indaloop

AI × PROCESSI × AZIENDE

CASO STUDIO · REPORT COMPLETO

DDT IN INGRESSO

Meno dati da ricopiare, più tempo per gestire il magazzino

Come indaloop ha affiancato una torneria di precisione nella gestione dei documenti di trasporto.

> *Un caso reale di una torneria di precisione italiana. Per volontà dell'azienda, ne omettiamo il nome e i dettagli identificativi. Il progetto riguarda un'attività quotidiana: registrare i DDT dei fornitori e collegarli agli ordini corretti nel gestionale.*

Circa 120 documenti a settimana. Per ciascuno, qualcuno doveva leggere il foglio, cercare l'ordine, trovare la riga giusta, copiare i dati e archiviare il PDF. Un lavoro necessario, che occupava circa dodici ore alla settimana e si complicava quando arrivavano più materiali insieme.

indaloop ha lavorato con l'azienda per cambiare questo passaggio. Oggi l'operatore carica il documento, controlla i dati letti dal sistema e approva le modifiche proposte al gestionale. Il progetto ha richiesto **2 settimane** di lavoro, di cui 3 giorni serviti a verificare che una soluzione con OCR e modello linguistico open source, sull'hardware disponibile, non dava risultati adeguati sui documenti più impegnativi.

La scelta successiva è nata da quel confronto. Al centro del progetto sono rimasti il tempo di chi lavora in magazzino, la correttezza delle registrazioni e la possibilità di controllare ogni aggiornamento prima di applicarlo.

~120	6 min	35 s	€0,15
DDT a settimana, da decine di fornitori	per DDT prima del progetto: circa 12 ore a settimana	elaborazione media nel test, più il controllo dell'operatore	consumo AI per documento, circa €80 al mese

01

IL PROGETTO IN BREVE

AZIENDA	Torneria di precisione italiana
PROCESSO	Registrazione dei DDT in ingresso sugli ordini di acquisto e di conto lavoro nel gestionale
PRIMA	Lettura, ricerca dell'ordine, inserimento manuale e archiviazione: circa sei minuti per DDT, circa dodici ore a settimana
DOPO	Circa 35 secondi medi di elaborazione nel test, più uno o due minuti di controllo dell'operatore sui documenti ordinari; i casi complessi richiedono più attenzione
RUOLO DI INDALOOP	Analisi del processo, confronto delle alternative, integrazione con il gestionale e definizione dei controlli
DURATA	Due settimane dall'avvio alla messa in uso; tre giorni dedicati alla verifica della soluzione OCR + LLM open source
CONTROLLO UMANO	Ogni aggiornamento del gestionale richiede l'approvazione dell'operatore
CONSUMO AI	Circa €0,15 per documento, nell'ordine di €80 al mese ai volumi indicati
HARDWARE ACQUISTATO	Nessuno

I tempi di elaborazione provengono dal confronto interno descritto nell'approfondimento tecnico. Il tempo di controllo dipende dalla lunghezza del documento e dalle anomalie: il valore medio non descrive ogni DDT e non sostituisce una misurazione continuativa del lavoro di magazzino.

02

IL PROBLEMA: IL MATERIALE ARRIVA, MA VA ANCHE REGISTRATO NEL POSTO GIUSTO

In una torneria il materiale si muove continuamente. Arrivano barre e utensili; lotti di particolari escono per trattamenti termici, zincatura, rettifica o rivestimento e rientrano giorni dopo. Arrivano anche componenti finiti dai terzisti. Ogni movimento è accompagnato da un DDT, con articoli, quantità e riferimenti agli ordini o alle spedizioni precedenti.

L'azienda riceve circa 120 DDT a settimana, da decine di fornitori. Cambiano i formati: PDF leggibili, scansioni, fotografie di fogli spiegazzati scattate allo scarico.

Prima dell'intervento, l'operatore leggeva ciascun documento, apriva il gestionale, cercava l'ordine e la riga corretta, inseriva numero del DDT, data di consegna ed eventuale prezzo, aggiornava lo stato e allegava il PDF. Infine salvava una copia nella cartella condivisa con un nome riconoscibile.

