Control starts with Transparency: Cloudflare’s position on AI Crawlers and Bots

As a provider of Internet services, Cloudflare has a strong interest in assuring both that publishers’ rights are respected and that developers have access to innovative new features, including LLMs. Balancing the tensions that emerge requires either trust or, in its absence, transparency.

Transparency is preferable when operating at Internet scale. It provides accountability with, for example, tamper evident logs. A prominent example is Certificate Transparency for the certificate authority ecosystem. If bot identities were similarly transparent, publishers could have confidence in how their content will be treated, and knowingly offer access to train LLMs.

The Robots Exclusion Protocol (robots.txt) is a poor basis for building that confidence for many reasons that will doubtless be discussed at the Workshop. We instead take the following position:

> Crawlers and bots operating at Internet scale are different from humans. To operate freely on the Internet, humans need privacy. Bots, however, are meant to serve human needs. To ensure correct operation on the Internet, humans need transparency from bots.

Cloudflare believes that bot identification is the foundation on which to build trust between publishers and LLM trainers who wish to express purpose of and limitations for crawlers. For identification purposes, however, User-Agent identifiers, IP addresses and similar techniques are many combinations of fragile, spoofable, incorrect, and hard to scale.

Instead, we advocate for identification using well established cryptographic techniques. To this end, we are developing a system we refer to as “cryptographically verified bots” (CVB), based upon a proposed change to the TLS specification called “Request mTLS.” CVB-enabled crawlers use a new TLS feature to signal a willingness to self-identify. It is our position that since bots serve humans, legitimate bots can be most effective by making themselves known.

Our method for crawler self-identification is currently under consideration for adoption in the IETF TLS Working Group, and a strict improvement over alternative non-cryptographic identifiers under consideration. In brief, the proposal is to add a flag to the TLS handshake that allows the client to signal that it is configured with a certificate. The server can then ask only clients with certificates to authenticate themselves.

While this does not address all issues in the space – in particular, how client certificates will be issued and how trust in them will be managed – it serves as a necessary first step towards bot transparency, allowing service to both unauthenticated human traffic and authenticated bot traffic to the same resources.

See the Request mTLS Internet-Draft for details.

We believe that cryptographically verified bots are a crucial component of a healthy Internet. Strong identity provides a basis for building trust and making the behavioral commitments that are necessary to both respect the rights of publishers and promote innovative new tools for developers.