

AI Crawler & Technical Accessibility Checklist

The discoverability layer: can AI systems actually reach, parse and attribute your content — regardless of how good it is?

This checklist covers Technical Accessibility — the one Dimension of Citation Integrity™ that overlaps most with familiar technical-SEO work. A source can be authoritative and well-written and still be functionally invisible to automated systems. Work through this with (or hand it to) your developer.

Crawler access

- robots.txt does not block reputable AI crawlers from pages you want cited.
- If you maintain an llms.txt file, it accurately reflects what you want AI systems to access and how.
- You've checked for accidental blanket disallows left over from a staging environment.

Canonical, stable URLs

- Every page you care about has one canonical, non-session-based URL.
- URLs don't change on re-publish, redesign, or A/B test rotation without a proper 301 redirect.
- Pagination and tracking parameters don't fragment one page into many crawlable variants.

Renderable content

- Core text content is present in the initial HTML response, not injected client-side only via JavaScript.
- Content behind tabs, accordions or “read more” toggles is present in the DOM, not loaded on click only.
- Critical content isn't delivered only inside images, video, or embedded PDFs with no text layer.

Structured data & attribution

- Structured data (schema.org / JSON-LD) marks up author, organization, and publish/update dates where relevant.
- Article or Organization schema is present and matches what's visibly on the page (no mismatched metadata).
- Rate limits and bot-management rules don't inadvertently throttle or block legitimate AI crawlers.

Licensing & access barriers

- Paywalls or login walls don't sit in front of content you actually want cited.
- Terms-of-service or licensing text doesn't unintentionally prohibit the kind of automated access you want.

What to hand your developer

A condensed version of the above, phrased as engineering tickets rather than checklist items:

- Audit robots.txt for AI-crawler user agents (GPTBot, ClaudeBot, PerplexityBot, Google-Extended, etc.) and confirm none are unintentionally disallowed on pages meant to be public.
- Confirm server-side rendering (or a pre-render/SSR fallback) delivers full text content without requiring client-side JavaScript execution.
- Add or validate Article/Organization JSON-LD on key pages: author, datePublished, dateModified, publisher.
- Eliminate session IDs, tracking parameters, or A/B-test variants from canonical URLs; add canonical tags where duplicates are unavoidable.
- Review bot-management / WAF rules and rate limits for false positives against known AI crawler user agents.
- If an llms.txt file exists, confirm it stays in sync with robots.txt rather than drifting out of date.

Want this checked automatically? Run a free Citation Integrity™ Audit, or ask about API access for ongoing monitoring.

This is an educational self-assessment tool from CiteAbility™, not a formal Citation Integrity™ Score, Audit or a guarantee that any AI system will cite this content.

Note: this checklist describes the Technical Accessibility Dimension in general terms. It does not promise more than CiteAbility™'s product currently verifies for any specific check.