Circa sei minuti per documento significavano dodici ore settimanali dedicate a riportare informazioni già presenti altrove. A questo si aggiungevano le verifiche quando qualcosa non tornava: un articolo sull'ordine sbagliato, una quantità interpretata male, una riga dimenticata nella seconda pagina, un documento registrato due volte.

Ripetere centinaia di ricerche e trascrizioni rende possibili errori anche quando chi lavora conosce bene il proprio mestiere. La richiesta del titolare riassumeva l'obiettivo: *«Voglio che le persone mantengano la decisione e smettano di ricopiare i dati.»*

Il rientro dal trattamento termico

Uno dei documenti più impegnativi arrivava una volta a settimana dal trattamentista: tre o quattro pagine, dodici o quindici lotti e diversi riferimenti ai DDT con cui erano usciti. I riferimenti erano stampati fra i gruppi di righe, non tutti nell'intestazione.

Registrarlo richiedeva quasi mezz'ora. Ogni lotto andava ricondotto alla spedizione in uscita e poi alla riga d'ordine. Prima del progetto, una registrazione sull'ordine sbagliato aveva lasciato un lotto fermo al controllo in accettazione per una settimana, mentre un altro ordine risultava arrivato nel gestionale e veniva sollecitato dal cliente.

Il documento era lungo, i codici simili e mancava un controllo che rendesse evidente l'incongruenza prima della registrazione. Era uno dei casi su cui valutare l'utilità dell'intervento.

03

IL LAVORO DI INDALOOP: CAPIRE IL PROCESSO, PROVARE, INTEGRARE

Il problema comprendeva più della lettura del PDF. Per essere utile, il sistema doveva ricondurre i dati al giusto ordine, gestire riferimenti distribuiti su più pagine e rendere visibili le eccezioni a chi conosceva il materiale ricevuto.

Abbiamo separato questi compiti: al modello AI la lettura del documento; al software le regole di confronto e la preparazione degli aggiornamenti; all'operatore la verifica e l'approvazione. Questa separazione ha permesso di cambiare il componente di lettura senza riscrivere le regole di abbinamento al gestionale.

In una fase iniziale abbiamo pensato di utilizzare la combinazione OCR + LLM open source sulla workstation disponibile. I test effettuati hanno dimostrato limiti di quella configurazione.

Nel confronto successivo abbiamo scelto la lettura diretta dei PDF tramite Claude. Sul campione della torneria offriva una lettura più completa e tempi più brevi, senza richiedere nuovo hardware. La scelta riguardava quel processo, quei documenti e quei volumi; altre condizioni possono giustificare una soluzione locale.

04

COME SI LAVORA OGGI

L'operatore trascina uno o più PDF nella pagina dell'applicazione. Ogni documento ha il proprio spazio di lavoro. Il sistema estrae intestazione, righe articolo e riferimenti agli ordini o alle spedizioni in uscita.

Il documento originale resta visibile accanto ai dati estratti. L'operatore può correggere la lettura; i dubbi individuati dal sistema vengono segnalati. Il software confronta poi i dati verificati con il gestionale e mostra, riga per riga, l'abbinamento proposto, i campi da aggiornare e il motivo della proposta.

Il gestionale cambia soltanto dopo l'approvazione. Il PDF viene quindi allegato all'ordine e archiviato con un nome standard.

Lo stesso percorso gestisce i due flussi oggetto del progetto:

- **Acquisti:** barre, inserti e componenti vengono collegati all'ordine indicato nel DDT.

- **Rientri da conto lavoro:** i particolari tornati da trattamento, rettifica o rivestimento vengono ricondotti al DDT di uscita e da lì all'ordine corrispondente.

La causale di trasporto aiuta a distinguere i flussi. L'operatore trova le proposte nello stesso ambiente e può concentrarsi sulle differenze da chiarire.

Lo stesso rientro, con il nuovo processo

Il documento del trattamentista richiede ora circa tre minuti di lavoro dell'operatore. Dopo il caricamento, in una quarantina di secondi il sistema presenta quattordici lotti con i rispettivi riferimenti e i campi da aggiornare.

