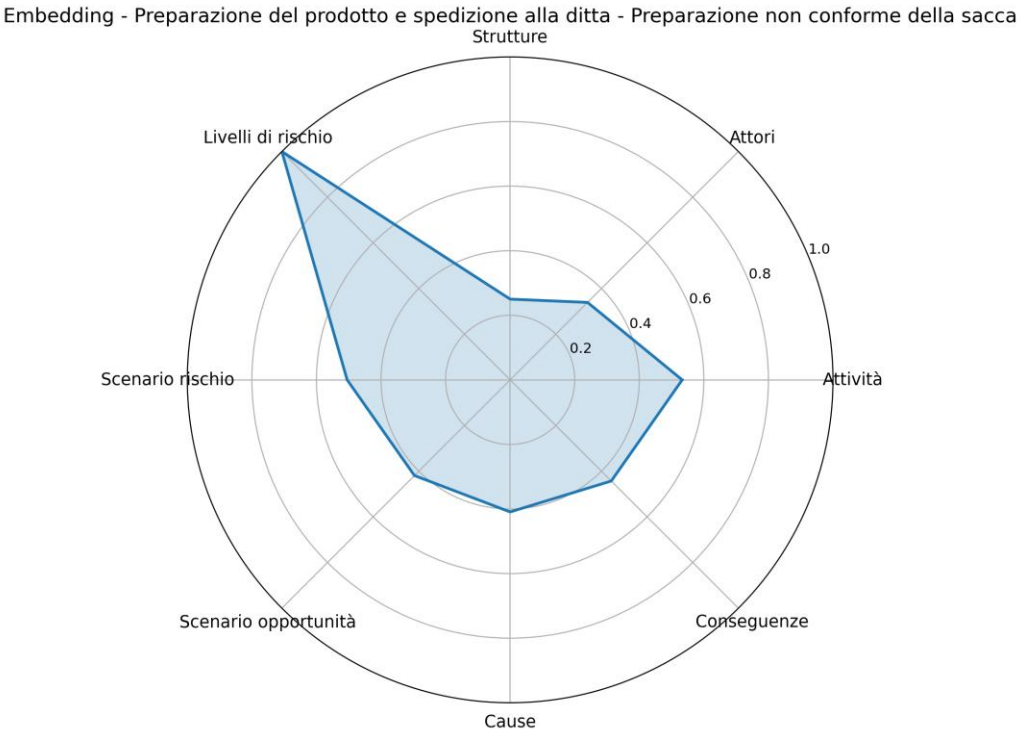
Un primo esempio riguarda la fase 8, relativa alla preparazione del prodotto e alla spedizione alla ditta. La riga di riferimento descrive un rischio legato alla preparazione non conforme della sacca rispetto agli standard richiesti dalla ditta produttrice. In questo caso, embedding e LLM-as-a-judge selezionano lo stesso macro-rischio astratto, ma con aree radar differenti. I risultati sono riportati in Tabella 4.12 e in Figura 4.3.

Elemento	**Embedding**	**LLM-as-a-judge**
Riferimento — scenario di rischio	Sacca preparata in maniera non conforme rispetto agli standard richiesti dalla ditta produttrice.	
Macro-rischio astratto selezionato	Danneggiamento strutturale della sacca criopreservata o degradazione termica del prodotto cellulare durante preparazione e confezionamento nel dry shipper, dovuto a non conformità	Danneggiamento strutturale della sacca criopreservata o degradazione termica del prodotto cellulare durante preparazione e confezionamento nel dry shipper, dovuto a non conformità

	alle SOP o uso di contenitori non idonei.	alle SOP o uso di contenitori non idonei.
Area radar normalizzata	0,228	0,355

Tabella 4.12 - Esempio di confronto nella fase 8, preparazione del prodotto e spedizione al TE: stesso macro-rischio selezionato dai due approcci.

In questo caso, entrambi i metodi individuano lo stesso macro-rischio come best match. Tuttavia, il radar plot del judge presenta un'area maggiore, perché assegna punteggi più elevati alle dimensioni semanticamente più rilevanti, come attività, scenario di rischio, scenario delle opportunità, cause e conseguenze. L'embedding riconosce una somiglianza tra i testi, ma produce un profilo più contenuto. Questo esempio mostra quindi una situazione in cui entrambi gli approcci sono coerenti nella selezione, ma il judge valuta il match come più forte dal punto di vista concettuale.



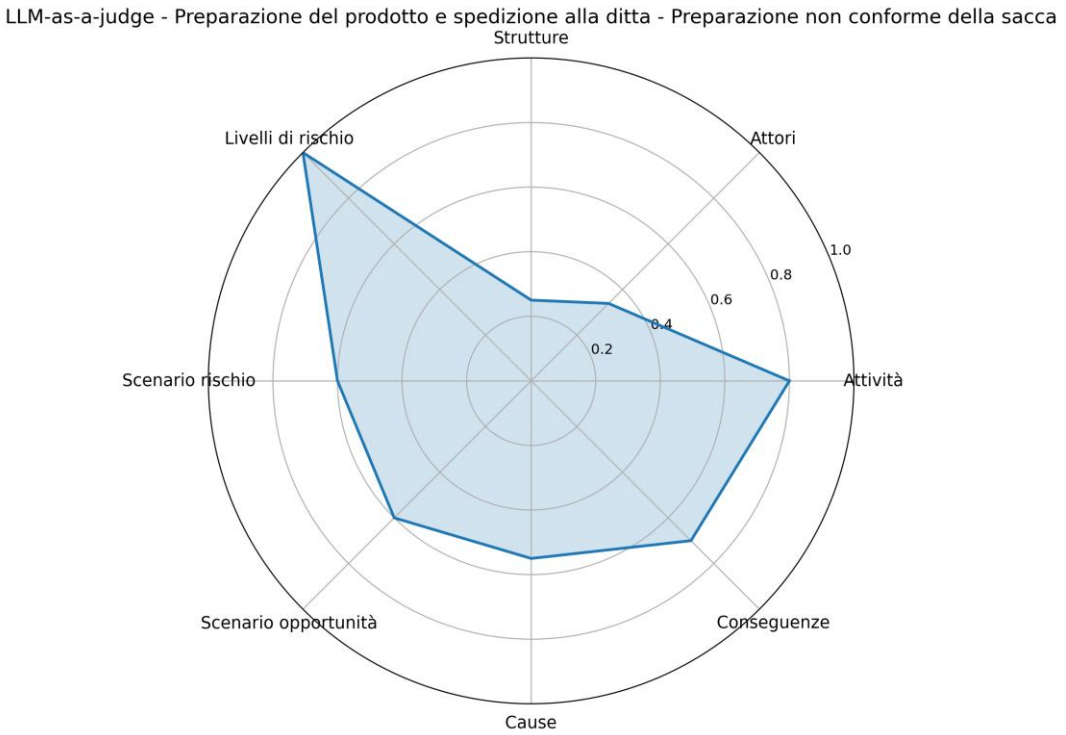


Figura 4.3 - Radar plot del confronto relativo alla fase 8, preparazione del prodotto e spedizione a TE, per lo scenario di rischio relativo alla preparazione non conforme della sacca.

Un secondo esempio riguarda la fase sei, ovvero la Procedura Aferetica e mostra una criticità più evidente dell'approccio embedding. In questo caso, la riga del riferimento riguarda la stampa dell'etichetta contenente i dati anagrafici del paziente e il rischio di un'errata identificazione e tracciabilità del prodotto. Lo scenario delle opportunità riportato nel riferimento prevede inoltre un upgrade del sistema Emonet. I due approcci selezionano come best match due macro-rischi differenti. I risultati sono mostrati in Tabella 4.13 e in Figura 4.4.

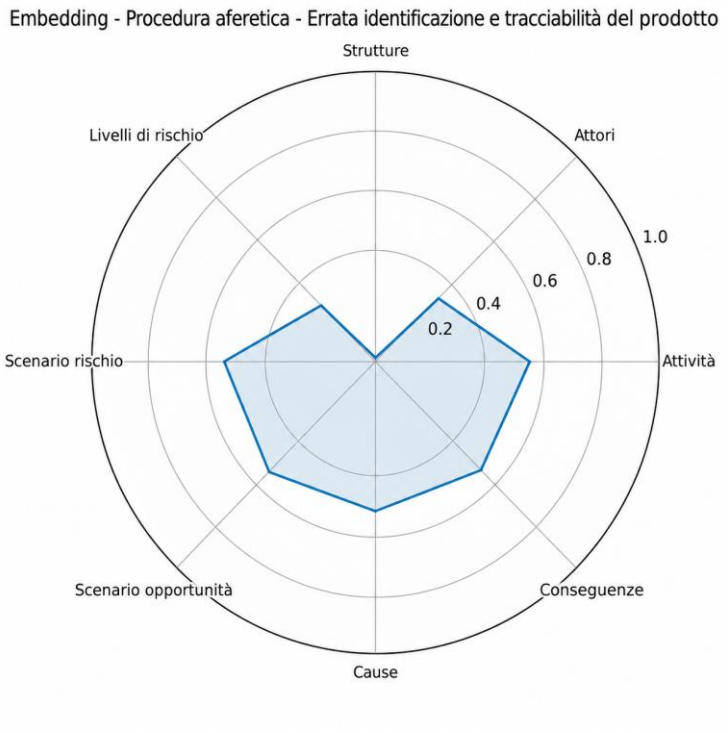
Elemento	**Embedding**	**LLM-as-a-judge**
Riferimento — scenario di rischio	Errata stampa della prima etichetta e produzione dell'etichetta con scorretta identificazione della tracciabilità del prodotto.	
Macro-rischio astratto selezionato	Inadeguata informazione, mancata o incompleta acquisizione del consenso informato consapevole, o carenze nel monitoraggio clinico del paziente sottoposto ad aferesi, con conseguente mancata prevenzione o gestione tempestiva di complicanze immediate.	Errori o discrepanze nell'identificazione del paziente, nell'etichettatura delle sacche o nella registrazione dei dati nei sistemi informatici di tracciabilità durante la raccolta e il trasferimento, determinando il rischio di scambio di prodotto o non conformità regolatoria.
Area radar normalizzata	0,136	0,317

Tabella 4.13 - Esempio di confronto nella fase 6, procedura aferetica: diversa selezione del best match tra embedding e LLM-as-a-judge.

In questo esempio, l'embedding seleziona un macro-rischio relativo alla valutazione clinica, al consenso informato e al monitoraggio del paziente. La riga appartiene alla stessa fase e presenta quindi una vicinanza generale con il contesto aferetico, ma non rappresenta in modo specifico il problema principale del riferimento, che riguarda invece la corretta identificazione del paziente, l'etichettatura e la tracciabilità del prodotto. L'LLM-as-a-judge seleziona invece un macro-rischio maggiormente aderente al contenuto del riferimento, perché riguarda direttamente errori di identificazione, etichettatura e registrazione nei sistemi informatici di tracciabilità. Anche lo scenario delle opportunità risulta maggiormente coerente: il riferimento riporta come opportunità l'upgrade di Emonet, mentre la riga scelta dal judge propone l'implementazione di sistemi digitali di tracciabilità con codice a barre e scansione elettronica.

Questo esempio evidenzia una criticità dell'approccio embedding. La similarità vettoriale può infatti individuare una riga semanticamente vicina al contesto generale della fase, senza necessariamente riconoscere il nucleo specifico del rischio. L'LLM-as-a-judge, invece, in questo caso distingue meglio il problema di identificazione e tracciabilità rispetto ad altri rischi appartenenti alla stessa fase.

Anche il radar del judge mostra tuttavia una discrepanza importante: i livelli di rischio presentano uno score pari a 0, poiché il macro-rischio selezionato è classificato come Compliance / Regolamenti interni, mentre il riferimento appartiene alla categoria Clinico-sanitari / Identificazione del paziente. Questo evidenzia come una buona corrispondenza semantica non implichi necessariamente un allineamento completo su tutte le dimensioni della tabella.



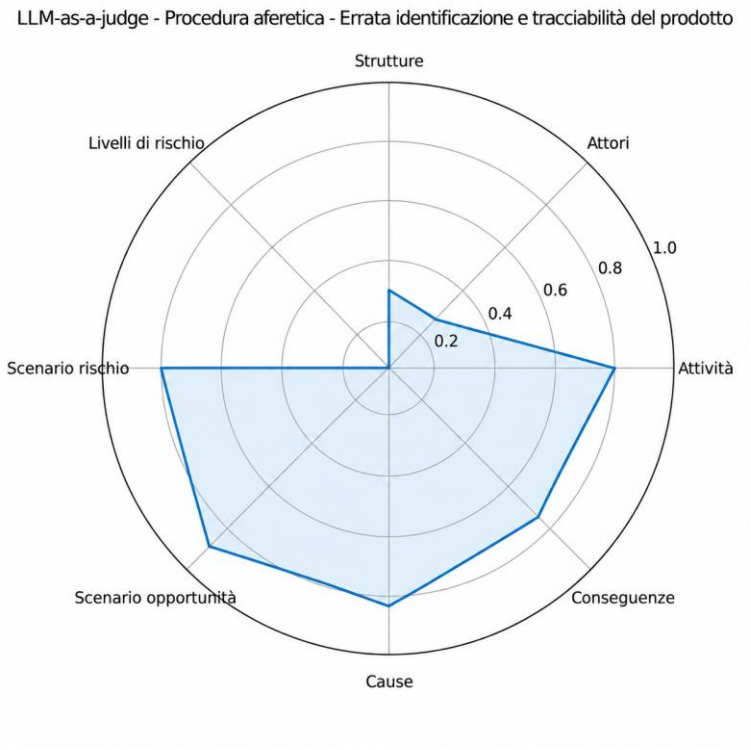


Figura 4.4 — Radar plot del confronto relativo alla procedura aferetica. L'embedding seleziona un macro-rischio relativo al consenso informato e al monitoraggio del paziente, mentre il LLM-as-a-judge individua un rischio maggiormente centrato sull'identificazione, sull'etichettatura e sulla tracciabilità del prodotto.

Un terzo esempio riguarda ancora la fase sei, ovvero la Procedura Aferetica e rappresenta invece un caso in cui nessuno dei due approcci individua un abbinamento pienamente coerente con la riga del riferimento. Il riferimento descrive un rischio di interruzione delle attività di leucoaferesi in presenza di situazioni di emergenza, associato in particolare alla disponibilità e all'affidabilità delle apparecchiature utilizzate. Tra le opportunità viene infatti indicata la sostituzione dei separatori con apparecchiature di nuova generazione. I risultati sono riportati in Tabella 4.14 e in Figura 4.5.

Elemento	**Embedding**	**LLM-as-a-judge**
Riferimento — scenario di rischio	Interruzione delle attività per situazioni di emergenza.	
Macro-rischio astratto selezionato	Compromissione della qualità, quantità o integrità del materiale cellulare raccolto durante la procedura aferetica a causa di errori procedurali, uso improprio di anticoagulanti/criopreservanti o difetti di sigillatura delle sacche, con conseguente necessità di ripetere la raccolta.	Carenze nei controlli interni del Sistema di Gestione Qualità o ritardi nella registrazione dei dati clinici e di follow-up nei database obbligatori, con conseguente perdita di accreditamento o sanzioni da parte degli enti regolatori.

