# Quale LLM scegliere per l’OSINT? Il modello migliore dipende dal lavoro che deve fare

Maria Cattini | 15/09/2026 | Intelligenza Artificiale

Scegliere un modello linguistico per un’attività OSINT partendo dalla domanda «qual è il migliore?» porta quasi sempre nella direzione sbagliata. Un analista può dover esaminare migliaia di righe di conversazioni, estrarre entità, leggere uno screenshot, classificare fotografie, trascrivere una registrazione o lavorare senza connessione Internet. Sono problemi diversi e richiedono caratteristiche diverse.

La scelta, quindi, dovrebbe partire dal materiale e dal compito, non dalla classifica del momento.

Il documento preso come base per questa analisi propone diversi modelli testati in scenari specifici e individua alcuni criteri ricorrenti: ampiezza della finestra di contesto, multimodalità, capacità di produrre dati strutturati, tendenza alle allucinazioni e risorse hardware necessarie. I risultati numerici citati nel materiale derivano da prove dell'autore del documento e vanno letti come esperienze di test, non come benchmark indipendenti o universalmente riproducibili.

## Prima del modello viene il compito

Per l’OSINT una finestra di contesto ampia può essere importante quando bisogna lavorare con dossier, log o conversazioni molto lunghe. Nel materiale di partenza viene indicata come soglia minima 32.000 token, con una preferenza per 128.000 nei lavori più impegnativi.

La quantità di testo gestibile, però, non basta. Se l'obiettivo è automatizzare una parte del flusso di lavoro, conta anche la capacità di restituire informazioni in una struttura prevedibile, per esempio JSON o tabelle. Se invece le fonti comprendono screenshot, fotografie o filmati, serve un modello multimodale.

C'è poi un requisito ancora più importante: l'affidabilità.

Un LLM può estrarre e organizzare informazioni molto rapidamente, ma un risultato ben scritto non è automaticamente un risultato vero. Nell'OSINT questo problema è particolarmente serio perché un'entità inventata, un collegamento inesistente o una deduzione trasformata in fatto possono contaminare le fasi successive dell'indagine.

Per questo il tasso di allucinazioni non dovrebbe essere considerato una semplice voce in una scheda tecnica: gli output importanti devono essere verificati sulle fonti originali.

## Grandi quantità di testo: Llama 3.1 70B e Qwen2.5 72B

Quando il materiale è prevalentemente testuale e molto esteso, il documento segnala due modelli di grandi dimensioni.

**Llama 3.1 70B Instruct** viene indicato per attività come l'estrazione di nomi, numeri di telefono e altre entità da grandi conversazioni. Nel test descritto nel materiale, condotto su 10.000 righe di chat, avrebbe individuato il 99% dei numeri telefonici e il 97% dei nomi univoci senza creare contatti

inesistenti. Il supporto linguistico al di fuori dell'inglese viene invece indicato come un limite.

**Qwen2.5 72B Instruct** viene preferito nel documento quando le fonti sono multilingue. Nelle prove riportate su dossier contenenti russo, ucraino e inglese, avrebbe raggiunto il 96% nell'estrazione dei fatti, con circa il 2% di allucinazioni. Viene inoltre apprezzata la capacità di produrre JSON pulito, utile quando l'output deve passare successivamente a parser o altri strumenti.

Sono modelli potenti, ma il costo computazionale conta. Il materiale indica configurazioni basate su due RTX 3090/4090 con quantizzazione Q4_K_M oppure, per Llama, un Mac Studio M2 Ultra con 192 GB di memoria unificata. Non sono quindi necessariamente la scelta più razionale per qualsiasi attività OSINT.

## Screenshot e fotografie richiedono un altro tipo di modello

Un dossier investigativo raramente contiene soltanto testo. Può comprendere schermate di chat, documenti fotografati, immagini provenienti dai social o fotogrammi estratti da un filmato.