In un caso, tredici righe erano pronte e una richiedeva attenzione: erano rientrati 480 pezzi sui 500 inviati. L'operatore ha approvato le tredici righe, chiarito con il trattamentista lo scarto di venti pezzi e gestito la differenza. La riga anomala non è stata aggiornata automaticamente come se la quantità fosse corretta.

> **«Prima erano i documenti lunghi a preoccuparmi. Ora sono quelli su cui risparmio più tempo. Non ricopio più i dati: controllo quello che il sistema ha trovato e dico sì o no.»**

Il cambiamento è visibile nel lavoro: meno passaggi fra foglio e gestionale, meno trascrizioni, più attenzione alle consegne che richiedono una decisione.

05

I CONTROLLI CHE RENDONO IL SISTEMA UTILIZZABILE

Un documento letto bene può comunque essere associato all'ordine sbagliato. Per questo abbiamo progettato i controlli anche dopo l'estrazione dei dati.

- **La proposta approvata è quella applicata.** Il sistema usa i dati corretti dall'operatore e applica il piano mostrato, senza ricalcolare abbinamenti dopo il sì.
- **I valori esistenti vengono riletti prima della scrittura.** I controlli sui campi già compilati proteggono dalle registrazioni ripetute. I numeri di DDT presenti non vengono sovrascritti; le sostituzioni ammesse di date e prezzi sono evidenziate nel piano.
- **Le quantità diverse richiedono una decisione.** Se il documento indica 480 pezzi e l'ordine 500, la riga resta esclusa dall'aggiornamento proposto e mostra entrambi i valori.
- **Il controllo qualità mantiene le proprie regole.** Nel processo configurato, soltanto i fornitori approvati possono portare una riga allo stato di rilascio previsto; gli altri passano dal controllo in accettazione anche quando l'abbinamento è riuscito.
- **Un gestionale irraggiungibile blocca il confronto.** L'errore di collegamento viene distinto da un articolo o da un ordine che non è stato trovato.

Queste protezioni riducono le occasioni di errore e rendono verificabili le modifiche.

06

COSTI DELLE SCELTE PROGETTUALI

Il consumo AI indicato per il progetto è di circa **€0,15 per documento**. Con circa 120 documenti a settimana, l'ordine di grandezza è di €80 al mese. Il costo effettivo varia con pagine, contenuto e numero di chiamate necessarie.

Questa cifra riguarda il servizio AI. Il costo del progetto comprende anche analisi, sviluppo, integrazione, prove e assistenza. Queste attività esistono anche usando un modello locale: per confrontare le alternative occorre separare il lavoro comune dai costi aggiuntivi di ciascuna strada.

Nel caso della torneria non è stato acquistato hardware. Proseguire con la soluzione locale avrebbe richiesto di valutare più memoria e ulteriore lavoro di messa a punto, senza una qualità finale già dimostrata. Le stime considerate descrivono un'alternativa valutata, non spese sostenute.

07

QUANDO UN INTERVENTO SIMILE PUÒ ESSERE UTILE

Il caso riguarda una torneria, ma il processo è comune a molte PMI manifatturiere. Vale la pena esaminarlo quando gli stessi documenti arrivano ogni settimana da mittenti diversi, qualcuno ne ricopia il contenuto e un errore di registrazione ha conseguenze sulle consegne, sugli ordini o sulle verifiche amministrative.

Da questo lavoro abbiamo ricavato alcuni criteri pratici:

1. **Partire dal lavoro ripetuto.** Individuare quali informazioni vengono copiate ogni giorno, da chi e con quali conseguenze quando qualcosa non torna.
2. **Separare lettura, regole e approvazione.** Rendere controllabili gli abbinamenti e le scritture permette anche di sostituire il modello senza rifare tutto il processo.
3. **Misurare sui documenti difficili.** I risultati sui fogli più semplici non bastano a descrivere il lavoro reale.
4. **Mostrare le modifiche prima di applicarle.** Chi conosce il processo deve poter capire la proposta, correggerla e fermarla.
5. **Fermarsi dove manca l'evidenza.** Un'anomalia da chiarire è preferibile a una registrazione apparentemente corretta su un ordine sbagliato.
6. **Confrontare il costo complessivo.** Consumo AI, hardware, integrazione e manutenzione vanno valutati rispetto ai volumi e ai requisiti dell'azienda.