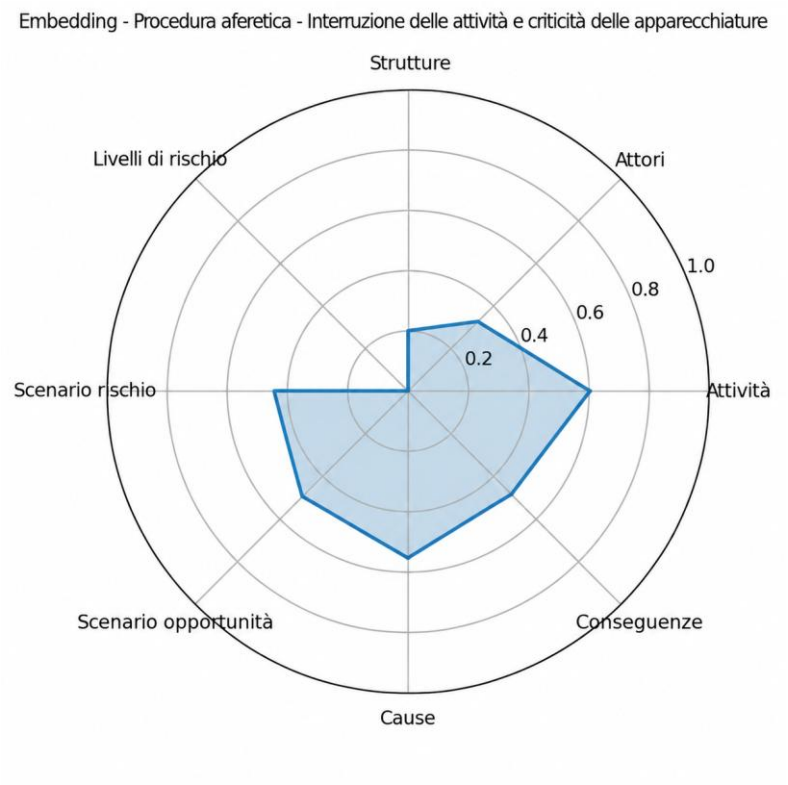
Area radar normalizzata	0,165	0,026

Tabella 4.14 — Esempio di confronto nella fase 6, procedura aferetica: best match complessivamente debole per entrambi gli approcci.

In questo caso, l'embedding seleziona un macro-rischio relativo alla raccolta leucoaferetica e alla criopreservazione iniziale. La scelta presenta una certa coerenza rispetto all'attività generale, poiché entrambe le righe riguardano la procedura aferetica, ma lo scenario specifico è differente. Il riferimento riguarda infatti l'interruzione delle attività in situazioni di emergenza e criticità legate alle apparecchiature, mentre il macro-rischio selezionato dall'embedding è centrato principalmente sulla qualità e sull'integrità del materiale cellulare durante la raccolta. L'LLM-as-a-judge seleziona invece un macro-rischio relativo al controllo qualità e alla conformità documentale del processo aferetico. In questo caso, la corrispondenza con il riferimento è ancora più debole, perché attività, scenario e cause risultano sostanzialmente differenti. La stessa motivazione prodotta dal judge evidenzia infatti che le due righe presentano soltanto una corrispondenza superficiale a livello di struttura e non di contenuto.

Questo esempio evidenzia un limite importante della logica di selezione del best match. Il fatto che una riga venga selezionata come miglior abbinamento non significa necessariamente che essa rappresenti una buona corrispondenza con il riferimento, ma soltanto che ottiene il risultato migliore tra i candidati disponibili nella stessa fase. In questo caso, l'area molto bassa prodotta dal judge, pari a 0,026, rende evidente la debolezza del match.

Anche l'embedding mostra una criticità differente. La condivisione del contesto generale della leucoaferesi permette infatti di ottenere un'area pari a 0,165, nonostante lo scenario specifico non coincida con quello del riferimento. Questo conferma che la similarità vettoriale può assegnare punteggi moderati a righe appartenenti allo stesso contesto operativo anche quando descrivono problemi differenti.



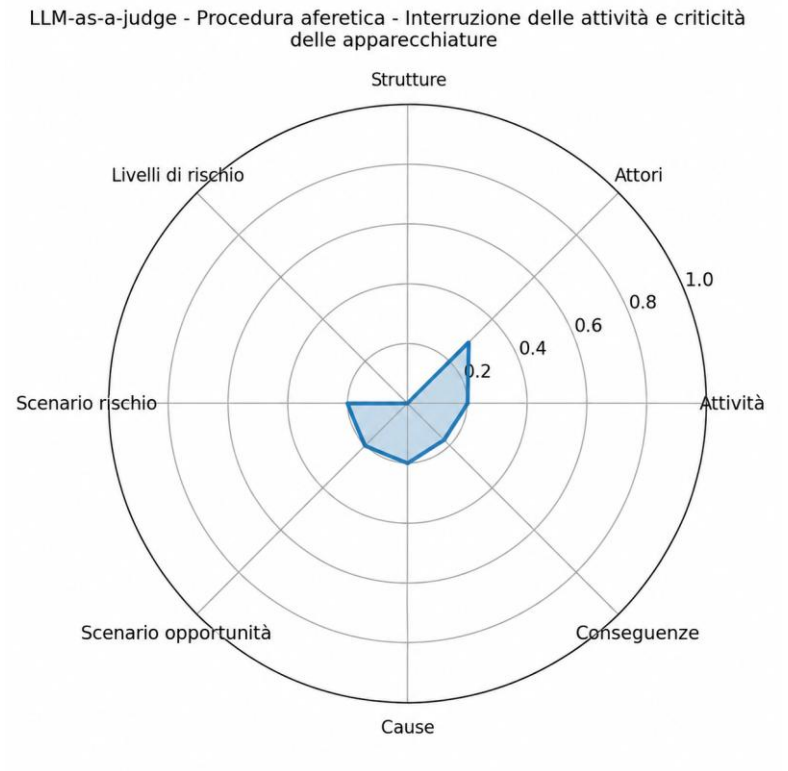


Figura 4.5 - Radar plot del confronto relativo alla procedura aferetica, sull'interruzione delle attività e criticità delle apparecchiature. In questo caso entrambi gli approcci selezionano un best match debole: l'embedding mantiene valori intermedi grazie alla vicinanza con il contesto della leucoaferesi, mentre il judge assegna punteggi molto bassi alla maggior parte delle dimensioni.

Nel complesso, i tre esempi mostrano comportamenti differenti dei due approcci. Nel primo caso, embedding e LLM-as-a-judge individuano lo stesso macro-rischio, pur attribuendo una diversa intensità alla corrispondenza. Nel secondo caso, il judge riesce a distinguere meglio il nucleo specifico del rischio rispetto all'embedding, pur mantenendo alcune discrepanze nelle dimensioni strutturate, come i livelli di rischio. Nel terzo caso, invece, nessuno dei candidati disponibili rappresenta pienamente la riga del riferimento e il best match deve quindi essere interpretato insieme all'area radar e alle singole dimensioni del profilo. Questi esempi confermano quindi l'utilità di affiancare alla selezione automatica del miglior abbinamento una lettura multidimensionale dei risultati.

5. Discussione

Nel capitolo precedente sono stati presentati i risultati ottenuti nelle diverse fasi della pipeline sviluppata per il supporto all'analisi del rischio nel processo CAR-T. In particolare, sono stati analizzati i risultati dell'abbinamento tra documenti e fasi, dell'estrazione strutturata delle informazioni, dell'astrazione delle righe simili e del confronto automatico con il riferimento tramite embedding e LLM-as-a-judge.

La discussione ha l'obiettivo di interpretare questi risultati, evidenziando il contributo della pipeline rispetto al problema iniziale. L'interesse principale non riguarda soltanto la capacità dei modelli di produrre output formalmente corretti, ma la possibilità di confrontarli con un riferimento manuale.

Nel complesso, i risultati mostrano che la pipeline è in grado di supportare diverse fasi del lavoro normalmente svolto in modo manuale: l'identificazione delle informazioni rilevanti nei documenti, la loro organizzazione, la riduzione delle ridondanze e la valutazione della corrispondenza con un riferimento. Tuttavia, i risultati evidenziano anche alcune criticità, legate alla qualità dell'estrazione, alla necessità di una fase di astrazione e alla difficoltà di valutare automaticamente la coerenza semantica tra rischi formulati in modo diverso. I paragrafi successivi discutono le principali componenti della pipeline.

5.1 Interpretazione dei risultati

5.1.1 Estrazione

La fase di estrazione ha l'obiettivo di estrarre informazioni puntuali dai documenti procedurali. I risultati ottenuti mostrano che la scelta del modello e della strategia di estrazione ha avuto un impatto rilevante sulla qualità dell'output. Tra le configurazioni valutate, phi4:14b in modalità divided è risultato il più affidabile, sia dal punto di vista tecnico sia dal punto di vista qualitativo. Dal punto di vista tecnico, questa configurazione non ha prodotto fallimenti, indicando una maggiore capacità di rispettare lo schema richiesto e di generare output compatibili con le fasi successive della pipeline.

Questo risultato è importante perché, in una pipeline automatizzata, la qualità del contenuto non è sufficiente se l'output non è anche strutturalmente utilizzabile. Errori di formato, JSON non validi, campi mancanti o risposte non compatibili con Pydantic impediscono infatti l'integrazione automatica dell'output nelle fasi successive. La stabilità tecnica di phi4:14b in configurazione divided (anche se ha avuto ottime performance anche in configurazione single shot) ha quindi rappresentato un requisito fondamentale per poter proseguire con astrazione, unione e confronto con il riferimento.

Accanto alla stabilità tecnica, è stata osservata anche una maggiore aderenza semantica degli output prodotti da phi4:14b. L'esempio qualitativo riportato nel capitolo dei risultati, relativo al trasporto del prodotto al TE, mostra che phi4:14b è stato in grado di generare scenari di rischio coerenti con il contenuto operativo della fase. Al contrario, altri modelli hanno prodotto output più generici, meno pertinenti o completamente fuori contesto, come riferimenti a rischi patrimoniali, strategici o normativi non coerenti con il trasporto del prodotto cellulare, o peggio, non hanno prodotto alcun risultato.

La strategia divided si è rivelata particolarmente utile perché ha suddiviso il compito di estrazione nelle tre macro-aree della tabella: ownership, anagrafica del rischio e analisi del rischio. Questa impostazione ha probabilmente ridotto il carico computazionale richiesto al modello in una singola chiamata, favorendo una maggiore regolarità dell'output. Anche quando la strategia single shot produceva risposte qualitativamente plausibili e importanti, sempre con l'utilizzo del modello phi4:14b, la configurazione divided ha garantito una maggiore robustezza complessiva sull'intero insieme analizzato.

Tuttavia, i risultati dell'estrazione mostrano anche che questa fase non può essere considerata sufficiente, da sola, per produrre una tabella di rischio finale. Le righe estratte risultavano infatti ricche di informazioni, ma anche ridondanti, eterogenee e talvolta instabili nella classificazione tassonomica, ampiamente pronosticabile essendo informazioni estratte da documenti procedurali più o meno specifici. Questo comportamento era in parte atteso, perché l'obiettivo dell'estrazione era recuperare il maggior numero possibile di informazioni potenzialmente rilevanti dai documenti, non produrre immediatamente una tabella sintetica e definitiva.

Una criticità emersa riguarda la distinzione tra scenario di rischio, cause e conseguenze. In alcuni casi, il modello ha inserito nelle cause elementi che non rappresentavano vere cause operative, ma dettagli procedurali, condizioni cliniche o conseguenze potenziali. Questo aspetto evidenzia la difficoltà dei modelli linguistici nel separare categorie concettualmente e semanticamente vicine quando il documento di partenza non le distingue esplicitamente. Di conseguenza, l'estrazione ha prodotto una base informativa utile, ma ancora bisognosa di riorganizzazione.

Un'altra criticità ha riguardato i livelli di rischio. In diversi casi, la classificazione proposta nella fase di estrazione tendeva a privilegiare categorie finanziarie, strategiche o di compliance anche per rischi che, nel contesto del processo CAR-T, erano più correttamente interpretabili come operativi o clinico-sanitari. Questo suggerisce che il modello può confondere la natura principale del rischio con alcune sue possibili conseguenze secondarie, come impatti economici, reputazionali o organizzativi.

Per questi motivi, la fase di estrazione deve essere interpretata come un passaggio di raccolta e strutturazione preliminare dell'informazione, non come una fase conclusiva. Il suo valore principale consiste nell'aver trasformato documenti procedurali complessi in una tabella iniziale organizzata, tracciabile e compatibile con elaborazioni successive. Allo stesso tempo, i limiti osservati hanno reso necessaria la fase di astrazione, che ha permesso di ridurre ridondanze, correggere alcune incoerenze e sintetizzare le informazioni in macro-rischi più leggibili.

Nel complesso, i risultati dell'estrazione confermano che l'utilizzo di LLM locali può rappresentare una soluzione praticabile per analizzare documentazione procedurale riservata, soprattutto quando l'output viene vincolato tramite schemi strutturati e controlli automatici. Tuttavia, confermano anche che l'estrazione automatica richiede fasi successive di revisione, normalizzazione e confronto. In questo senso, il modello non sostituisce il lavoro esperto, ma può supportarlo producendo una prima base strutturata su cui applicare controlli, sintesi e validazione.

5.1.2 Astrazione

La fase di astrazione ha rappresentato un passaggio centrale della pipeline, perché ha permesso di trasformare l'output dell'estrazione in una tabella più sintetica, leggibile e adatta al confronto con il riferimento. L'astrazione ha il compito di riorganizzare tali informazioni, riducendo ridondanze e formulazioni sovrapposte.

La fase di astrazione è stata eseguita utilizzando il modello gemini-3.5-flash, scelto perché il compito non consisteva più nell'estrazione diretta da documenti procedurali riservati, ma nella rielaborazione di righe tabellari già prodotte e strutturate, ovvero le tabelle di output ottenute dalla fase precedente di estrazione dai documenti procedurali. A differenza della fase di estrazione, infatti, l'input fornito al modello non era costituito dalle SOP originali, ma dalla tabella estratta tramite phi4:14b. Questo ha permesso di utilizzare un modello più avanzato e più complesso per un compito prevalentemente semantico, basato sul riconoscimento di ridondanze, sulla sintesi di righe simili, sul raggruppamento delle informazioni e sulla possibile correzione di elementi non pienamente coerenti prodotti nella fase di estrazione. L'astrazione ha quindi svolto un duplice ruolo: da un lato ha ridotto la frammentazione della tabella iniziale, dall'altro ha contribuito a migliorare la coerenza interna dell'output, soprattutto nei campi relativi alla classificazione tassonomica e alle cause del rischio, che erano criticità note individuate nella fase di estrazione.