Per questo tipo di materiale viene proposto **Llama 3.2 11B Vision Instruct**, modello decisamente più piccolo dei precedenti. Nel test riportato avrebbe riconosciuto correttamente il 94% di indirizzi, numeri di carta e password presenti in screenshot di applicazioni di messaggistica. Il documento lo indica come utilizzabile anche con GPU da 8-12 GB di VRAM in quantizzazione Q4_K_M.

Qui serve però una cautela metodologica importante. Il fatto che un modello riesca a leggere un'informazione sensibile non significa che sia opportuno fornirgliela. Prima di utilizzare servizi cloud per analizzare database compromessi, credenziali, conversazioni private o altri dati personali occorre valutare provenienza del materiale, base giuridica del trattamento, condizioni del servizio e conseguenze dell'eventuale trasferimento dei dati. Quando possibile, l'elaborazione locale può ridurre una parte di questi rischi, ma non rende automaticamente lecito qualsiasi trattamento.

## Video: Qwen2-VL e MiniCPM-V per compiti differenti

Per l'analisi di materiale video, il documento indica **Qwen2-VL 7B Instruct**, attribuendogli la capacità di individuare fotogrammi significativi, descrivere sequenze temporali e riconoscere elementi presenti nelle immagini. In un test su dieci minuti di una manifestazione avrebbe estratto gli slogan con una precisione del 95%. Viene però segnalato il rischio di perdere dettagli piccoli nei video più lunghi.

Questo limite suggerisce una regola utile anche oltre il singolo modello: non chiedere necessariamente all'AI di «capire tutto il video» in una sola operazione.

Un flusso più controllabile può consistere nel dividere il materiale, individuare segmenti e fotogrammi rilevanti, estrarre gli elementi di interesse e tornare poi alle immagini originali per la verifica.

**MiniCPM-V 2.6**, invece, viene proposto per operazioni meno complesse e più ripetitive, come classificare grandi quantità di fotografie in base alla presenza di determinati oggetti o documenti. Richiede meno risorse, ma non viene presentato come modello adatto a deduzioni articolate.

È una distinzione importante: classificare non significa interpretare. E interpretare non significa dimostrare.

## Audio: prima trascrivere, poi analizzare

Per le registrazioni audio, il materiale suggerisce un'architettura a più passaggi.

**Whisper large-v3** viene utilizzato per trasformare il parlato in testo. La trascrizione può essere successivamente affidata a un LLM per individuare nomi, luoghi, riferimenti temporali o altri elementi utili. Nei test descritti, su registrazioni pulite in russo il sistema avrebbe prodotto meno del 5% di errori per parola.

Viene citato anche **Qwen-Audio-Chat 7B** per analizzare elementi come intonazione, emozioni e suoni ambientali, con una limitazione a frammenti di 30 secondi secondo il materiale fornito. L'analisi delle emozioni richiede comunque particolare prudenza: un'etichetta prodotta da un modello non dimostra lo stato emotivo o l'intenzione di una persona.

Per un'indagine OSINT è più utile trattarla come possibile segnale da contestualizzare che come fatto accertato.

## E se bisogna lavorare sul campo?

Non sempre l'analista dispone di una workstation con decine di gigabyte di memoria video.

Nel documento viene raccontato anche l'utilizzo di **Phi-3.5-vision-instruct**, modello da 4,2 miliardi di parametri, su un GPD MicroPC 2 con Intel i3-N300 e 16 GB di RAM. Attraverso OpenVINO e la grafica integrata Intel Iris viene descritto come sufficientemente efficace per attività circoscritte, per esempio leggere rapidamente un documento o controllare una fotografia. Non è invece indicato per ragionamenti complessi in più passaggi.

È forse l'esempio che chiarisce meglio il criterio di scelta: il modello più grande non è necessariamente quello più utile. Se bisogna effettuare una prima selezione sul campo, un sistema piccolo e locale può essere più pratico di un modello molto più potente ma difficile da eseguire.