Conferme d'ordine dei fornitori, fatture da confrontare con gli ordini, certificati dei materiali e ordini clienti ricevuti via e-mail possono presentare problemi simili. Ognuno richiede le proprie regole e una verifica su documenti rappresentativi.

APPENDICE

APPROFONDIMENTO TECNICO: COME ABBIAMO VALUTATO LA LETTURA DEI DOCUMENTI

Questa parte raccoglie configurazioni, risultati del confronto e vincoli tecnici. Le osservazioni riguardano il campione e l'hardware del progetto; non costituiscono una classifica generale fra modelli locali e servizi esterni.

La prima ipotesi: OCR e LLM open source sulla workstation esistente

La configurazione iniziale eseguiva due componenti: un modello OCR open source per trasformare la scansione in testo e un modello linguistico open source da 8 miliardi di parametri per estrarne i dati. Condividevano una scheda grafica da 16 GB, già presente nella workstation usata per le simulazioni CAM.

Il vantaggio atteso era usare l'hardware disponibile e mantenere i documenti in sede, senza una tariffa AI per documento. Restavano comunque consumo elettrico, gestione dei servizi e tempo di sviluppo.

I primi dieci DDT, puliti e di una sola pagina, davano risultati incoraggianti. I test effettuati con i successivi 100 DDT hanno fatto emergere diversi problemi. La sintesi emersa dal lavoro era concreta: stavamo dedicando più tempo a correggere il comportamento del lettore che a rendere utilizzabile il processo.

Il vincolo di memoria e la finestra di contesto

La finestra di contesto è lo spazio disponibile al modello per elaborare istruzioni, testo e risposta. Si misura in token, unità di testo che non corrispondono a un numero fisso di parole. Nella configurazione provata, il bilancio doveva comprendere anche lo spazio necessario per generare la risposta strutturata.

Con OCR e modello linguistico sulla stessa scheda da 16 GB, e la memoria del contesto alla precisione adottata nel test, la finestra utilizzabile era di circa 8.000 token. Il limite effettivo dipendeva da modello, memoria disponibile e configurazione del servizio, non soltanto dal massimo dichiarato dal modello.

BILANCIO INDICATIVO DI UNA CHIAMATA NELLA CONFIGURAZIONE PROVATA	TOKEN
Istruzioni e schema dei campi	~2.000
Testo OCR di una pagina densa, con struttura della tabella	2.500–3.500
Risposta strutturata per una tabella di circa trenta righe	~3.000
Una pagina, ingresso e risposta	**7.500–8.500**
Due pagine, con la stessa riserva indicativa per la risposta	**10.000–12.000**
Finestra utilizzabile sulla scheda condivisa	**~8.000**

La stima su due pagine non tiene conto di un eventuale aumento delle righe da restituire, che può richiedere ancora più spazio per la risposta.

1. I documenti lunghi superavano lo spazio disponibile

Già una pagina densa poteva avvicinarsi al limite. Due pagine lo superavano nei casi considerati. Fra questi c'erano proprio i rientri settimanali da trattamento e i DDT che chiudevano più ordini: documenti sui quali l'aiuto sarebbe stato più utile.

2. La divisione per pagine interrompeva i riferimenti

Inviare una pagina alla volta, ripetendo l'intestazione, non risolveva i casi in cui il riferimento alla spedizione era stampato fra gruppi di articoli. Una riga a pagina tre poteva dipendere da un riferimento a metà pagina due. La strategia provata perdeva questa relazione. Gestirla avrebbe richiesto di ricostruire e trasmettere lo stato fra le pagine, con ulteriori controlli.

3. Alcune righe centrali venivano omesse o confuse

Nel modello locale provato, intestazione e righe iniziali erano più regolari delle parti centrali di tabelle lunghe, dove comparivano omissioni, fusioni o ripetizioni. Un risultato formalmente leggibile poteva quindi essere incompleto. Questo rischio richiedeva verifiche sul contenuto, oltre alla validità della struttura restituita.

4. Alcune risposte venivano troncate

Su tabelle lunghe si osservavano ripetizioni fino all'esaurimento dello spazio disponibile per la risposta. Una risposta strutturata troncata veniva trattata come errore di estrazione; il documento originale restava disponibile per una nuova prova o per l'inserimento manuale. I tentativi aggiungevano tempo al processo.

