Categorizing and filtering LLM bot/crawler traffic in GA4

This guide explains how to effectively categorize and track traffic from Large Language Models (LLMs). The goal is to refine GA4's default categorizations to specify the referring LLM, as well as filter out traffic associated to crawler bots dispersed by LLMs to gather information.

How LLM Bot Traffic is Currently Categorized:
- Referral: If the bot includes the LLM's domain as a referrer.
- Direct: If the bot removes referrer information.
- (not set) / Unassigned: When GA4 cannot determine the source, common for poorly formed bot requests.

GA4's Bot Handling:
- GA4 automatically excludes traffic from known bots and spiders, such as established search engine crawlers, using Google's methods and industry lists like the one maintained by the IAB (updated monthly). However, it does not reliably filter out newer or malicious bots, or bots that do not clearly identify themselves. These can be miscategorized as referral or direct traffic.

How to ensure all traffic from LLMs gets categorized as such:
- Go to Admin > Channel Groups.
- Create a new channel (e.g., "AI Chatbots" or "LLM Traffic").
- Set the condition to filter traffic based on **Source**, using a regular expression for known bot/LLM domains, e.g.:
  chatgpt\.com|chat\.openai\.com|gemini\.google\.com|claude\.ai|perplexity\.ai|you\.com|copilot\.microsoft\.com|bing\.com\/chat

How to filter out traffic associated to LLM crawler bots:
- In GTM, implement a custom JS variable that returns *true* if **navigator.userAgent** matches that of known LLM crawlers (e.g., GPTBot, ClaudeBot, CCBot, PerplexityBot, Amazonbot, copilot-bing, etc).
- Apply this variable as an exception rule within the triggers of all GA4 tags.

How to create a report that shows session traffic by source:
- Navigate to Explore and create a Blank exploration.
- Under Dimensions, add Source and your new Channel Group. Drag these into the Rows section
- Under Metrics, add Sessions. Drag this into the Values section.