## Un metodo pratico per scegliere

Prima di scaricare o configurare un modello, conviene descrivere l'attività con precisione.

Se abbiamo centinaia di migliaia di righe di testo, servono soprattutto contesto, estrazione affidabile e output strutturato. Con documenti in lingue differenti acquista peso la capacità multilingue. Per screenshot e fotografie diventa indispensabile la componente visiva. Per i video occorre valutare non soltanto se il modello dichiara di supportarli, ma quanto materiale riesce effettivamente ad analizzare senza perdere dettagli. Per l'audio può essere più efficace separare trascrizione e analisi. Se i dati non devono lasciare la macchina, infine, hardware e possibilità di esecuzione locale diventano requisiti centrali.

Solo dopo aver definito questi vincoli ha senso confrontare i modelli.

E prima di impiegarne uno su un'indagine reale è opportuno costruire un piccolo dataset di controllo del quale conosciamo già le risposte. Possiamo così misurare quanti elementi vengono trovati, quanti vengono persi e, soprattutto, quanti vengono inventati.

Quest'ultimo dato è spesso più importante della capacità del modello di produrre una risposta convincente.

## L'LLM non è la fonte

C'è infine un confine che nell'OSINT non dovrebbe essere confuso.

Un modello linguistico può essere molto efficace per ridurre il rumore: classificare migliaia di elementi, estrarre entità, organizzare informazioni, individuare ricorrenze o suggerire dove guardare. Può quindi diventare un potente strumento di triage.

La verifica, però, deve tornare alla fonte.

Se il modello individua una targa in un fotogramma, bisogna controllare il fotogramma. Se estrae un nome da una conversazione, bisogna ritrovare quel nome nella conversazione originale. Se propone una relazione tra due entità, bisogna capire quali elementi documentali sostengano davvero quella relazione.

La domanda utile, allora, non è semplicemente quale LLM scegliere per l'OSINT. È quale parte del lavoro possiamo affidargli senza confondere velocità di elaborazione e qualità della prova. Scegliere un modello linguistico per un’attività OSINT partendo dalla domanda «qual è il migliore?» porta quasi sempre nella direzione sbagliata. Un analista può dover esaminare migliaia di righe di conversazioni, estrarre entità, leggere uno screenshot, classificare fotografie, trascrivere una registrazione o lavorare senza connessione Internet. Sono problemi diversi e richiedono caratteristiche diverse.

La scelta, quindi, dovrebbe partire dal materiale e dal compito, non dalla classifica del momento.

Il documento preso come base per questa analisi propone diversi modelli testati in scenari specifici e individua alcuni criteri ricorrenti: ampiezza della finestra di contesto, multimodalità, capacità di produrre dati strutturati, tendenza alle allucinazioni e risorse hardware necessarie. I risultati numerici citati nel materiale derivano da prove dell'autore del documento e vanno letti come esperienze di test, non come benchmark indipendenti o universalmente riproducibili.

## Prima del modello viene il compito

Per l’OSINT una finestra di contesto ampia può essere importante quando bisogna lavorare con dossier, log o conversazioni molto lunghe. Nel materiale di partenza viene indicata come soglia minima 32.000 token, con una preferenza per 128.000 nei lavori più impegnativi.

La quantità di testo gestibile, però, non basta. Se l'obiettivo è automatizzare una parte del flusso di lavoro, conta anche la capacità di restituire informazioni in una struttura prevedibile, per esempio JSON o tabelle. Se invece le fonti comprendono screenshot, fotografie o filmati, serve un modello multimodale.

C'è poi un requisito ancora più importante: l'affidabilità.

Un LLM può estrarre e organizzare informazioni molto rapidamente, ma un risultato ben scritto non è automaticamente un risultato vero. Nell'OSINT questo problema è particolarmente serio perché un'entità inventata, un collegamento inesistente o una deduzione trasformata in fatto possono contaminare le fasi successive dell'indagine.