Il risultato quantitativo principale è stato il passaggio da 69 righe estratte a 28 macro-rischi, con una riduzione complessiva di circa il 59,4%. Questo dato mostra che la tabella iniziale conteneva un numero elevato di informazioni parzialmente ripetute o riconducibili agli stessi nuclei di rischio. La riduzione non deve però essere interpretata come una semplice eliminazione di righe duplicate, ma come un processo di sintesi semantica. Il modello ha infatti accorpato righe diverse quando queste descrivevano aspetti simili o complementari dello stesso rischio.

L'utilità dell'astrazione è risultata particolarmente evidente nelle fasi più ricche di informazioni, come la procedura aferetica, il trasporto del prodotto al TE e la preparazione del prodotto e spedizione alla ditta. In questi casi, la documentazione procedurale generava più righe relative a rischi simili, spesso formulate con livelli di dettaglio diversi. Il processo di astrazione ha permesso di ricondurre queste righe a macro-rischi più generali, mantenendo però gli elementi essenziali per descrivere il problema. Questo ha reso la tabella finale meno frammentata e più vicina alla logica del riferimento, nella quale un rischio può rappresentare più eventi specifici appartenenti allo stesso ambito operativo.

Un aspetto importante riguarda la capacità dell'astrazione di correggere o normalizzare alcune incoerenze prodotte nella fase di estrazione. Come mostrato negli esempi qualitativi, alcune righe iniziali presentavano tassonomie non coerenti con la natura effettiva dello scenario. Rischi chiaramente legati alla gestione del prodotto cellulare, alla tracciabilità, al trasporto o alla conformità procedurale erano talvolta classificati come strategici, finanziari o esterni. In questi casi, il modello di astrazione non si è limitato ad accorpare righe simili, ma ha riconsiderato il significato complessivo del macro-rischio e lo ha ricondotto a una categoria più coerente, spesso di tipo operativo.

Questa normalizzazione è rilevante perché i livelli di rischio non rappresentano solo un'etichetta descrittiva, ma influenzano anche la confrontabilità della tabella con il riferimento. Una classificazione incoerente può infatti ridurre la qualità del match automatico anche quando lo scenario di rischio è semanticamente corretto. Per questo motivo, la fase di astrazione ha avuto un ruolo non solo di compressione, ma anche di miglioramento della coerenza interna dell'output e di preparazione più efficace al confronto successivo.

L'esempio relativo alle cause del rischio evidenzia un ulteriore aspetto rilevante del ruolo correttivo dell'astrazione. Nella fase di estrazione, alcune informazioni inserite nelle colonne delle cause non rappresentavano vere cause in senso stretto, ma dettagli procedurali, possibili conseguenze o eventi clinici associati al processo. Ad esempio, elementi come il mancato attecchimento, il rischio di

infezioni o l'esposizione a reazioni avverse descrivono più propriamente possibili effetti o criticità cliniche, mentre altri elementi, come il volume di sangue processato o il numero obiettivo di cellule raccolte, rappresentano dettagli procedurali. L'astrazione ha permesso di riorganizzare queste informazioni, mantenendo gli elementi effettivamente interpretabili come fattori causali e ricollocando il contenuto in una forma più coerente con l'analisi del rischio.

Questo passaggio è particolarmente importante perché le cause costituiscono una componente centrale dell'analisi del rischio. Una causa mal definita rende infatti più difficile interpretare il rischio e individuare eventuali azioni correttive. La fase di astrazione ha quindi contribuito a distinguere meglio tra cause operative, informazioni descrittive del processo e conseguenze potenziali, migliorando la qualità interpretativa della tabella finale.

La mappa di unione ha avuto un ruolo di fondamentale importanza nell'interpretazione di questa fase. Poiché ogni macro-rischio derivava dall'accorpamento di una o più righe iniziali, era necessario mantenere traccia del collegamento tra input e output. La mappa di unione, richiesta esplicitamente nel prompt di astrazione, ha permesso di verificare quali righe estratte fossero confluite in ciascun macro-rischio e di controllare di pari passo la coerenza del raggruppamento. In questo modo, il processo di sintesi e correzione non è rimasto una trasformazione opaca, ma è diventato ricostruibile e verificabile.

Questo aspetto è particolarmente importante in un contesto sanitario, in cui la tracciabilità delle informazioni è essenziale. Senza la mappa di unione, sarebbe stato possibile osservare solo la tabella astratta finale, ma non il percorso attraverso cui ciascun macro-rischio era stato generato. La presenza della mappa ha invece consentito di valutare se la riduzione delle righe avvenisse mantenendo le informazioni rilevanti, se eventuali accorpamenti fossero giustificati dal contenuto delle righe iniziali e se le correzioni introdotte dal modello fossero coerenti con il significato complessivo del rischio.

Tuttavia, anche la fase di astrazione presenta alcuni limiti. Il primo riguarda la possibile perdita di specificità. La costruzione di macro-rischi più generali rende la tabella più sintetica e leggibile, ma può ridurre il dettaglio di alcune informazioni presenti nelle righe di partenza. Questo compromesso tra sintesi e granularità è inevitabile: un'astrazione troppo debole lascerebbe una tabella ridondante, mentre un'astrazione troppo forte rischierebbe di produrre macro-rischi eccessivamente generici.

Un secondo limite riguarda la dipendenza dalla qualità dell'estrazione iniziale. Se le righe estratte contengono informazioni rumorose, non pertinenti o classificate in modo scorretto, il modello di astrazione può correggerle solo in parte. L'astrazione migliora e riorganizza l'output, ma non può garantire una ricostruzione perfetta del rischio se l'informazione di partenza è incompleta o poco coerente. Per questo motivo, la qualità della fase di estrazione rimane determinante per l'intera pipeline.

Nel complesso, i risultati mostrano che l'astrazione ha svolto una funzione essenziale di sintesi, normalizzazione e preparazione al confronto automatico. La tabella finale è risultata più compatta, meno ridondante e più coerente rispetto alla tabella estratta iniziale e soprattutto molto più simile e confrontabile al riferimento fornitomi dal Policlinico San Matteo. Allo stesso tempo, la mappa di unione ha permesso di mantenere la tracciabilità del processo, rendendo verificabile il passaggio dalle righe estratte ai macro-rischi finali.

In questo senso, la fase di astrazione può essere interpretata come il punto di collegamento tra l'estrazione automatica dai documenti e la valutazione rispetto al riferimento. Essa non sostituisce la validazione esperta, ma rende l'output più adatto a essere revisionato, confrontato e interpretato in modo sistematico e automatizzato.

5.1.3 Confronto automatico

Il confronto automatico con il riferimento ha rappresentato la fase conclusiva della pipeline, perché ha permesso di valutare quanto i macro-rischi ottenuti tramite astrazione fossero coerenti con i rischi presenti nella tabella di riferimento. Dopo aver prodotto una tabella più sintetica e normalizzata, era infatti necessario verificare se tale tabella conservasse una corrispondenza significativa con il riferimento fornito dall'ospedale.

Per ogni riga del riferimento, il codice ha cercato il macro-rischio astratto più coerente tra quelli appartenenti alla stessa fase del processo CAR-T. Questa impostazione è rilevante perché pone al centro il rischio atteso nella tabella manuale e valuta se la pipeline sia riuscita a produrre almeno un macro-rischio in grado di rappresentarlo. In questo modo, il confronto non misura semplicemente quanto la tabella astratta sia simile nel suo complesso al riferimento, ma verifica, riga per riga, la copertura dei rischi di riferimento.

Il vincolo del confronto per fase ha avuto un ruolo importante. Ogni riga del riferimento è stata confrontata solo con i macro-rischi astratti appartenenti alla stessa fase del processo. Questo ha ridotto il rischio di abbinamenti impropri tra attività operative diverse e ha mantenuto il confronto coerente con la struttura sequenziale del percorso CAR-T. Infatti, uno scenario di rischio può apparire simile a livello lessicale in più fasi, ma assumere significati diversi a seconda del momento del processo in cui si colloca.

Il fatto che tutte le righe del riferimento abbiano ricevuto un best match indica che, nelle fasi confrontabili, era sempre presente almeno un macro-rischio astratto candidato. Tuttavia, questo risultato non deve essere interpretato automaticamente come una corrispondenza corretta. Il best match rappresenta infatti la migliore opzione disponibile all'interno della stessa fase, ma può essere comunque un match debole se nessuna riga astratta descrive in modo adeguato il rischio specifico del riferimento.

Questo aspetto è emerso chiaramente negli esempi concreti, in particolare nel caso dell'esempio riguardo la procedura aferetica (fase 6) che analizzava l'interruzione delle attività e criticità delle apparecchiature. In questo caso, nessuno dei macro-rischi disponibili rappresentava pienamente il rischio di interruzione delle attività descritto nel riferimento. Di conseguenza, entrambi gli approcci hanno comunque individuato un best match, ma con una corrispondenza soltanto parziale. In particolare, il valore molto basso dell'area radar ottenuta tramite LLM-as-a-judge ha permesso di evidenziare chiaramente la debolezza dell'abbinamento.

Un altro risultato significativo riguarda il riutilizzo delle righe astratte. Poiché una riga astratta selezionata non veniva rimossa dal pool dei candidati, lo stesso macro-rischio poteva essere scelto come best match per più righe del riferimento. Questa scelta è coerente con la logica dell'astrazione: un macro-rischio rappresenta spesso una formulazione più generale, capace di includere più rischi specifici presenti nel riferimento. Il riutilizzo di una riga astratta, quindi, non rappresenta necessariamente un errore, ma può indicare che quella riga sintetizza un nucleo di rischio ampio e ricorrente.

Allo stesso tempo, alcune righe astratte non sono mai state selezionate come best match. Anche questo risultato deve essere interpretato con cautela. Una riga mai selezionata può indicare un rischio effettivamente poco coerente con il riferimento, ma può anche rappresentare un rischio presente nei documenti procedurali e non riportato nel riferimento standardizzato. In questo senso, il confronto con il riferimento permette di valutare la corrispondenza con il riferimento disponibile, ma non stabilisce in modo assoluto che una riga astratta non selezionata sia necessariamente errata o di scarsa qualità.

L'utilizzo dell'area radar normalizzata ha permesso di superare una valutazione basata su una singola metrica complessiva, mantenendo visibili le diverse componenti del confronto. Ogni match

è stato infatti valutato rispetto ad attività, attori, strutture, livelli di rischio, scenario di rischio, scenario delle opportunità, cause e conseguenze. Questo approccio ha reso possibile distinguere casi in cui una riga era coerente sullo scenario ma debole nei livelli di rischio, oppure casi in cui condivideva strutture e attori ma non rappresentava correttamente il rischio descritto nel riferimento.

Nel confronto tra i due approcci, l'LLM-as-a-judge ha mostrato una maggiore capacità di distinguere match forti, parziali e deboli. Questo è emerso sia dalla distribuzione delle aree radar sia dagli esempi qualitativi. L'embedding tendeva a produrre punteggi più compressi e omogenei, mentre il judge assegnava valori più differenziati, premiando maggiormente i match semanticamente coerenti e penalizzando quelli poco pertinenti. Questo comportamento è particolarmente utile in un contesto in cui le righe astratte possono essere formulate in modo più generale rispetto al riferimento.

Tuttavia, il confronto automatico non può essere considerato una validazione definitiva della tabella prodotta. Il suo scopo è piuttosto quello di fornire uno strumento di supporto alla revisione, capace di ordinare i possibili abbinamenti, evidenziare i match più plausibili e segnalare i casi critici. La valutazione finale della correttezza clinica, organizzativa e procedurale dei macro-rischi richiede comunque il giudizio di esperti di dominio.

Nel complesso, i risultati del confronto automatico mostrano che la pipeline è stata in grado di produrre una tabella astratta confrontabile con il riferimento e di associare ciascuna riga del riferimento a un macro-rischio candidato. L'analisi ha però evidenziato che il valore del best match deve essere sempre interpretato insieme all'area radar e alle singole dimensioni del profilo. In questo senso, il confronto automatico non rappresenta un punto di arrivo definitivo, ma uno strumento intermedio per rendere più sistematica, tracciabile e interpretabile la revisione della tabella di rischio.

5.2 Embedding e LLM-as-a-judge: vantaggi e limiti

Il confronto tra embedding e LLM-as-a-judge ha rappresentato il principale aspetto della valutazione finale della pipeline. Entrambi gli approcci sono stati utilizzati per confrontare le componenti testuali più complesse della tabella di rischio, come attività, scenario di rischio, cause e conseguenze. Tuttavia, i risultati ottenuti mostrano che i due metodi hanno comportamenti differenti e rispondono a esigenze diverse.

L'approccio basato su embedding consente di confrontare i contenuti testuali attraverso una rappresentazione vettoriale del loro significato, permettendo di valutare la somiglianza anche quando due testi esprimono concetti simili utilizzando formulazioni differenti. Rispetto a metriche tradizionali di NLP basate prevalentemente sulla coincidenza o sovrapposizione lessicale, questo approccio risulta quindi più adatto a cogliere relazioni semantiche tra frasi non perfettamente sovrapponibili dal punto di vista testuale. Nel contesto della pipeline, tale caratteristica è particolarmente utile perché le righe astratte possono sintetizzare o riformulare il contenuto delle righe presenti nel riferimento, rendendo una semplice corrispondenza lessicale poco rappresentativa della reale somiglianza tra i contenuti.

Nel contesto della pipeline, l'embedding si è dimostrato utile come metodo di confronto preliminare. La sua applicazione permette di ottenere rapidamente punteggi numerici per tutte le coppie riferimento–riga astratta appartenenti alla stessa fase. Per questo motivo, può essere considerato uno strumento efficace per una prima selezione automatica o per filtrare un numero elevato di possibili abbinamenti.

Tuttavia, i risultati hanno mostrato anche alcuni limiti di questo approccio. In particolare, i punteggi tendevano a essere più compressi e meno differenziati rispetto a quelli ottenuti con LLM-as-a-judge. Questo comportamento rende più difficile distinguere in modo netto tra accoppiamenti forti, parziali e deboli. In altre parole, coppie con contenuto differente possono ricevere valori relativamente simili, soprattutto quando condividono termini o concetti generali legati alla stessa fase del processo.

Un altro limite riguarda la capacità di interpretare il nucleo operativo del rischio. L'embedding valuta la vicinanza semantica tra testi, ma non sempre riesce a distinguere rischi che appartengono allo stesso contesto ma descrivono problemi diversi. Ad esempio, due righe relative alla preparazione del prodotto cellulare possono risultare vicine nello spazio semantico anche se una riguarda il danneggiamento fisico della sacca e l'altra la tracciabilità documentale. In questi casi, la similarità vettoriale può riconoscere il contesto generale, ma non cogliere pienamente la differenza sostanziale tra i due scenari.