5. Comparivano valori non presenti o interpretazioni instabili

Fra gli errori osservati c'erano una data assente dal documento e campi completati con valori plausibili. Lo stesso numero poteva essere restituito con separatori diversi in chiamate successive. Il modello usava una compressione a quattro bit per rientrare in memoria, ma il suo eventuale contributo a questi errori non è stato isolato nel confronto e non gli attribuiamo una causa dimostrata.

6. Le correzioni aumentavano il lavoro da mantenere

Durante la verifica sono state esplorate istruzioni più brevi, uno schema più piccolo, chiamate per pagina, recupero delle risposte troncate e nuovi tentativi con impostazioni diverse. Spostare l'OCR sul processore liberava memoria video, ma portava l'elaborazione da circa venti secondi a diversi minuti per pagina. Comprimere la memoria del contesto permetteva di arrivare intorno a 16.000 token: più spazio, senza risolvere tutte le omissioni. Un modello da 12–14 miliardi di parametri non rientrava insieme all'OCR nella configurazione provata.

La scheda condivisa imponeva inoltre un documento alla volta. Il servizio OCR, eseguito in un ambiente Linux su Windows, richiedeva attenzione al riavvio: nella configurazione iniziale non ripartiva automaticamente dopo un aggiornamento.

Modelli più grandi, nell'intervallo di 27–70 miliardi di parametri, e finestre di 32.000–128.000 token erano alternative da valutare con hardware adeguato. Avrebbero potuto alleviare i vincoli di contesto, ma la qualità sui DDT avrebbe comunque richiesto una nuova prova. Non erano risultati già ottenuti.

Il confronto: sostituire il lettore mantenendo le regole

indaloop aveva separato il componente di lettura dalle regole di abbinamento e scrittura. Il confronto ha quindi usato gli stessi documenti con tre configurazioni.

CONFIGURAZIONE	CHE COSA PERMETTE DI CONFRONTARE
OCR locale + LLM locale da 8B	La configurazione iniziale
OCR locale + Claude, con le stesse istruzioni e gli stessi schemi	Il cambio di modello a parità di testo estratto
Claude con PDF diretto	L'effetto complessivo del percorso di lettura diretto, rispetto al testo prodotto dall'OCR separato

Il passaggio da testo OCR a PDF modifica anche le informazioni disponibili al modello, inclusa la disposizione visiva della pagina. Il risultato misura il percorso completo: non attribuisce ogni differenza a un singolo componente.

Il campione comprendeva **100 DDT reali, 31 fornitori e circa 650 righe articolo**. Sedici documenti erano rientri da conto lavoro su più pagine, con una parte rilevante delle righe. I casi difficili erano deliberatamente sovrarappresentati rispetto al flusso ordinario, per verificare dove le soluzioni incontravano i propri limiti.

CONFRONTO INTERNO SVOLTO DURANTE IL PROGETTO	OCR + LLM LOCALE 8B	OCR + CLAUDE	PDF DIRETTO
Documenti non completati per limiti di contesto	16 su 100	0 su 100	0 su 100
Righe giudicate corrette nel confronto interno	71%, **solo sugli 84 documenti completati**	93%, sul campione completo	98%, sul campione completo
Quota indicativa di documenti senza correzioni di lettura	Circa 2 su 10	Circa 7 su 10	Circa 9 su 10
Campi restituiti senza riscontro nel documento	4	0 osservati	0 osservati
Tempo medio dal caricamento alla proposta, escluso il controllo umano	4 min 10 s, sui documenti completati	1 min 30 s	35 s; circa 90 s nel caso più lento, un rientro di sei pagine

Il 71% e il 98% hanno basi diverse: il primo esclude i sedici documenti che la configurazione locale non ha completato. Non rappresentano quindi un miglioramento misurato sulle stesse righe. Le percentuali e le quote sono valori arrotondati del confronto interno, non una garanzia di accuratezza su altri documenti. Per replicare la misura servono lo stesso insieme di righe e criteri espliciti sui campi da considerare corretti. I mancati completamenti vanno sempre riportati separatamente.

