ROBOTS.TXT E CRAWLER AI: L'ESEMPIO COMMENTATO E I TOKEN

Un robots.txt che tiene il sito fuori dall'addestramento dei modelli e dentro la ricerca e le risposte AI, e la tabella dei nomi dalle pagine dei proprietari, lette il 10 novembre 2026.

IL FILE D'ESEMPIO

```
# robots.txt d'esempio: fuori dall'addestramento dei modelli, dentro la ricerca e le risposte AI.
# Nomi verificati sulle pagine dei proprietari il 10 novembre 2026:
https://itria.io/robots-txt-crawler-ai
# robots.txt è una richiesta, non un blocco (RFC 9309). Le modifiche arrivano in circa 24 ore.

# Tutti gli altri, motori di ricerca compresi: si entra.
User-agent: *
Allow: /

# ADDESTRAMENTO DEI MODELLI: fuori.
# OpenAI. Resti nella ricerca di ChatGPT, che usa OAI-SearchBot.
User-agent: GPTBot
# Anthropic. Resti nella ricerca di Claude, che usa Claude-SearchBot.
User-agent: ClaudeBot
# Google, Gemini. Non tocca la Ricerca né le AI Overview.
User-agent: Google-Extended
# Apple. Applebot continua a scansionare per Siri, Spotlight e Safari.
User-agent: Applebot-Extended
# Common Crawl, l'archivio aperto che molti usano per addestrare.
User-agent: CCBot
# Mistral.
User-agent: MistralAI-Training
# Amazon: servizi e addestramento. La ricerca di Alexa usa Amzn-SearchBot.
User-agent: Amazonbot
Disallow: /

# RICERCA E RISPOSTE AI: dentro. Servono solo se sopra c'è un blocco generale,
# ma scriverle rende la scelta leggibile a chi apre il file.
User-agent: OAI-SearchBot
User-agent: ChatGPT-User
User-agent: Claude-SearchBot
User-agent: Claude-User
User-agent: PerplexityBot
User-agent: Perplexity-User
User-agent: Amzn-SearchBot
User-agent: DuckAssistBot
User-agent: MistralAI-Index
User-agent: MistralAI-User
Allow: /

Sitemap: https://www.tuosito.it/sitemap.xml
```

I TOKEN, DALLE PAGINE DEI PROPRIETARI

Token	Azienda	Scopo	Rispetta robots.txt
GPTBot	OpenAI	Addestramento	Sì
OAI-SearchBot	OpenAI	Ricerca	Sì
ChatGPT-User	OpenAI	Richiesta dell'utente	Non sempre: le regole «may not apply»
OAI-AdsBot	OpenAI	Verifica degli annunci	La pagina non lo dice
ClaudeBot	Anthropic	Addestramento	Sì
Claude-SearchBot	Anthropic	Ricerca	Sì
Claude-User	Anthropic	Richiesta dell'utente	Sì
Google-Extended	Google	Addestramento (Gemini) e grounding	Sì (token senza user agent proprio)

Googlebot	Google	Ricerca, con AI Overview e AI Mode	Sì
Google-CloudVertexBot	Google	Agenti Vertex AI chiesti dal proprietario del sito	Sì
Google-Agent	Google	Richiesta dell'utente (agenti)	In genere no
Google-GeminiNotebook	Google	Richiesta dell'utente (fonti date a Gemini Notebook)	In genere no
PerplexityBot	Perplexity	Ricerca	Sì
Perplexity-User	Perplexity	Richiesta dell'utente	In genere no
Applebot	Apple	Ricerca (Spotlight, Siri, Safari), anche per l'AI	Sì
Applebot-Extended	Apple	Addestramento (solo esclusione)	Sì
Amazonbot	Amazon	Misto: servizi Amazon e addestramento	Sì (robots.txt in cache fino a 30 giorni)
Amzn-SearchBot	Amazon	Ricerca (Alexa e ricerche Amazon)	Sì
Amzn-User	Amazon	Richiesta dell'utente	Non sempre
CCBot	Common Crawl	Archivio aperto del web, usato da terzi anche per addestrare	Sì
DuckAssistBot	DuckDuckGo	Ricerca (risposte AI in tempo reale)	Sì, dopo 72 ore
MistralAI-User	Mistral	Richiesta dell'utente	Sì, secondo la pagina
MistralAI-Index	Mistral	Ricerca	Sì
MistralAI-Training	Mistral	Addestramento	Sì
bingbot	Microsoft	Ricerca Bing e risposte di Copilot	Sì

RFC 9309, IETF: «These rules are not a form of access authorization.» Un robots.txt chiede, non chiude: i fetcher che lavorano su richiesta di un utente dichiarano di poterlo non rispettare.

Il foglio Excel e la guida completa, con le fonti:
itria.io/robots-txt-crawler-ai

Foglio libero, costruito da Itria sulle pagine pubbliche dei proprietari dei crawler, lette il 10 novembre 2026. I nomi e le regole cambiano: ricontrollate le pagine ufficiali prima di pubblicare il robots.txt.

Mattia Esposito lavora sul posizionamento delle aziende, con analisi di mercato per capire dove i concorrenti sono scoperti. Due righe su WhatsApp al 350 574 3578 o una mail a info@itria.io.