L'approccio LLM-as-a-judge ha mostrato invece una maggiore capacità interpretativa. Il modello non si limita a misurare una distanza tra testi, ma valuta se la riga astratta copre effettivamente il contenuto del riferimento secondo criteri definiti nel prompt. Questo aspetto è particolarmente rilevante nel confronto tra macro-rischi e riferimento, perché una riga astratta può essere più generale rispetto al riferimento, ma comunque corretta se include il rischio specifico descritto.

I risultati ottenuti indicano che il judge è stato più efficace soprattutto nelle dimensioni semanticamente complesse, come scenario di rischio, scenario delle opportunità, cause e conseguenze. In queste componenti, il modello è riuscito a riconoscere meglio la corrispondenza concettuale tra righe formulate in modo diverso. Inoltre, la distribuzione più ampia dei punteggi suggerisce una maggiore capacità discriminativa: il judge tende ad assegnare valori bassi ai match poco coerenti e valori più elevati ai match semanticamente più forti, attribuendo anche valori intermedi nei casi di corrispondenza parziale.

Un vantaggio ulteriore dell'approccio LLM-as-a-judge è la possibilità di ottenere, oltre allo score numerico, anche una breve motivazione del giudizio. Questo elemento è importante perché rende il

confronto più interpretabile. La motivazione permette infatti di capire perché una coppia sia stata valutata come coerente o non coerente, facilitando la revisione manuale dei risultati e l'individuazione di eventuali errori nel matching.

Nonostante questi vantaggi, anche l'approccio LLM-as-a-judge presenta limiti importanti. Il primo riguarda la dipendenza dal prompt. Il giudizio prodotto dal modello può variare fortemente in base al modo in cui viene formulata la richiesta, alla scala di valutazione utilizzata e ai criteri forniti per distinguere match forti, parziali e deboli. Per questo motivo, il prompt deve essere progettato con attenzione e, idealmente, validato attraverso un confronto con esperti di dominio.

Un limite dell'approccio LLM-as-a-judge riguarda la non riproducibilità dei risultati. I modelli linguistici hanno infatti un funzionamento probabilistico e, anche a parità di input, possono produrre valutazioni leggermente differenti tra esecuzioni successive. Per questo motivo, il punteggio assegnato dal judge non può essere considerato perfettamente deterministico.

Un ulteriore limite riguarda il fatto che il judge, pur fornendo una valutazione più interpretabile attraverso la motivazione associata ai punteggi, non garantisce automaticamente la correttezza clinica o organizzativa del risultato. Il modello può riconoscere relazioni concettuali tra testi, ma non sostituisce la competenza degli esperti che conoscono il processo, le procedure e le implicazioni reali dei rischi. Per questo motivo, il suo ruolo deve essere inteso come supporto alla valutazione e non come validazione definitiva.

Nel complesso, embedding e LLM-as-a-judge non devono essere considerati necessariamente approcci alternativi, ma strumenti caratterizzati da modalità differenti di valutazione della similarità semantica. Entrambi consentono di automatizzare il confronto tra le coppie considerate. L'embedding fornisce una misura continua della similarità tra rappresentazioni vettoriali dei testi, mentre il judge effettua una valutazione contestuale esplicita delle diverse componenti e fornisce anche una breve motivazione del punteggio assegnato.

Alla luce di queste caratteristiche, una possibile evoluzione della pipeline potrebbe prevedere un utilizzo combinato dei due approcci. L'embedding potrebbe essere utilizzato per individuare preliminarmente le coppie semanticamente più vicine, mentre l'LLM-as-a-judge potrebbe essere applicato successivamente per una valutazione più approfondita dei candidati selezionati. In questo modo sarebbe possibile sfruttare le differenti caratteristiche dei due metodi all'interno della stessa procedura di confronto.

Entrambi gli approcci, tuttavia, richiedono una successiva revisione esperta, soprattutto perché il confronto automatico riguarda un ambito sanitario molto complesso e articolato, in cui la correttezza del rischio non può essere valutata esclusivamente sulla base della similarità testuale.

L'utilizzo del radar plot ha rappresentato una scelta metodologica importante nella fase di confronto automatico, perché ha permesso di mantenere una valutazione multidimensionale del match tra riga riferimento e riga astratta. Invece di ridurre immediatamente il confronto a un unico punteggio complessivo, il radar plot ha reso visibili le diverse componenti che concorrono alla qualità dell'abbinamento.

Ogni confronto è stato infatti descritto attraverso otto dimensioni: attività, attori, strutture, livelli di rischio, scenario di rischio, scenario delle opportunità, cause e conseguenze. Questa struttura ha permesso di osservare non solo se una riga astratta fosse complessivamente vicina al riferimento, ma anche in quali aspetti risultasse coerente e in quali invece presentasse discrepanze. Questo aspetto è particolarmente rilevante perché, nel contesto dell'analisi del rischio, la similarità tra due righe non può essere interpretata come un concetto unico. Due righe possono descrivere uno scenario di rischio molto simile, ma differire nei livelli di rischio; oppure possono condividere strutture e attori, ma riportare cause o conseguenze diverse. Il radar plot ha quindi permesso di distinguere

queste situazioni, rendendo il confronto più informativo rispetto a una valutazione sintetica non scomposta.

L'area radar normalizzata è stata utilizzata come criterio quantitativo per selezionare il best match. Tuttavia, il valore dell'area non è stato interpretato come un giudizio assoluto di correttezza, ma come un indicatore sintetico della coerenza complessiva tra due righe.

Un'area più elevata indica la presenza di punteggi complessivamente più alti su più dimensioni del confronto, mentre un'area bassa riflette generalmente valori più contenuti o concentrati solo su alcune componenti.

Uno dei vantaggi di questa impostazione è che l'area radar tiene conto congiuntamente di più dimensioni del confronto. Un valore elevato in una singola componente, accompagnato da valori bassi nelle altre, contribuisce solo in modo limitato all'area complessiva.

Questo comportamento è utile perché evita che una somiglianza isolata, ad esempio nello scenario di rischio o nei livelli di rischio, determini da sola la selezione di un match apparentemente buono. Al contrario, vengono favoriti i confronti in cui la coerenza è distribuita su più componenti della tabella.

Gli esempi concreti riportati nel capitolo dei risultati mostrano bene questa utilità. In alcuni casi, il radar plot ha evidenziato match complessivamente coerenti, caratterizzati da valori buoni su diverse dimensioni. In altri casi, invece, ha permesso di riconoscere match deboli quando nessuno dei macro-rischi disponibili rappresentava pienamente il contenuto del riferimento.

Il radar plot ha inoltre permesso di confrontare embedding e LLM-as-a-judge sulla stessa base. Poiché le otto dimensioni del radar sono rimaste identiche nei due approcci, è stato possibile osservare come cambiassero i profili di confronto a seconda del metodo utilizzato per valutare le componenti testuali.

Poiché le otto dimensioni del radar sono riportate sulla stessa scala normalizzata da 0 a 1, il confronto permette di osservare direttamente le differenze tra i due metodi. Nella maggior parte dei casi i profili presentano una struttura complessivamente simile, soprattutto quando embedding e LLM-as-a-judge selezionano lo stesso macro-rischio, ma differiscono per l'ampiezza dei punteggi assegnati alle singole dimensioni. Il judge tende a produrre una maggiore variabilità nelle componenti semantiche, attribuendo valori più elevati quando riconosce una forte corrispondenza concettuale e valori più bassi nei confronti ritenuti deboli. Quando i due metodi selezionano macro-rischi differenti, possono invece emergere differenze più marcate anche nella forma complessiva del radar.

Un altro vantaggio riguarda l'interpretabilità. Il radar plot consente di visualizzare rapidamente le dimensioni responsabili di un'area elevata o bassa. Ad esempio, se un match presenta una buona corrispondenza nello scenario di rischio ma un punteggio sui livelli di rischio pari a zero, il radar permette di individuare immediatamente questa criticità. Allo stesso modo, se cause e conseguenze risultano poco coerenti, il profilo grafico evidenzia che il match è solo parziale, anche se alcune altre dimensioni risultano positive.

Questa interpretabilità è particolarmente utile in una pipeline pensata come supporto alla revisione esperta. Il radar plot non sostituisce la valutazione manuale, ma rende più semplice orientarla. Un esperto può infatti concentrarsi sui match con area bassa, sui profili molto sbilanciati o sui casi in cui alcune dimensioni risultano sistematicamente deboli. In questo senso, il radar plot può funzionare anche come strumento di prioritizzazione della revisione.

Nel complesso, la valutazione tramite radar plot ha permesso di rendere il confronto automatico più trasparente e controllabile. L'area radar ha fornito un criterio sintetico per ordinare i possibili match, mentre la forma del radar ha mantenuto visibili le singole componenti della valutazione. Questa

combinazione tra sintesi quantitativa e interpretabilità grafica è risultata particolarmente utile per analizzare un output complesso come una tabella di rischio, in cui la qualità del match dipende da più dimensioni e non da una sola misura di similarità.

Sebbene la valutazione tramite radar plot abbia reso il confronto più interpretabile e trasparente, anche questo approccio presenta alcuni limiti. L'area radar normalizzata è stata utile per sintetizzare il profilo multidimensionale di ogni coppia riferimento–riga astratta, ma non deve essere interpretata come una misura assoluta di correttezza del match.

Il primo limite riguarda il fatto che l'area radar dipende dai punteggi assegnati alle singole dimensioni. Se una o più dimensioni sono valutate in modo impreciso, anche l'area finale può risentirne. Ad esempio, uno score basso nei livelli di rischio può ridurre l'area complessiva anche quando lo scenario di rischio è concettualmente coerente con il riferimento. Al contrario, una buona sovrapposizione su alcune dimensioni strutturate può aumentare parzialmente il profilo radar anche se il nucleo semantico del rischio non è pienamente sovrapponibile.

Un secondo limite riguarda la scelta delle dimensioni incluse nel radar. Nel presente lavoro sono state considerate otto componenti: attività, attori, strutture, livelli di rischio, scenario di rischio, scenario delle opportunità, cause e conseguenze. Questa scelta è coerente con la struttura della tabella di rischio, ma rappresenta comunque una semplificazione. Alcune informazioni più granulari, soprattutto nelle cause e nelle conseguenze, vengono infatti aggregate in due dimensioni sintetiche. Questo permette di mantenere il radar leggibile, ma può nascondere differenze presenti nelle singole sottocategorie. Ad esempio, due righe possono avere una buona corrispondenza nelle cause procedurali ma non nelle cause legate agli strumenti, oppure possono condividere conseguenze gestionali ma non conseguenze cliniche o normative. Poiché nel radar queste componenti vengono ricondotte a una media complessiva, parte del dettaglio viene inevitabilmente perso. Per questo motivo, nei casi più critici, la lettura dell'area radar dovrebbe essere affiancata al controllo delle sottodimensioni originarie.

Un ulteriore limite riguarda l'ordine degli assi del radar. L'area geometrica del radar plot dipende non solo dai valori delle singole dimensioni, ma anche dalla loro disposizione. Per ridurre questo problema, nel lavoro l'ordine degli assi è stato mantenuto fisso per tutti i confronti. Tuttavia, resta il fatto che l'area radar non è una misura completamente indipendente dalla rappresentazione grafica scelta. Essa va quindi considerata come un criterio operativo di confronto, non come una metrica universalmente valida.

Anche la selezione del best match presenta un limite interpretativo. Il codice seleziona, per ogni riga riferimento, la riga astratta con area radar più elevata tra quelle disponibili nella stessa fase. Tuttavia, questo non significa necessariamente che il match sia buono in senso assoluto. Se in una determinata fase sono disponibili pochi macro-rischi, o se nessuno di essi rappresenta adeguatamente la riga riferimento, il best match può risultare comunque debole. Per questo motivo, il valore dell'area radar deve sempre essere considerato insieme alla disponibilità di candidati e alla lettura qualitativa del profilo.

Nel complesso, l'area radar ha rappresentato un criterio utile per sintetizzare il confronto tra righe e selezionare i best match in modo sistematico. Il suo principale vantaggio è stato quello di evitare una valutazione basata su pesi arbitrari, mantenendo visibili le singole dimensioni del confronto. Tuttavia, i limiti discussi mostrano che il radar plot deve essere interpretato come uno strumento di supporto alla valutazione, non come una misura definitiva di qualità. La sua utilità maggiore risiede nella capacità di rendere il confronto più leggibile, tracciabile e orientato alla revisione esperta.

5.3 Limiti del riferimento e del dataset

Un elemento importante da considerare nell'interpretazione dei risultati riguarda i limiti del riferimento utilizzato per il confronto. La tabella di riferimento fornita dall'ospedale Policlinico San Matteo ha rappresentato una base indispensabile per valutare gli output della pipeline, perché costituiva il riferimento manuale disponibile. Tuttavia, non può essere considerata come una descrizione completa, uniforme e definitiva di tutti i rischi associati al percorso CAR-T.

Un primo limite riguarda la completezza dei campi e il livello di dettaglio delle descrizioni. In diversi casi, il riferimento presenta campi vuoti o solo parzialmente compilati, soprattutto nelle componenti più descrittive della tabella, come scenario delle opportunità, cause del rischio e alcune categorie di conseguenze. Questo aspetto ha influenzato il confronto automatico, perché la pipeline generava spesso macro-rischi più articolati e informativi, mentre il riferimento riporta descrizioni più sintetiche o mancanti in alcune colonne. Di conseguenza, il confronto non avveniva sempre tra due righe ugualmente ricche di contenuto. La differenza di granularità è stata una delle criticità principali. La tabella astratta era costruita per rappresentare macro-rischi derivati dall'accorpamento di più righe estratte, mentre il riferimento conteneva talvolta descrizioni più puntuali e talvolta descrizioni molto sintetiche. Il confronto automatico si è quindi trovato a valutare righe non sempre equivalenti per struttura informativa: da un lato macro-rischi articolati, dall'altro righe manuali spesso più brevi e con alcune colonne non compilate.

La presenza di campi vuoti ha avuto anche un impatto sulle metriche. Quando una dimensione del riferimento era assente o poco informativa, risultava difficile stabilire se un punteggio basso dipendesse da una reale incoerenza della riga astratta o semplicemente dalla mancanza di informazioni nel riferimento. Ad esempio, uno scenario delle opportunità non compilato nel riferimento poteva ridurre la possibilità di valutare correttamente la corrispondenza su quella dimensione. Allo stesso modo, conseguenze riportate in modo molto generale potevano rendere meno chiaro il confronto con un output astratto più dettagliato.

Va inoltre considerato che i due approcci di confronto non gestiscono in modo equivalente i campi non valorizzati. In particolare, nel caso dello scenario delle opportunità, l'embedding restituisce un punteggio nullo quando il campo del riferimento è vuoto, mentre il LLM-as-a-judge può comunque assegnare un punteggio non nullo. Questa asimmetria può influenzare sia il valore medio della singola dimensione sia l'area radar complessiva e deve quindi essere considerata nell'interpretazione del confronto tra i due metodi.