Nel percorso diretto tutti i documenti del campione rientravano nella finestra disponibile. La lettura del PDF, attraverso la pagina e l'eventuale livello testuale, evitava di dipendere esclusivamente dal testo dell'OCR locale, che poteva fondere colonne, spezzare codici o alterare separatori. Le immagini restavano comunque soggette a errori di lettura, come mostrano i casi descritti nel racconto.

La sostituzione ha riguardato il lettore. Le regole di abbinamento e scrittura sono rimaste le stesse.

ASPETTO OPERATIVO	CONFIGURAZIONE LOCALE PROVATA	LETTURA PDF TRAMITE SERVIZIO ESTERNO
Componenti AI da gestire in sede	OCR, ambiente Linux su Windows, server del modello e GPU	Integrazione con le API; applicazione e collegamento al gestionale restano da gestire
Documenti lunghi	Limiti già su una pagina densa; sedici documenti non completati nel test	Tutti i 100 documenti del campione completati, compreso un rientro di sei pagine
Passaggio OCR locale	Circa 20 s per pagina	Assente come servizio separato
Parallelismo	Un documento alla volta sulla scheda condivisa	Possibile entro i limiti del servizio e dell'applicazione
Riavvio dei servizi AI locali	Intervento manuale nella configurazione iniziale	Nessun servizio AI locale da riavviare
Valori senza riscontro osservati nel test	4 campi sui documenti completati	Nessuno nei 100 documenti; il campione non dimostra che il rischio sia nullo
Costo del consumo AI	Nessuna tariffa per chiamata; restano energia e gestione	Circa €0,15 per documento nel progetto

Il modello legge, il software applica le regole, la persona approva

Le prove iniziali hanno rafforzato una scelta di progetto: limitare le decisioni affidate al modello e rendere verificabile ciò che avviene dopo la lettura. L'AI trascrive il contenuto. Il software interpreta la struttura restituita, controlla quantità e riferimenti e prepara gli abbinamenti.

GESTITO DALLE REGOLE DEL SOFTWARE SUI DATI ESTRATTI	AFFIDATO AL MODELLO AI
Individuazione e interpretazione della tabella nella struttura restituita	Lettura del documento e riconoscimento di numero DDT, data, mittente e causale
Mappatura delle colonne quando riconoscibili	Supporto alla mappatura delle celle quando le regole non bastano
Interpretazione dei separatori numerici, con segnalazione delle ambiguità	Riconoscimento del fornitore fra candidati già individuati dal software
Relazione fra righe di riferimento e righe articolo	
Distinzione fra articoli e intestazioni di sezione	
Confronto delle quantità con l'ordine	

Lo schema di risposta ammette un campo assente quando l'informazione non è sul documento. Una struttura non valida o troncata viene trattata come errore. Questo aiuta a controllare la forma del risultato, ma non garantisce che ogni valore restituito sia corretto.

Oltre alla lettura, le richieste al modello sono circoscritte a opzioni trovate dal software, come una lista di fornitori candidati. Il sistema verifica che la risposta appartenga alle opzioni ammesse; resta possibile scegliere un candidato sbagliato, motivo per cui abbinamento e proposta vengono mostrati all'operatore.

Un valore numerico ambiguo viene lasciato da chiarire. La registrazione procede solo quando il piano è verificabile e viene approvato: è questa relazione fra lettura, regole e controllo umano che rende il modello sostituibile all'interno del processo.

> *Nota sul caso: il progetto riguarda un'azienda reale, che ha chiesto di non essere identificata. Il lavoro si è svolto in due settimane complessive; la verifica iniziale di OCR e LLM open source ha occupato i primi tre giorni. I valori quantitativi sono arrotondati e i risultati del confronto si riferiscono al campione descritto, deliberatamente concentrato sui documenti più impegnativi. Tempi, qualità e costi possono variare con documenti, volumi e configurazione. indaloop è disponibile a illustrare il percorso di estrazione, i controlli e il confronto tecnico.*

indaloop

AI × PROCESSI × AZIENDE

AI che legge, regole che controllano, persone che decidono.

Altri casi e strumenti gratuiti su indaloop.ai