Per questo il tasso di allucinazioni non dovrebbe essere considerato una semplice voce in una scheda tecnica: gli output importanti devono essere verificati sulle fonti originali.

## Grandi quantità di testo: Llama 3.1 70B e Qwen2.5 72B

Quando il materiale è prevalentemente testuale e molto esteso, il documento segnala due modelli di grandi dimensioni.

**Llama 3.1 70B Instruct** viene indicato per attività come l'estrazione di nomi, numeri di telefono e altre entità da grandi conversazioni. Nel test descritto nel materiale, condotto su 10.000 righe di chat, avrebbe individuato il 99% dei numeri telefonici e il 97% dei nomi univoci senza creare contatti inesistenti. Il supporto linguistico al di fuori dell'inglese viene invece indicato come un limite.

**Qwen2.5 72B Instruct** viene preferito nel documento quando le fonti sono multilingue. Nelle prove riportate su dossier contenenti russo, ucraino e inglese, avrebbe raggiunto il 96% nell'estrazione dei fatti, con circa il 2% di allucinazioni. Viene inoltre apprezzata la capacità di produrre JSON pulito, utile quando l'output deve passare successivamente a parser o altri strumenti.

Sono modelli potenti, ma il costo computazionale conta. Il materiale indica configurazioni basate su due RTX 3090/4090 con quantizzazione Q4_K_M oppure, per Llama, un Mac Studio M2 Ultra con 192 GB di memoria unificata. Non sono quindi necessariamente la scelta più razionale per qualsiasi attività OSINT.

## Screenshot e fotografie richiedono un altro tipo di modello

Un dossier investigativo raramente contiene soltanto testo. Può comprendere schermate di chat, documenti fotografati, immagini provenienti dai social o fotogrammi estratti da un filmato.

Per questo tipo di materiale viene proposto **Llama 3.2 11B Vision Instruct**, modello decisamente più piccolo dei precedenti. Nel test riportato avrebbe riconosciuto correttamente il 94% di indirizzi, numeri di carta e password presenti in screenshot di applicazioni di messaggistica. Il documento lo indica come utilizzabile anche con GPU da 8-12 GB di VRAM in quantizzazione Q4_K_M.

Qui serve però una cautela metodologica importante. Il fatto che un modello riesca a leggere un'informazione sensibile non significa che sia opportuno fornirgliela. Prima di utilizzare servizi cloud per analizzare database compromessi, credenziali, conversazioni private o altri dati personali occorre valutare provenienza del materiale, base giuridica del trattamento, condizioni del servizio e conseguenze dell'eventuale trasferimento dei dati. Quando possibile, l'elaborazione locale può ridurre una parte di questi rischi, ma non rende automaticamente lecito qualsiasi trattamento.

## Video: Qwen2-VL e MiniCPM-V per compiti differenti

Per l'analisi di materiale video, il documento indica **Qwen2-VL 7B Instruct**, attribuendogli la capacità di individuare fotogrammi significativi, descrivere sequenze temporali e riconoscere elementi presenti nelle immagini. In un test su dieci minuti di una manifestazione avrebbe estratto gli slogan con una precisione del 95%. Viene però segnalato il rischio di perdere dettagli piccoli nei video più lunghi.

Questo limite suggerisce una regola utile anche oltre il singolo modello: non chiedere necessariamente all'AI di «capire tutto il video» in una sola operazione.

Un flusso più controllabile può consistere nel dividere il materiale, individuare segmenti e fotogrammi rilevanti, estrarre gli elementi di interesse e tornare poi alle immagini originali per la verifica.

**MiniCPM-V 2.6**, invece, viene proposto per operazioni meno complesse e più ripetitive, come classificare grandi quantità di fotografie in base alla presenza di determinati oggetti o documenti. Richiede meno risorse, ma non viene presentato come modello adatto a deduzioni articolate.