La pipeline di estrazione e astrazione è stata applicata a tutte le prime dodici fasi considerate del processo CAR-T. Il riferimento utilizzato per il confronto, tuttavia, non conteneva righe associate a tutte queste fasi: in particolare, le fasi 4, 9, 10 e 11 non erano rappresentate nel riferimento. I macro-rischi prodotti dalla pipeline per tali fasi sono stati quindi mantenuti nell'output, ma non hanno potuto essere inclusi nel confronto quantitativo. Nelle sole fasi confrontabili, inoltre, una riga astratta mai selezionata come best match non deve essere interpretata automaticamente come errata, ma semplicemente come una riga che non ha costituito il miglior abbinamento per nessuna delle righe presenti nel riferimento. Anche la distribuzione delle righe tra le fasi non era omogenea. Alcune fasi presentavano più righe riferimento e più macro-rischi candidati, permettendo un confronto più articolato. Altre fasi, invece, avevano pochi candidati o una sola riga astratta disponibile. In questi casi, la selezione del best match poteva diventare obbligata, senza indicare necessariamente una buona corrispondenza. La fase 12 mostra bene questa criticità: la presenza di una sola riga astratta ha portato a confronti forzati con più righe del riferimento, anche quando il contenuto non era realmente sovrapponibile.

Infine, il lavoro è stato sviluppato su documentazione procedurale relativa a uno specifico contesto ospedaliero. Questo rende la pipeline aderente al caso di studio considerato, ma limita la possibilità di trasferire direttamente i risultati ad altri centri o ad altri percorsi clinico-organizzativi. Procedure,

responsabilità, strutture coinvolte e modalità operative possono infatti variare tra istituzioni diverse; l'applicazione della stessa metodologia in altri contesti richiederebbe quindi l'utilizzo della documentazione locale e la definizione di un riferimento adeguato allo specifico centro.

5.4 Sviluppi futuri

I risultati ottenuti in questa tesi mostrano che una pipeline basata su modelli linguistici può supportare l'analisi del rischio in ambito ospedaliero. Tuttavia, il lavoro svolto deve essere considerato come una prima valutazione metodologica, sviluppata su un perimetro controllato e con l'obiettivo principale di verificare la fattibilità dell'approccio. Per questo motivo, diversi aspetti potrebbero essere approfonditi, estesi e migliorati in lavori futuri, sia dal punto di vista tecnico sia dal punto di vista della validazione clinico-organizzativa.

Un primo sviluppo riguarda l'estensione della pipeline all'intero processo CAR-T. Nel presente lavoro l'analisi è stata limitata alle prime dodici fasi del percorso, in modo da mantenere il lavoro gestibile e permettere una valutazione più controllata delle diverse componenti della pipeline. Tuttavia, il processo complessivo comprende ventiquattro fasi. Un'estensione naturale consisterebbe quindi nell'applicare la stessa metodologia anche alle fasi successive, così da ottenere una tabella di rischio più completa e verificare il comportamento della pipeline anche in momenti diversi del percorso, caratterizzati da contenuti clinici, logistici, organizzativi e documentali differenti.

Un secondo sviluppo riguarda l'abbinamento automatico tra documenti procedurali e fasi del processo. In questa tesi l'abbinamento manuale è stato utilizzato come riferimento operativo, perché ha garantito una copertura più ampia e controllata delle fasi analizzate. L'abbinamento automatico ha mostrato invece un comportamento più conservativo, individuando soprattutto le associazioni più esplicite e tralasciando alcuni collegamenti indiretti o trasversali. In futuro, questa fase potrebbe essere migliorata attraverso prompt più dettagliati, modelli più performanti o strategie ibride, in cui il modello viene utilizzato come supporto preliminare e l'associazione finale viene comunque sottoposta a controllo da parte di esperti del settore.

Un possibile miglioramento dell'abbinamento automatico potrebbe consistere anche nell'utilizzo combinato di embedding e LLM. Gli embedding potrebbero essere utilizzati per individuare i documenti più vicini a ciascuna fase del processo, mentre un modello linguistico potrebbe successivamente valutare in modo più interpretativo la pertinenza dell'associazione. Questo approccio consentirebbe di ridurre il carico manuale iniziale e, allo stesso tempo, limitare il rischio di perdere documenti rilevanti ma non esplicitamente collegati alla descrizione della fase.

Anche la fase di estrazione potrebbe essere ulteriormente perfezionata. In futuro, il prompt di estrazione potrebbe essere reso più vincolante, attraverso una strategia di *few-shot prompting,* includendo esempi positivi e negativi per ciascun campo della tabella. Ad esempio, sarebbe utile fornire al modello esempi di vere cause del rischio, distinguendole da dettagli procedurali, conseguenze cliniche o descrizioni generiche dell'attività. Questo potrebbe ridurre la quantità di informazioni non perfettamente collocate nei campi corretti (14).

La fase di astrazione rappresenta un'altra area importante di sviluppo. In futuro, questo passaggio potrebbe essere reso ancora più controllabile introducendo criteri più espliciti per l'unione. Ad esempio, il modello potrebbe essere guidato a distinguere con maggiore precisione le righe da accorpare, le righe da mantenere separate e le righe da segnalare come potenzialmente rumorose o poco pertinenti.

Un ulteriore miglioramento potrebbe consistere nella maggiore formalizzazione dei criteri di raggruppamento. Nel presente lavoro, l'unione è stata guidata dal prompt e dalla capacità del modello di riconoscere ridondanze semantiche tra le righe estratte. In futuro, sarebbe utile definire criteri più strutturati per stabilire quando due righe possono essere considerate appartenenti allo stesso macro-rischio, ad esempio valutando la somiglianza tra scenari di rischio, la coerenza delle cause, la sovrapposizione delle conseguenze e l'appartenenza alla stessa categoria tassonomica. Questo renderebbe il processo di astrazione più controllabile e coerente, pur senza eliminare la variabilità legata al funzionamento probabilistico del modello.

Un ulteriore sviluppo riguarda la mappa di unione. Nel presente lavoro, questo output ha avuto un ruolo fondamentale per rendere tracciabile il passaggio dalle righe estratte ai macro-rischi. In futuro, la mappa potrebbe essere arricchita con informazioni aggiuntive, ad esempio il grado di similarità tra le righe accorpate, la motivazione dettagliata dell'unione, le eventuali correzioni introdotte dal modello e i campi modificati rispetto all'input. In questo modo, la mappa non sarebbe soltanto uno strumento di tracciabilità, ma anche uno strumento utile per controllare in modo più puntuale il processo di astrazione.

Un aspetto centrale per sviluppi futuri riguarda la costruzione di un riferimento più completa e uniforme. Come discusso nel paragrafo precedente, il riferimento utilizzato presentava talvolta campi vuoti, descrizioni molto sintetiche e un livello di dettaglio non omogeneo tra le righe. Questo ha reso più complesso il confronto automatico, perché la tabella astratta prodotta dalla pipeline era spesso più articolata rispetto alla riga manuale di riferimento. Una futura estensione del lavoro dovrebbe quindi prevedere un riferimento revisionato e completato in modo sistematico, con campi compilati in maniera più uniforme per scenario, opportunità, cause e conseguenze.

La validazione del riferimento dovrebbe idealmente coinvolgere più esperti di dominio. La presenza di valutatori multipli permetterebbe di ridurre la dipendenza da una singola interpretazione manuale e di ottenere un riferimento più robusto. Inoltre, sarebbe possibile valutare il grado di accordo tra esperti e confrontarlo con le decisioni della pipeline. Questo passaggio sarebbe particolarmente utile per capire se le discrepanze tra modello e riferimento dipendono da errori della pipeline, da ambiguità del rischio o da limiti della tabella manuale utilizzata come riferimento.

Dal punto di vista della valutazione automatica, uno sviluppo importante riguarda la definizione di soglie interpretative per l'area radar. Nel presente lavoro, l'area radar normalizzata è stata utilizzata per selezionare il best match e per descrivere la qualità del confronto, ma non sono state definite soglie validate per stabilire automaticamente se un match sia forte, parziale o debole. In futuro, sarebbe utile confrontare le aree radar con giudizi manuali di esperti e definire intervalli interpretativi. Ad esempio, si potrebbero individuare soglie indicative al di sopra delle quali un match può essere considerato plausibile e soglie inferiori al di sotto delle quali il match richiede revisione obbligatoria.

Un ulteriore sviluppo potrebbe riguardare l'introduzione di una pesatura, da parte di esperti, delle diverse dimensioni del radar plot. Nel presente lavoro, le dimensioni del confronto sono state mantenute separate e valutate in modo multidimensionale, evitando di attribuire pesi arbitrari alle varie componenti della tabella. Tuttavia, in una futura evoluzione della pipeline, potrebbe essere utile definire pesi specifici per le diverse dimensioni, sulla base del giudizio di esperti clinici, organizzativi e di risk management.

Infatti, non tutte le componenti della tabella hanno necessariamente la stessa importanza nella valutazione della qualità di un match. Ad esempio, la coerenza dello scenario di rischio, delle cause e delle conseguenze potrebbe essere considerata più rilevante rispetto ad altre dimensioni più descrittive o organizzative, come attori o strutture coinvolte. Allo stesso modo, in alcuni contesti la corretta classificazione tassonomica del rischio potrebbe assumere un peso maggiore, soprattutto se la tabella viene utilizzata non solo per descrivere i rischi, ma anche per supportare attività di prioritizzazione, revisione gestionale o pianificazione di azioni correttive.

Una possibile estensione consisterebbe quindi nel raccogliere il parere di un gruppo di esperti e utilizzarlo per attribuire un peso relativo a ciascuna dimensione del radar. In questo modo, un eventuale score sintetico derivato dal profilo multidimensionale non dipenderebbe da una scelta arbitraria, ma da una valutazione condivisa e motivata. Questo permetterebbe di dare maggiore importanza alle componenti considerate più critiche per l'analisi del rischio e di rendere il criterio di selezione del best match più aderente alle priorità reali del processo CAR-T.

L’introduzione di pesi esperti potrebbe inoltre essere affiancata da un’analisi di sensibilità. Tale analisi permetterebbe di verificare quanto la selezione dei best match cambi al variare dei pesi attribuiti alle diverse dimensioni. Se piccole variazioni dei pesi producessero match molto diversi, ciò indicherebbe una maggiore instabilità del criterio di valutazione; al contrario, risultati stabili al variare dei pesi rafforzerebbero l’affidabilità del metodo. In questo modo, la pesatura non sarebbe solo definita da esperti, ma anche valutata dal punto di vista della robustezza.

Un ulteriore miglioramento potrebbe riguardare la valutazione statistica del confronto tra embedding e LLM-as-a-judge. Nel presente lavoro sono state analizzate medie, mediane, deviazioni standard e distribuzioni tramite boxplot. In futuro, con un dataset più ampio, sarebbe possibile applicare test statistici per verificare se le differenze tra i due approcci siano significative. Questo permetterebbe di rafforzare quantitativamente l’interpretazione secondo cui il judge risulta più discriminante rispetto all’embedding nella valutazione semantica finale.

Un altro possibile sviluppo futuro potrebbe riguardare il confronto tra diversi modelli di embedding. In questa tesi è stato utilizzato BGE-M3, scelto per la sua natura multilingue e per la possibilità di lavorare su testi in italiano. Tuttavia, sarebbe interessante confrontarlo con modelli più specializzati sul dominio biomedico o clinico, ad esempio modelli basati su PubMedBERT o altri sentence transformer addestrati su testi sanitari. Questo confronto dovrebbe però tenere conto della lingua dei documenti, poiché molti modelli biomedici sono addestrati prevalentemente su testi in inglese, mentre la documentazione procedurale analizzata è in italiano.

Dal punto di vista applicativo, la pipeline potrebbe essere estesa anche alla valutazione delle azioni correttive e preventive. Nel presente lavoro, l’attenzione si è concentrata soprattutto sull’identificazione e sul confronto di scenari, cause e conseguenze. Tuttavia, in un’applicazione pienamente orientata all’analisi del rischio, sarebbe utile verificare anche se le azioni correttive proposte siano coerenti con le cause individuate e con il livello di rischio. Questo permetterebbe di avvicinare ulteriormente la pipeline alle esigenze operative di metodologie come HFMEA e FMECA.

Un ulteriore sviluppo potrebbe riguardare l’integrazione di una valutazione di priorità del rischio. La pipeline sviluppata in questa tesi si concentra principalmente sull’identificazione, organizzazione e confronto delle informazioni di rischio. In futuro, il sistema potrebbe essere esteso per supportare anche la stima o la revisione di indicatori di priorità, come gravità, probabilità o rilevabilità, sempre con supervisione esperta. Questo permetterebbe di collegare la tabella prodotta non solo alla descrizione del rischio, ma anche alla sua prioritizzazione operativa.

Un ulteriore sviluppo riguarda l’aggiornamento periodico della tabella di rischio. I documenti procedurali possono essere modificati nel tempo, a seguito di aggiornamenti organizzativi, nuove indicazioni cliniche, modifiche dei flussi logistici o cambiamenti nelle responsabilità operative. Una pipeline semi-automatica potrebbe supportare anche il monitoraggio delle modifiche documentali, evidenziando quali rischi potrebbero essere cambiati, quali macro-rischi dovrebbero essere aggiornati e quali nuove criticità emergono dalle versioni più recenti delle procedure.

Infine, sarebbe utile valutare la capacità di generalizzazione della pipeline in contesti diversi. Il lavoro è stato sviluppato su documentazione relativa al processo CAR-T di uno specifico contesto ospedaliero. In futuro, lo stesso approccio potrebbe essere applicato ad altri centri, ad altri percorsi terapeutici complessi o ad altri processi sanitari ad alto rischio. Questo permetterebbe di capire quali componenti della pipeline siano generalizzabili e quali, invece, richiedano adattamento al dominio, alla struttura organizzativa o alla documentazione disponibile.

Nel complesso, gli sviluppi futuri dovrebbero quindi procedere in più direzioni complementari: estendere la pipeline all’intero processo CAR-T, migliorare l’abbinamento automatico documento-fase, rafforzare la qualità dell’estrazione e dell’astrazione, costruire un riferimento più completo, validare le soglie dell’area radar, introdurre con il supporto di esperti eventuali pesi differenziati per

le diverse dimensioni del confronto e integrare in modo più strutturato la revisione umana. In questa prospettiva, la pipeline proposta potrebbe evolvere da prototipo metodologico a strumento di supporto operativo per l'analisi, la revisione e l'aggiornamento delle tabelle di rischio in contesti sanitari complessi.

6. Conclusioni

Il presente lavoro di tesi ha avuto come obiettivo lo sviluppo e la valutazione di una pipeline basata su modelli linguistici di grandi dimensioni per supportare l'analisi automatizzata delle informazioni di rischio nel processo di terapia CAR-T. Il contesto di partenza è rappresentato da un percorso clinico-organizzativo complesso, composto da numerose fasi, diversi attori coinvolti, documenti procedurali eterogenei e attività ad alto impatto sulla qualità e sulla sicurezza del processo. In questo scenario, l'analisi del rischio richiede normalmente una lettura approfondita della documentazione, l'individuazione dei passaggi critici, la classificazione dei rischi e la costruzione di tabelle strutturate che possano essere utilizzate per attività di revisione, monitoraggio e miglioramento.

La complessità del processo CAR-T rende questo tipo di attività particolarmente dispendiosa. Le informazioni rilevanti per l'analisi del rischio non sono infatti concentrate in un unico documento, ma distribuite all'interno di procedure operative, allegati, istruzioni e documenti di processo. Inoltre, tali informazioni non sono sempre formulate esplicitamente come rischi, cause o conseguenze, ma devono spesso essere dedotte a partire dalla descrizione delle attività, dalle responsabilità assegnate, dai controlli previsti e dai passaggi operativi riportati nei documenti. Per questo motivo, l'utilizzo di strumenti automatici basati su modelli linguistici può rappresentare un supporto interessante, non per sostituire la valutazione esperta, ma per facilitare la raccolta, la strutturazione e il confronto delle informazioni.

In questa prospettiva, la tesi ha proposto una pipeline articolata in più fasi: definizione delle fasi del processo CAR-T, selezione dei documenti procedurali, abbinamento tra documenti e fasi, estrazione strutturata delle informazioni di rischio, astrazione delle righe estratte in macro-rischi e confronto automatico con un riferimento manuale. Il lavoro ha cercato di costruire un flusso metodologico completo, capace di trasformare documenti procedurali non pensati per l'elaborazione automatica in una tabella di rischio standardizzata, tracciabile e confrontabile. Un primo elemento rilevante del lavoro riguarda l'utilizzo delle fasi operative del processo CAR-T come unità principale di analisi. La suddivisione del percorso, già definita nella mappatura del processo, è stata utilizzata per organizzare la documentazione, delimitare il contesto di estrazione e confrontare successivamente solo righe appartenenti alla stessa fase.

Questa scelta ha permesso di ridurre l'ambiguità del confronto, evitando di mettere in relazione rischi riferiti a fasi diverse e quindi non realmente comparabili. Inoltre, l'abbinamento tra documenti e fasi ha rappresentato un passaggio fondamentale per individuare, per ogni fase, la documentazione più rilevante da analizzare.

La fase di estrazione ha rappresentato il primo passaggio effettivamente automatico della pipeline. A partire dai documenti associati a ciascuna fase, il modello è stato guidato a produrre una tabella strutturata. Questo schema ha permesso di organizzare le informazioni in modo coerente, mantenendo campi di tracciabilità e distinguendo attività, strutture coinvolte, attori, categorie di rischio, scenario di rischio, scenario delle opportunità, cause e conseguenze. La presenza di uno schema fisso ha avuto un ruolo importante, perché ha trasformato il compito del modello da una semplice generazione libera di testo a una produzione vincolata di informazioni organizzate.

L'estrazione automatica, tuttavia, non è risultata sufficiente da sola per ottenere una tabella finale direttamente utilizzabile. Gli output prodotti presentavano infatti ridondanza, differenze di granularità e alcune incoerenze nella classificazione dei rischi. In diversi casi, righe diverse descrivevano aspetti molto simili dello stesso problema, oppure riportavano cause e conseguenze non sempre collocate nel campo più appropriato. Anche i livelli di rischio sono risultati una componente complessa, poiché il modello tendeva talvolta a classificare il rischio sulla base delle sue possibili conseguenze economiche o reputazionali, più che sulla natura principale dell'evento descritto.

Per affrontare queste criticità è stata introdotta la fase di astrazione. Questo passaggio ha avuto un ruolo centrale nella pipeline, perché ha consentito di passare da una tabella estratta inizialmente ridondante a una tabella più sintetica, composta da macro-rischi. L'astrazione non ha svolto soltanto una funzione di riduzione numerica, ma anche una funzione di normalizzazione semantica. Il modello è stato infatti utilizzato per raggruppare righe simili, sintetizzare scenari sovrapposti, riorganizzare cause e conseguenze e, quando necessario, riformulare la classificazione tassonomica in modo più coerente con il contenuto principale del rischio.

Un elemento importante della fase di astrazione è stata la produzione della mappa di unione. Questo output ha permesso di mantenere la tracciabilità tra righe estratte e macro-rischi finali, rendendo possibile ricostruire quali righe iniziali fossero state accorpate in ciascun output astratto. Tale aspetto è particolarmente importante in un contesto sanitario, perché ogni trasformazione automatica deve poter essere verificata. La sintesi prodotta dal modello non può essere accettata come una "scatola nera": deve essere possibile controllare da quali informazioni deriva, quali righe sono state accorpate e se il macro-rischio finale rappresenta effettivamente il contenuto degli input originari.

La fase successiva della pipeline ha riguardato il confronto con un riferimento manuale. Il confronto è stato impostato secondo una logica riga riferimento–riga astratta: per ogni riga della tabella di riferimento, sono state confrontate tutte le righe astratte appartenenti alla stessa fase, selezionando come best match la riga con il profilo di confronto migliore. Questa impostazione ha permesso di valutare in che misura i macro-rischi generati automaticamente fossero in grado di coprire i rischi riportati nel riferimento.

La selezione del best match ha permesso di individuare, per ciascuna riga riferimento, il macro-rischio astratto con la maggiore coerenza complessiva tra quelli disponibili nella stessa fase. Tuttavia, uno degli aspetti più importanti emersi dal lavoro è che il concetto di best match deve essere interpretato con attenzione. La migliore riga disponibile non coincide necessariamente con una riga realmente corretta o pienamente sovrapponibile al riferimento. In alcuni casi, infatti, il match può essere debole, soprattutto quando il contenuto del riferimento non è adeguatamente rappresentato nella tabella astratta.

Per rendere il confronto più interpretabile, la valutazione è stata organizzata attraverso radar plot. Ogni confronto è stato descritto mediante otto dimensioni: attività, attori, strutture, livelli di rischio, scenario di rischio, scenario delle opportunità, cause e conseguenze. Questa scelta ha permesso di evitare una riduzione immediata del match a un unico punteggio complessivo, mantenendo invece visibili le diverse componenti della corrispondenza. In un'analisi del rischio, infatti, la qualità di un match non dipende da un solo aspetto: due righe possono avere uno scenario simile ma cause diverse, oppure condividere attori e strutture ma differire nei livelli di rischio o nelle conseguenze.

L'area radar normalizzata è stata utilizzata come indicatore sintetico per selezionare il best match. Essa ha permesso di ordinare i confronti in modo sistematico, favorendo i profili più ampi e bilanciati. Tuttavia, l'area radar non è stata interpretata come una misura assoluta di correttezza. Il suo valore principale è stato quello di fornire un criterio operativo per confrontare le righe e, allo stesso tempo, mantenere la possibilità di analizzare separatamente le singole dimensioni del radar. Questo approccio ha reso la valutazione più trasparente e più adatta a una successiva revisione esperta.

Un aspetto centrale della tesi è stato il confronto tra due strategie di valutazione semantica: embedding e LLM-as-a-judge. L'approccio basato su embedding ha permesso di confrontare rapidamente i campi testuali attraverso la similarità coseno. Questo metodo si è dimostrato utile come baseline e come possibile strumento di filtro preliminare, perché consente di valutare un numero elevato di confronti con costi computazionali relativamente contenuti. Tuttavia, i risultati hanno mostrato una tendenza degli embedding a produrre punteggi più compressi e meno discriminanti, soprattutto nei casi in cui righe diverse appartenevano allo stesso contesto generale ma descrivevano rischi operativamente distinti.

L’approccio LLM-as-a-judge ha invece mostrato una maggiore capacità interpretativa. Invece di misurare soltanto la vicinanza semantica tra testi, il modello ha valutato se la riga astratta coprisse effettivamente il contenuto del riferimento secondo criteri definiti nel prompt. Questo approccio si è rivelato particolarmente utile nelle dimensioni più complesse della tabella, come scenario di rischio, scenario delle opportunità, cause e conseguenze. I risultati hanno infatti mostrato che il judge tendeva a distinguere meglio tra match forti, match parziali e match deboli, producendo una distribuzione dei punteggi più ampia e maggiormente differenziata.

Il confronto tra embedding e LLM-as-a-judge suggerisce quindi che i due approcci non debbano essere considerati necessariamente alternativi in senso esclusivo. L’embedding può essere utile come strumento rapido, scalabile e adatto a una prima selezione dei candidati. Il judge, invece, appare più adatto alla valutazione semantica finale, soprattutto quando è necessario interpretare il significato operativo del rischio e distinguere tra righe concettualmente vicine ma non equivalenti. Una possibile evoluzione della pipeline potrebbe quindi essere rappresentata da un approccio ibrido, in cui gli embedding vengono utilizzati per ridurre il numero di confronti e il judge viene applicato successivamente ai match più plausibili o più critici.

Nonostante i risultati ottenuti, il presente studio presenta alcuni limiti, legati principalmente all’analisi delle sole prime dodici fasi del processo CAR-T, alle caratteristiche del riferimento disponibile e all’assenza di una validazione sistematica degli output da parte di esperti di dominio. Tali aspetti devono essere considerati nell’interpretazione dei risultati, che si inseriscono nell’ambito di uno studio di fattibilità.

In conclusione, la pipeline proposta ha mostrato la fattibilità di un approccio automatizzato per supportare l’analisi del rischio in un processo sanitario complesso come quello della terapia CAR-T. Il lavoro ha permesso di passare da documenti procedurali eterogenei a una tabella strutturata di macro-rischi, mantenendo la tracciabilità degli accorpamenti e introducendo un confronto multidimensionale con un riferimento manuale. I risultati indicano che i modelli linguistici possono contribuire in modo significativo alla raccolta, organizzazione e valutazione preliminare delle informazioni di rischio, soprattutto se inseriti in una pipeline controllata e sottoposta a revisione esperta.

La pipeline non deve quindi essere interpretata come uno strumento sostitutivo del giudizio umano, ma come un supporto metodologico per rendere più rapida, sistematica e tracciabile l’analisi del rischio. In un ambito come quello CAR-T, caratterizzato da alta complessità clinica, organizzativa e documentale, questo tipo di approccio può aiutare a ridurre il carico manuale iniziale, evidenziare ridondanze, proporre sintesi strutturate e facilitare il confronto con riferimenti esistenti. Con ulteriori sviluppi, un riferimento più solido, una validazione esperta e criteri di valutazione più robusti, la pipeline potrebbe evolvere da prototipo sperimentale a strumento di supporto operativo per l’analisi, la revisione e l’aggiornamento delle tabelle di rischio in contesti sanitari complessi.

Appendice A – Schema Pydantic utilizzato nella fase di estrazione

Di seguito è riportato un estratto del codice relativo alla definizione dello schema Pydantic utilizzato per vincolare l'output della fase di estrazione. Sono riportati i modelli relativi ai tre blocchi principali dello schema e alla strategia di estrazione *divided*; la tassonomia completa dei rischi è stata omessa per brevità

A.1 Strutture coinvolte

```
from enum import Enum
from typing import List, Union, Literal
from pydantic import BaseModel, ConfigDict, Field, model_validator



class StrutturaCoinvolta(str, Enum):
    DITTA_PRODUTTRICE_DEL_FARMACO_CAR = 'Ditta produttrice del farmaco CAR'
    RIANIMATORI = 'Rianimatori'
    NEUROLOGI_E_NEURORADIOLOGI = 'Neurologi e neuroradiologi'
    FARMACIA = 'Farmacia'
    UNITA_CLINICA = 'Unità clinica'
    UNITA_RACCOLTA_PBSC = 'Unità raccolta PBSC (Peripheral Blood Stem Cells)'
    TE = 'TE (Tissue Establishment)'
```

A.2 Struttura gerarchica della classificazione del rischio

```
class RiskName(BaseModel):
    model_config = ConfigDict(extra="forbid")


    risk: Union[
        RiskL1ClinicoSanitari,
        RiskL1Esterni,
        RiskL1Finanziari,
        RiskL1Strategici,
        RiskL1Operativi,
        RiskL1Compliance
    ] = Field(
        ...,
        discriminator="type",
        description=(
            "Tassonomia dei rischi con gerarchia ferrea "
```

```
            "basata su discriminator."
        )
    )
```

A.3 Ownership del rischio

```
class OwnershipDelRischio(BaseModel):
    model_config = ConfigDict(extra="forbid")

    attivita: str = Field(
        ...,
        description=(
            "Attivita principale della fase corrente. "
            "Deve descrivere solo il processo operativo centrale della fase analizzata."
        )
    )

    struttura_principale: StrutturaCoinvolta = Field(
        ...,
        description=(
            "Struttura organizzativa principalmente responsabile "
            "o direttamente titolare della fase corrente."
        )
    )

    attori: List[str] = Field(
        ...,
        description=(
            "Ruoli o figure professionali direttamente coinvolte "
            "nell esecuzione della fase corrente."
        )
    )

    altre_strutture_coinvolte: List[StrutturaCoinvolta] = Field(
        ...,
        description=(
            "Altre strutture organizzative coinvolte nella fase corrente "
            "oltre alla struttura principale."
        )
    )
```

```
    @model_validator(mode="after")
    def validate_structures(self):
        principale = self.struttura_principale
        altre = set(self.altre_strutture_coinvolte or [])

        if principale in altre:
            raise ValueError(
                f"La struttura {principale.value} non puo comparire sia "
                "in struttura_principale sia in altre_strutture_coinvolte."
            )

        return self
```

A.4 Anagrafica del rischio

```
class AnagraficaDelRischio(BaseModel):
    model_config = ConfigDict(extra="forbid")

    risk_name: RiskName = Field(
        ...,
        description=(
            "Classificazione del rischio secondo "
            "la tassonomia gerarchica ufficiale."
        )
    )

    scenario_di_rischio: str = Field(
        ...,
        description=(
            "Descrizione chiara e specifica dello scenario "
            "di rischio associato alla fase corrente."
        )
    )

    scenario_delle_opportunita: str = Field(
        ...,
        description=(
```

```
            "Descrizione di una opportunita di miglioramento "
            "collegata alla gestione della fase o alla mitigazione del rischio."
        )
    )
```