È una distinzione importante: classificare non significa interpretare. E interpretare non significa dimostrare.

## Audio: prima trascrivere, poi analizzare

Per le registrazioni audio, il materiale suggerisce un'architettura a più passaggi.

**Whisper large-v3** viene utilizzato per trasformare il parlato in testo. La trascrizione può essere successivamente affidata a un LLM per individuare nomi, luoghi, riferimenti temporali o altri elementi utili. Nei test descritti, su registrazioni pulite in russo il sistema avrebbe prodotto meno del 5% di errori per parola.

Viene citato anche **Qwen-Audio-Chat 7B** per analizzare elementi come intonazione, emozioni e suoni ambientali, con una limitazione a frammenti di 30 secondi secondo il materiale fornito. L'analisi delle emozioni richiede comunque particolare prudenza: un'etichetta prodotta da un modello non dimostra lo stato emotivo o l'intenzione di una persona.

Per un'indagine OSINT è più utile trattarla come possibile segnale da contestualizzare che come fatto accertato.

## E se bisogna lavorare sul campo?

Non sempre l'analista dispone di una workstation con decine di gigabyte di memoria video.

Nel documento viene raccontato anche l'utilizzo di **Phi-3.5-vision-instruct**, modello da 4,2 miliardi di parametri, su un GPD MicroPC 2 con Intel i3-N300 e 16 GB di RAM. Attraverso OpenVINO e la grafica integrata Intel Iris viene descritto come sufficientemente efficace per attività circoscritte, per esempio leggere rapidamente un documento o controllare una fotografia. Non è invece indicato per ragionamenti complessi in più passaggi.

È forse l'esempio che chiarisce meglio il criterio di scelta: il modello più grande non è necessariamente quello più utile. Se bisogna effettuare una prima selezione sul campo, un sistema piccolo e locale può essere più pratico di un modello molto più potente ma difficile da eseguire.

## Un metodo pratico per scegliere

Prima di scaricare o configurare un modello, conviene descrivere l'attività con precisione.

Se abbiamo centinaia di migliaia di righe di testo, servono soprattutto contesto, estrazione affidabile e output strutturato. Con documenti in lingue differenti acquista peso la capacità multilingue. Per screenshot e fotografie diventa indispensabile la componente visiva. Per i video occorre valutare non soltanto se il modello dichiara di supportarli, ma quanto materiale riesce effettivamente ad analizzare senza perdere dettagli. Per l'audio può essere più efficace separare trascrizione e analisi. Se i dati non devono lasciare la macchina, infine, hardware e possibilità di esecuzione locale diventano requisiti centrali.

Solo dopo aver definito questi vincoli ha senso confrontare i modelli.

E prima di impiegarne uno su un'indagine reale è opportuno costruire un piccolo dataset di controllo del quale conosciamo già le risposte. Possiamo così misurare quanti elementi vengono trovati, quanti vengono persi e, soprattutto, quanti vengono inventati.

Quest'ultimo dato è spesso più importante della capacità del modello di produrre una risposta convincente.

## L'LLM non è la fonte

C'è infine un confine che nell'OSINT non dovrebbe essere confuso.

Un modello linguistico può essere molto efficace per ridurre il rumore: classificare migliaia di elementi, estrarre entità, organizzare informazioni, individuare ricorrenze o suggerire dove guardare. Può quindi diventare un potente strumento di triage.

La verifica, però, deve tornare alla fonte.

Se il modello individua una targa in un fotogramma, bisogna controllare il fotogramma. Se estrae un nome da una conversazione, bisogna ritrovare quel nome nella conversazione originale. Se propone una relazione tra due entità, bisogna capire quali elementi documentali sostengano davvero quella relazione.

La domanda utile, allora, non è semplicemente quale LLM scegliere per l'OSINT. È quale parte del lavoro possiamo affidargli senza confondere velocità di elaborazione e qualità della prova.