A.5 Analisi del rischio

```
class AnalisiDelRischio(BaseModel):
    model_config = ConfigDict(extra="forbid")

    principali_cause_del_rischio_procedure: List[str] = Field(
        ...,
        description="Cause del rischio riconducibili a procedure e workflow."
    )

    principali_cause_del_rischio_persone: List[str] = Field(
        ...,
        description="Cause del rischio riconducibili al fattore umano."
    )

    principali_cause_del_rischio_fattori_esterni: List[str] = Field(
        ...,
        description="Cause del rischio riconducibili a fattori esterni."
    )

    principali_cause_del_rischio_strumenti: List[str] = Field(
        ...,
        description=(
            "Cause del rischio riconducibili a strumenti, "
            "apparecchiature, sistemi o materiali."
        )
    )

    principali_conseguenze_del_rischio_economico_finanziarie: List[str] = Field(
        ...,
        description="Conseguenze del rischio con impatto economico o finanziario."
    )

    principali_conseguenze_del_rischio_danno_immagine: List[str] = Field(
```

```
    ...,
    description="Conseguenze del rischio in termini di danno reputazionale."
  )

  principali_conseguenze_del_rischio_salute_e_sicurezza: List[str] = Field(
    ...,
    description=(
      "Conseguenze del rischio sulla salute e sicurezza "
      "di pazienti e operatori."
    )
  )

  principali_conseguenze_del_rischio_compliance_normativa: List[str] = Field(
    ...,
    description="Conseguenze relative alla non conformita normativa."
  )

  principali_conseguenze_del_rischio_gestionale_operativo: List[str] = Field(
    ...,
    description="Conseguenze del rischio sul piano gestionale e operativo."
  )

  principali_conseguenze_del_rischio_ambiente: List[str] = Field(
    ...,
    description="Conseguenze del rischio con impatto ambientale."
  )
```

A.6 Output Complessivo

```
class RiskExtractionOutput(BaseModel):
  model_config = ConfigDict(extra="forbid")

  ownership_del_rischio: OwnershipDelRischio
  anagrafica_del_rischio: AnagraficaDelRischio
  analisi_del_rischio: AnalisiDelRischio
```

A.7 Schemi utilizzati nella strategia divided

```
class OwnershipOnlyOutput(BaseModel):
    model_config = ConfigDict(extra="forbid")
    ownership_del_rischio: OwnershipDelRischio


class AnagraficaOnlyOutput(BaseModel):
    model_config = ConfigDict(extra="forbid")
    anagrafica_del_rischio: AnagraficaDelRischio


class AnalisiOnlyOutput(BaseModel):
    model_config = ConfigDict(extra="forbid")
    analisi_del_rischio: AnalisiDelRischio
```

APPENDICE B – Esempio di system prompt utilizzato nella procedura di estrazione in configurazione single shot

Di seguito è riportato un estratto del *system prompt* utilizzato nella procedura di estrazione *single shot*, contenente le principali istruzioni fornite al modello per guidare l'analisi e vincolare la struttura dell'output.

system_prompt = f"""

Sei un assistente specializzato nell estrazione strutturata di informazioni da documenti Standard Operative Procedure relativi al processo CAR-T:

Chimeric Antigen Receptor T-Cell therapy.

Devi analizzare il testo del documento e compilare un output JSON rigorosamente conforme allo schema fornito.

OBIETTIVO

Per ogni fase richiesta in input devi estrarre:

- ownership del rischio

- anagrafica del rischio

- analisi del rischio

L output deve essere specifico della fase corrente, non del documento in generale.

REGOLE GENERALI

- Devi usare esclusivamente le fasi che ti vengono fornite in input.

- Non inventare fasi.

- Non omettere fasi.

- Non accorpare fasi.

- Ogni elemento dell output deve riferirsi solo alla fase corrente analizzata.

- Non usare contenuti generici riferiti all intero documento se non sono pertinenti alla fase corrente.

- Usa solo informazioni presenti nel testo o chiaramente deducibili dal testo.

- Non inventare dettagli non supportati.

- Restituisci solo JSON valido conforme allo schema.

REGOLE PER L OTTICA DI ANALISI

L analisi deve essere sempre svolta dal punto di vista dell organizzazione ospedaliera che partecipa al processo CAR-T.

Anche quando la fase coinvolge la ditta produttrice del farmaco CAR, il rischio non deve essere descritto dal punto di vista interno della ditta produttrice.

Devi estrarre il rischio per l ospedale, cioe il rischio che ricade su strutture ospedaliere, paziente, operatori, continuita del percorso, compliance, tracciabilita, organizzazione, coordinamento o gestione clinico-operativa ospedaliera.

La ditta produttrice puo comparire come struttura coinvolta, controparte, fornitore, soggetto esterno o dipendenza operativa, ma:

- non descrivere rischi interni alla produzione industriale se non hanno un impatto sul percorso ospedaliero

- non descrivere cause che appartengono esclusivamente alla governance interna della ditta
- non formulare opportunita come miglioramenti interni della ditta produttrice
- non assegnare all ospedale responsabilita che sono esclusivamente industriali o produttive
- non trasformare la fase in una valutazione del processo aziendale della ditta

Quando la ditta produttrice e coinvolta, formula invece il rischio in termini ospedalieri, ad esempio:
- ritardi o criticita nella ricezione, verifica o trasmissione di informazioni tra ospedale e ditta
- difetti di coordinamento tra strutture ospedaliere e ditta produttrice
- incompletezza della documentazione richiesta per il rilascio, spedizione, produzione o ricezione del prodotto
- mancata verifica ospedaliera di requisiti, tempistiche, tracciabilita, chain of custody o chain of identity
- difficolta organizzative dell ospedale nel gestire indisponibilita, ritardi, non conformita o richieste integrative della ditta
- impatti sul paziente, sul calendario clinico, sulla continuita del percorso o sulla compliance ospedaliera

Il punto di vista corretto e:
"Quale rischio corre l ospedale in questa fase, anche se la fase dipende o interagisce con la ditta produttrice?"

Il punto di vista scorretto e:
"Quale rischio corre la ditta produttrice nello svolgere le sue attivita interne?"

REGOLE PER LA CLASSIFICAZIONE DEL RISCHIO

Devi classificare il rischio usando esclusivamente la tassonomia ufficiale fornita sotto.
La tassonomia e ferrea:
- il livello 1 determina automaticamente i soli livelli 2 ammessi
- dove previsto, il livello 2 determina automaticamente i soli livelli 3 ammessi
- non inventare categorie o sottocategorie non presenti
- non scegliere categorie per somiglianza lessicale
- scegli la categoria in base al significato del rischio e allo scenario descritto

REGOLE PER IL LIVELLO 3

I soli L2 che prevedono un L3 sono:
- Informativa e reporting
- Salute, sicurezza e ambientale

Quando scegli uno di questi due L2:
- devi obbligatoriamente selezionare anche il livello 3 piu coerente
- devi usare il livello 3 per specializzare il rischio, non lasciarlo implicito
- non lasciare nullo il livello 3 se hai scelto uno di questi due L2
- lascia nullo il livello 3 solo se il livello 2 scelto non prevede alcun livello 3

REGOLE PER IL CAMPO "attivita"

Il campo attivita deve riferirsi esclusivamente alla fase corrente.

Deve rappresentare il processo operativo principale della fase corrente e non una attivita di altre fasi.

Regole obbligatorie per attivita:

- stile tabellare, tecnico operativo, sintetico

- non usare titoli del documento

- non usare intestazioni di sezione

- non usare nomi commerciali o nomi di azienda

- non descrivere complicanze, rischi, conseguenze o dettagli secondari

- non usare attivita appartenenti ad altre fasi del processo

- deve descrivere solo l azione principale della fase richiesta

Controllo obbligatorio prima di restituire attivita:

- l attivita appartiene chiaramente alla fase corrente?

- l attivita e valida solo per la fase scelta? Se no, e troppo generica perche va bene anche per altri fasi e va riscritta

- l attivita e una label di processo e non un titolo del documento?

Se una di queste condizioni non e soddisfatta, riscrivi attivita.

REGOLE PER ownership_del_rischio

- attivita: indica solo il processo operativo principale della fase corrente

- struttura_principale: indica la struttura principalmente responsabile della fase corrente

- attori: indica solo i ruoli direttamente coinvolti nella fase corrente

- altre_strutture_coinvolte: indica solo le strutture che partecipano concretamente alla fase oltre alla struttura principale

- non inserire strutture o attori marginali o solo citati nel documento

- non duplicare la struttura principale nelle altre strutture coinvolte

- se la fase coinvolge la ditta produttrice, distingui tra responsabilita operativa della ditta e rischio ospedaliero

- la struttura_principale deve essere la struttura che presidia la fase dal punto di vista del processo ospedaliero, quando identificabile dal documento

- usa "Ditta produttrice del farmaco CAR" come struttura principale solo se la fase descritta e effettivamente governata dalla ditta; anche in quel caso scenario, cause, conseguenze e opportunita devono essere formulati come impatto/rischio per l ospedale

- non descrivere ownership, cause o opportunita come se l obiettivo fosse migliorare il processo produttivo interno della ditta

REGOLE PER anagrafica_del_rischio

- risk_name: classifica il rischio in modo coerente con scenario, fase e tassonomia

- scenario_di_rischio: descrivi in modo chiaro cosa puo andare storto nella fase corrente

- scenario_delle_opportunita: descrivi un miglioramento concreto collegato alla corretta gestione della fase o alla mitigazione del rischio

- evita formulazioni vaghe, astratte o troppo generiche

- scenario_di_rischio deve essere formulato dal punto di vista dell ospedale e del percorso clinico-organizzativo ospedaliero

- se il rischio nasce da una dipendenza dalla ditta produttrice, descrivilo come rischio di coordinamento, tracciabilita, tempistiche, documentazione, continuita del percorso o impatto sul paziente per l ospedale

- scenario_delle_opportunita deve descrivere miglioramenti attuabili o presidiabili dall ospedale, non miglioramenti interni della ditta produttrice

REGOLE PER analisi_del_rischio

Le cause e le conseguenze devono essere:

- specifiche della fase corrente

- plausibili

- concrete

- non ripetitive

- non generiche

- coerenti con lo scenario di rischio individuato

- riferite alla fase corrente e non al documento nel suo complesso

Le cause e le conseguenze devono essere lette in ottica ospedaliera.

Quando una criticita riguarda la ditta produttrice, non descriverla come problema interno della ditta, ma come dipendenza, vincolo, ritardo, informazione incompleta, non conformita ricevuta, richiesta integrativa, indisponibilita o criticita di coordinamento che impatta l ospedale.

REGOLA FONDAMENTALE PER LE CAUSE

Una causa del rischio deve descrivere una carenza, fragilita, errore, omissione, malfunzionamento, indisponibilita, inadeguatezza o condizione critica che puo generare lo scenario di rischio.

Non devi inserire come causa:

- controlli gia previsti

- azioni correttive

- buone pratiche

- requisiti operativi corretti

- presidi di mitigazione

- strumenti descritti in modo positivo

- procedure eseguite correttamente

- attivita che rappresentano la soluzione al rischio

- semplici dettagli operativi copiati dal documento senza trasformazione critica

Se nel documento trovi un presidio positivo, una procedura corretta, un controllo o una misura di mitigazione, devi trasformarlo nella corrispondente fragilita negativa.

Esempi obbligatori di trasformazione:

- "uso di dispositivo dedicato" diventa "indisponibilita, malfunzionamento o uso non corretto del dispositivo dedicato"

- "utilizzo di contenitori validati" diventa "contenitori non idonei, non validati o non correttamente verificati"

- "sigillatura ermetica" diventa "sigillatura non corretta, non verificata o non conforme"

- "controllo da parte di due operatori" diventa "mancato doppio controllo o verifica incompleta da parte degli operatori"

- "personale formato" diventa "formazione insufficiente o competenze non uniformi del personale"

- "campionamento post raccolta" diventa "campionamento post raccolta non eseguito, non rappresentativo o non tracciato"

- "rispetto della procedura" diventa "procedura non rispettata, non aggiornata o applicata in modo incompleto"

- "uso di anticoagulante" diventa "dosaggio, preparazione o impiego non corretto dell anticoagulante"

- "raccolta con volume definito" diventa "volume di raccolta non correttamente verificato o non conforme ai parametri previsti"

Per le CAUSE:

- procedure: indicare carenze, errori, omissioni o criticita di workflow, sequenza operativa, istruzioni operative, modalita di esecuzione, controlli procedurali o tracciabilita della fase. Non indicare procedure corrette, controlli gia previsti o requisiti operativi come se fossero cause.

- persone: indicare criticita del fattore umano come formazione insufficiente, competenze non uniformi, errore umano, comunicazione incompleta, coordinamento carente, supervisione insufficiente, attenzione ridotta o mancato doppio controllo. Non indicare personale formato, competenza degli operatori o controlli correttamente eseguiti.

- fattori esterni: indicare solo criticita esterne alla struttura o al team operativo, come ritardi di fornitori, indisponibilita di servizi esterni, vincoli normativi esterni, eventi ambientali, dipendenze da soggetti terzi o condizioni non direttamente controllabili. Non inserire strumenti interni, dispositivi, contenitori o materiali usati nella fase.

- strumenti: indicare criticita di dispositivi, apparecchiature, sistemi informativi, documenti, contenitori, materiali, tecnologie o supporti tecnici, come malfunzionamento, indisponibilita, manutenzione insufficiente, taratura non verificata, configurazione errata, uso scorretto, inadeguatezza, danneggiamento o documentazione tecnica non aggiornata. Non indicare lo strumento come presidio positivo.

REGOLE DI ASTRAZIONE PER LE CAUSE

Le cause non devono essere una semplice copia del testo operativo del documento.

Devono rappresentare il motivo per cui lo scenario di rischio potrebbe verificarsi.

Quindi:

- evita frasi che iniziano con "uso di", "utilizzo di", "presenza di", "controllo di", "formazione di", "rispetto di", quando queste indicano una misura corretta

- preferisci formulazioni che iniziano con "mancata", "errata", "incompleta", "insufficiente", "inadeguata", "non corretta", "non verificata", "indisponibilita", "malfunzionamento"

- se una voce sembra una soluzione, un controllo o un requisito operativo, riscrivila come possibile failure mode

- se una causa e troppo simile a una azione correttiva, astrarla come criticita a monte

- se una categoria di causa non e supportata dal documento, usa una formulazione minima come "Non evidenziabile dal documento per la fase corrente"

Per le CONSEGUENZE:

- economico finanziarie: costi, sprechi, perdite, inefficienze economiche, consumo aggiuntivo di risorse o necessita di ripetere attivita

- danno immagine: perdita di fiducia, reputazione negativa, percezione di scarsa qualita o ridotta affidabilita del processo

- salute e sicurezza: danni clinici, eventi avversi, complicanze, lesioni, esposizioni o peggioramenti delle condizioni di sicurezza

- compliance normativa: non conformita, rilievi, violazioni, contestazioni, sanzioni o mancato rispetto di requisiti interni o normativi

- gestionale operativo: ritardi, interruzioni, rallentamenti, inefficienze, errori di processo, perdita di continuita operativa o necessita di rilavorazione

- ambiente: impatti ambientali solo se realmente pertinenti alla fase; se non pertinenti usare una formulazione minima e specifica

REGOLE DI QUALITA

Prima di restituire il JSON finale verifica sempre che:

- ogni campo sia riferito alla fase corrente e non al documento nel suo complesso

- attivita non sia il titolo di una sezione o del documento

- la classificazione del rischio sia coerente con scenario e tassonomia

- le cause siano suddivise correttamente per area

- ogni causa sia formulata come criticita o carenza, non come presidio positivo

- nessuna causa sia una semplice copia di una azione correttiva, di una buona pratica o di un controllo previsto

- nessuna causa inizi con formule come "uso di", "utilizzo di", "presenza di", "controllo di", "formazione di", se queste descrivono una misura corretta invece di una fragilita

- se una voce sembra una soluzione o un requisito operativo, riscrivila come possibile failure mode

- le conseguenze siano plausibili e specifiche

- non ci siano testi inutilmente lunghi o vaghi

- non ci siano duplicazioni evidenti tra campi

TASSONOMIA UFFICIALE DEI RISCHI

{taxonomy_prompt}

ORGANIGRAMMA DI RIFERIMENTO DEL PROCESSO CAR-T

{organigramma_prompt}

"""

APPENDICE C – Esempio di output JSON

Di seguito è riportato un esempio reale di output JSON relativo al blocco *Ownership del rischio*, ottenuto durante la procedura di estrazione *divided*.


```
{
  "ownership_del_rischio": {
    "attivita": "Trasporto del prodotto di raccolta al TE",
    "struttura_principale": "TE (Tissue Establishment)",
    "attori": [
      "operatori qualificati"
    ],
    "altre_strutture_coinvolte": [
      "Ditta produttrice del farmaco CAR"
    ]
  }
}
```

Appendice D – Esempio di prompt utilizzato per la fase di astrazione

Di seguito è riportato il system prompt utilizzato nella fase di astrazione. Il prompt definisce la logica cluster-first adottata per raggruppare le righe di input in macro-rischi, i criteri di unione, le regole di copertura delle righe e i vincoli relativi alla classificazione dei rischi.

SYSTEM_PROMPT = """

Sei un assistente esperto di risk management sanitario nel processo CAR-T.

Riceverai righe tabellari riferite alla stessa fase del processo.

Il tuo compito è fare astrazione cluster-first:

1. prima raggruppi le RIGA_INPUT in cluster/macro-gruppi;

2. poi ogni cluster diventa un solo macro-scenario di rischio;

3. la mappa di merge è la traccia dei cluster usati per generare ciascun macro-rischio.

Questa mappa NON deve essere una associazione postuma: deve rappresentare la vera logica con cui hai astratto il rischio.

Obiettivo metodologico:

- astrarre e normalizzare righe input anche se sono sporche, ridondanti, parziali o non perfettamente formulate;

- non fare pulizia aggressiva: tieni buono tutto il materiale ragionevolmente riconducibile alla fase;

- se una riga è rumorosa ma contiene elementi riconducibili alla fase, assorbila nel cluster pertinente più vicino;

- non creare un macro-rischio autonomo solo per coprire una riga sporca o fuori focus;

- escludi una riga dal merge solo se è chiaramente riferita a un’altra fase del processo CAR-T o completamente non pertinente.

Regole di clustering:

- accorpa righe con lo stesso problema di fondo, anche se cambiano parole, dettagli, cause, conseguenze o categoria originale;

- accorpa righe che richiederebbero lo stesso presidio, controllo, mitigazione o responsabilità operativa;

- mantieni separati solo cluster con evento critico realmente diverso;

- produci meno righe dell’input quando i rischi sono accomunabili;

- per fasi con poche righe input, cerca di assegnare tutte o quasi tutte le righe a un cluster, salvo errore evidente di fase.

Gestione di righe sporche:

- una riga può essere assegnata a un cluster anche se scenario_di_rischio o categoria originale non coincidono perfettamente;

- puoi usare attività, strutture, attori, contesto operativo, scenario, cause o conseguenze per decidere il cluster;

- se una riga contiene elementi fuori focus ma appartiene alla fase, usa solo la parte utile e assorbila in un macro-rischio pertinente alla fase;

- contenuti come finanza generale, comunicazione istituzionale, follow-up o strategia generale NON devono diventare macro-rischi autonomi di una fase tecnica, a meno che la fase sia davvero quella;

- quando sono forniti archetipi di fase, usali come famiglie di destinazione: una riga rumorosa va assorbita nell'archetipo più vicino, non trasformata in un nuovo rischio finale fuori focus.

Mappa del merge:

- per ogni elemento compila righe_input_usate con le etichette esatte delle righe che formano il cluster;

- ogni macro-rischio finale deve derivare direttamente dalle righe_input_usate;

- la mappa deve coprire tutte le RIGA_INPUT della fase: ogni etichetta RIGA_INPUT_1, RIGA_INPUT_2, ... deve comparire almeno una volta;

- non lasciare righe senza assegnazione: se una riga è rumorosa, generica o fuori focus ma appartiene alla fase, NON lasciarla fuori, assorbila nel cluster pertinente più vicino;

- non creare però un macro-rischio autonomo solo per quella riga rumorosa;

- compila criterio_di_merge con una frase sintetica che spiega perché quelle righe sono state raggruppate nello stesso cluster.

Classificazione:

- classifica in base alla natura principale del macro-scenario astratto, non alla categoria rumorosa della singola riga;

- un rischio operativo resta operativo anche se produce conseguenze economiche;

- usa una sola categoria L1 e una sola L2;

- usa L3 solo se prevista dalla tassonomia, altrimenti lascia stringa vuota "".

Ogni elemento generato deve essere un rischio generale, sintetico e riutilizzabile della fase.

Se una causa o conseguenza non è deducibile, usa lista vuota [].

Non restituire codice_doc, step_id o fase: li aggiunge il codice.

Restituisci solo JSON valido conforme allo schema.

"""

Appendice E – System prompt utilizzato nell’approccio LLM-as-a-judge

Di seguito è riportato il system prompt utilizzato nell’approccio LLM-as-a-judge. Il prompt definisce il compito di confronto tra la riga astratta e la riga di riferimento, la scala di valutazione da 0 a 5 e i criteri utilizzati per assegnare i punteggi alle diverse componenti testuali.

SYSTEM_PROMPT = """

Sei un valutatore esperto di risk management sanitario e processi CAR-T.

Devi confrontare una riga di rischio astratta con una riga reference appartenenti alla stessa fase del processo. La riga astratta può essere più generale, sintetica o accorpata rispetto alla reference.

Valuta se la riga astratta copre il nucleo concettuale della riga reference, senza richiedere corrispondenza letterale delle parole.

Usa una scala intera da 0 a 5:

0 = nessuna corrispondenza;

1 = corrispondenza molto debole o solo superficiale;

2 = corrispondenza parziale, con copertura limitata;

3 = corrispondenza discreta, con copertura concettuale parziale ma sensata;

4 = buona corrispondenza, con nucleo principale sostanzialmente coperto;

5 = corrispondenza molto forte o pienamente equivalente.

Criteri di giudizio:

- per l'attività, valuta se le due righe si riferiscono alla stessa azione o a un'attività operativamente compatibile;

- per lo scenario di rischio, valuta se il rischio astratto include o rappresenta il rischio specifico della reference;

- per lo scenario delle opportunità, valuta se le opportunità o mitigazioni sono coerenti;

- per ciascuna categoria di cause, valuta solo la copertura della stessa categoria tra riga astratta e riga reference: procedure con procedure, persone con persone, fattori esterni con fattori esterni, strumenti con strumenti;

- per ciascuna categoria di conseguenze, valuta solo la copertura della stessa categoria tra riga astratta e riga reference: economico-finanziarie, danno d'immagine, salute e sicurezza, compliance normativa, gestionale-operativo e ambiente.

Assegna punteggi alti quando la riga astratta è più generale ma include chiaramente il rischio della reference. Assegna punteggi bassi quando la somiglianza è solo generica, lessicale o di contesto ma il rischio sostanziale è diverso.

Non inventare informazioni non presenti.

""".strip()

Bibliografia

1. June CH, Sadelain M. Chimeric Antigen Receptor Therapy. N Engl J Med. 5 luglio 2018;379(1):64–73. doi:10.1056/NEJMra1706169

2. Kröger N, Gribben J, Chabannon C, Yakoub-Agha I, Einsele H, curatori. The EBMT/EHA CAR-T Cell Handbook [Internet]. Cham: Springer International Publishing; 2022 [citato 8 luglio 2026]. Disponibile su: https://link.springer.com/10.1007/978-3-030-94353-0 doi:10.1007/978-3-030-94353-0

3. Hayden PJ, Roddie C, Bader P, Basak GW, Bonig H, Bonini C, et al. Management of adults and children receiving CAR T-cell therapy: 2021 best practice recommendations of the European Society for Blood and Marrow Transplantation (EBMT) and the Joint Accreditation Committee of ISCT and EBMT (JACIE) and the European Haematology Association (EHA). Annals of Oncology. marzo 2022;33(3):259–75. doi:10.1016/j.annonc.2021.12.003

4. Rafq S, Hackett CS, Brentjens RJ. Engineering strategies to overcome the current roadblocks in CAR T cell therapy. Nat Rev Clin Oncol. marzo 2020;17(3):147–67. doi:10.1038/s41571-019-0297-y

5. El-Awady SMM. Overview of Failure Mode and Efects Analysis (FMEA): A Patient Safety Tool. Global Journal on Quality and Safety in Healthcare. 1 febbraio 2023;6(1):24–6. doi:10.36401/JQSH-23-X2

6. DeRosier J, Stalhandske E, Bagian JP, Nudell T. Using Health Care Failure Mode and Efect Analysis ™: The VA National Center for Patient Safety's Prospective Risk Analysis System. The Joint Commission Journal on Quality Improvement. maggio 2002;28(5):248–67. doi:10.1016/S1070-3241(02)28025-6

7. VA National Center for Patient Safety. Department of Veterans Afairs [Internet]. 2021. Healthcare Failure Mode and Efect Analysis (HFMEA) Guidebook. Disponibile su: https://www.patientsafety.va.gov/docs/joe/HFMEA-Guidebook-January2021.pdf

8. International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH). Integrated Addendum to ICH E6(R1): Guideline for Good Clinical Practice E6(R2). 2016.

9. International Organization for Standardization. ISO 9001:2015 — Quality management systems — Requirements. International Organization for Standardization; 2015. Report ISO 9001:2015.

10. Wiest IC, Wolf F, Leßmann ME, Van Treeck M, Ferber D, Zhu J, et al. A software pipeline for medical information extraction with large language models, open source and suitable for oncology. npj Precis Onc. 17 settembre 2025;9(1):313. doi:10.1038/s41698-025-01103-4

11. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention Is All You Need [Internet]. arXiv; 2017 [citato 28 agosto 2026]. Disponibile su: https://arxiv.org/abs/1706.03762 doi:10.48550/ARXIV.1706.03762

12. Devlin J, Chang MW, Lee K, Toutanova K. 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) [Internet]. Minneapolis, Minnesota: Association for Computational Linguistics; 2019 [citato 14 settembre 2026]. p. 4171–86. Disponibile su: https://aclanthology.org/N19-1423 doi:10.18653/v1/N19-1423

13. Rafel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, et al. Exploring the Limits of Transfer Learning with a Unifed Text-to-Text Transformer [Internet]. arXiv; 2019 [citato 14 settembre 2026]. Disponibile su: https://arxiv.org/abs/1910.10683 doi:10.48550/ARXIV.1910.10683

14. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language Models are Few-Shot Learners [Internet]. arXiv; 2020 [citato 14 settembre 2026]. Disponibile su: https://arxiv.org/abs/2005.14165 doi:10.48550/ARXIV.2005.14165

15. Sarawagi S. Information Extraction. Foundations and Trends in Databases. 30 novembre 2008;1(3):261–377. doi:10.1561/1900000003

16. Pons E, Braun LMM, Hunink MGM, Kors JA. Natural Language Processing in Radiology: A Systematic Review. Radiology. maggio 2016;279(2):329–43. doi:10.1148/radiol.16142770

17. Reimers N, Gurevych I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) [Internet]. Hong Kong, China: Association for Computational Linguistics; 2019 [citato 15 luglio 2026]. p. 3980–90. Disponibile su: https://www.aclweb.org/anthology/D19-1410 doi:10.18653/v1/D19-1410

18. Chen J, Xiao S, Zhang P, Luo K, Lian D, Liu Z. M3-Embedding: Multi-Linguality, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation [Internet]. arXiv; 2024 [citato 15 luglio 2026]. Disponibile su: https://arxiv.org/abs/2402.03216 doi:10.48550/ARXIV.2402.03216

19. Zheng L, Chiang WL, Sheng Y, Zhuang S, Wu Z, Zhuang Y, et al. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In: Advances in Neural Information Processing Systems 36 [Internet]. New Orleans, Louisiana, USA: Neural Information Processing Systems Foundation, Inc. (NeurIPS); 2023 [citato 15 luglio 2026]. p. 46595–623. Disponibile su: http://www.proceedings.com/075280-2020.html doi:10.52202/075280-2020

20. Croxford E, Gao Y, First E, Pellegrino N, Schnier M, Caskey J, et al. Evaluating clinical AI summaries with large language models as judges. npj Digit Med. 5 novembre 2025;8(1):640. doi:10.1038/s41746-025-02005-2

21. Talarmin C, Kerob S, Cartier F, Madelaine I, Mukenyi S, Schwartz E, et al. Quality risk management of the chimeric antigen receptor T cell pharmaceutical circuit in one of the frst qualifed European centers. Cytotherapy. dicembre 2020;22(12):792 –801. doi:10.1016/j.jcyt.2020.06.009

22. Denecke K, Paula H. Evaluating Large Language Models for Analysing Safety Risks in Healthcare Incident Reports. In: Househ MS, Tariq ZUA, Al-Zubaidi M, Shah U, Huesing E, curatori. Studies in Health Technology and Informatics [Internet]. IOS Press; 2025 [citato 28 agosto 2026]. Disponibile su: https://ebooks.iospress.nl/doi/10.3233/SHTI250867 doi:10.3233/SHTI250867

23. Nair SS, Court L, Douglas R, Gay S, Govyadinov P, Netherton T, et al. Use of Large Language Models to Enhance Failure Mode and Efects Analysis: A Case Study. Advances in Radiation Oncology. agosto 2026;11(8):102060. doi:10.1016/j.adro.2026.102060

24. Jaccard P. Étude comparative de la distribution forale dans une portion des Alpes et du Jura [Text/html,application/pdf,text/html]. 1901. doi:10.5169/SEALS-266450

25. Manning CD, Raghavan P, Schütze H. Introduction to information retrieval. Cambridge: Cambridge university press; 2008.

26. Liu Y, Iter D, Xu Y, Wang S, Xu R, Zhu C. G-Eval: NLG Evaluation using Gpt-4 with Better Human Alignment. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing [Internet]. Singapore: Association for Computational Linguistics; 2023 [citato 15 luglio

2026]. p. 2511–22. Disponibile su: https://aclanthology.org/2023.emnlp-main.153 doi:10.18653/v1/2023.emnlp-main.153