

Legal Discovery Tools for Large Language Model Larceny

v. 1.0, June 7, 2023

Submitted To

Chicago Media Party Hack-a-Thon, June 8-11, 2023

Submitted By

- [McKinley Park News](#)
- [Inside Publications](#), including:
 - Skyline* newspaper
 - News-Star* newspaper
 - Inside-Booster* newspaper
- [Enchúfate](#) magazine

This document comprises a prospectus and proposal for the hack-a-thon component of the Chicago Media Party event held on June 8-11, 2023, at the Polsky Exchange North facility near the University of Chicago's campus. It is submitted in advance and during the conference for prospective adoption and execution in the hack-a-thon, resulting in usable tools to assist local news publishers, including those listed above, who advocate for this hack-a-thon development and help with a problem of critical urgency.

Summary of Problem

Local news publishers like those above have seemingly had their content, products and services accessed and used without authorization in many Large Language Models (or "LLMs"): an essential technical component of contemporary artificial intelligence (or "AI") chat applications sponsored by enterprises that have launched competing products and services, realized revenue, and raised investment and other funding based on this stolen content and labor.

In order to effectively pursue legal action against Large Language Model developers, purveyors, vendors, customers, promoters, trainers and users who have accessed and used our content without authorization, local news publishers and our attorneys must discover as much information about this theft as possible, including:

- Which Large Language Model developers, vendors, purveyors, promoters and customers are accessing and using our content without authorization.
- How long this theft has been occurring.
- How much revenue, funding and/or labor value has been received by developers, purveyors, vendors, customers, trainers, promoters and users of Large Language Models that employ our content, products and services without authorization.

However, conducting direct discovery against Large Language Model theft is fraught with legal danger and disadvantage, including directly querying the chat interfaces that rely on stolen content to generate their answers. For example:

- Any access or use of an AI-powered service is likely compelled to conform to its terms of service, which likely limit legal options and redress for victims of larcenous Large Language Models.
- Any interaction with an AI interface may feed it information to be used toward the disadvantage and lack of remuneration for the user of the AI interface.

Until such a time when litigation and/or legislation mandates direct discovery against developers, purveyors, vendors, customers, promoters, trainers and users of Large Language Models that employ stolen content, local news publications and publishers like those above and our legal representatives would greatly benefit from a resource that provides the following **deliverables**:

1. Analysis of the terms of service of products and services employing larcenous Large Language Models via a method that does not engender any legal disadvantage or loss of rights for the victims who receive the results of this discovery. This should consider both the terms of use for public-facing services, as well as for technical and enterprise business uses, including use and access via application programming interfaces (APIs).

The hack-a-thon and Media Party will likely need to apply whatever qualified legal expertise and resources it has available toward informed development of this deliverable.

Deliverable: A report of the resulting analysis and legal options with a scope that considers all industry- and research-leading Large Language Model projects and services.

2. Establishment of a qualified technical method, if possible, to conduct discovery directly against Large Language Model-powered services without engendering legal disadvantage or loss of rights for the querant. This may also consider and include existing third-party tools that query LLMs for this purpose, but that still may need to have their legal provenance qualified.

Deliverable: A report on the resulting analysis, including whether any methods exist at all for legally safe direct discovery against this theft of content and expertise, as well as any scope and limitations on discovery.

3. Establishment of a database of developers, purveyors, vendors, customers, promoters, trainers and users of Large Language Models that include stolen content and expertise, and resulting metadata to be used toward serving legal notice and other purposes.

Deliverable: An online or offline database structure for tracking larcenous Large Language Model entity types and their details, instances of unauthorized content use, affected victims, litigation and legislation tracking, and related information. Data for such can come from various sources, including those encountered in Nos. 1 and 2, above.

Media Party's Role

The Media Party event itself offers opportunities for discovery to inform the results of this prospectus, since the event includes and spotlights individuals, organizations and entities developing, selling and using products and services based on Large Language Models that employ stolen content.

In addition to the opportunities offered at the conference for direct query, Media Party organizers and/or speakers may wish to format and present a poll to attendees asking about their use of Large Language Model-powered services, and if they qualify its provenance or not.

Indeed, Media Party organizers and their supporters may wish to consider adoption of an AI ethics policy that prohibits theft, as well as one that does not provide support, promotion and training for such. This hack-a-thon prospectus may also be useful as an instructional tool for attending larcenous Large Language Model developers, purveyors, vendors, customers and users who understand neither the immense damage their theft causes to local news nor the legal jeopardy they place themselves in by using stolen property toward their benefit and profit.

Contact

Justin Kerr
Publisher, *McKinley Park News*
justinkerr@mckinleypark.news
(312) 560-1115

Background

- “Proposed Legislation to Regulate AI Adds To Our Fight Against Theft of Local News”
McKinley Park News, June 5, 2023
<https://mckinleypark.news/info/letter-from-the-editor/4181-proposed-legislation-to-regulate-ai-adds-to-our-fight-against-theft-of-local-news>
- “Chicago Newspapers Join Fight Against AI’s Theft of Local News”
McKinley Park News, May 31, 2023
<https://mckinleypark.news/info/announcements/4140-chicago-newspapers-join-fight-against-ai-s-theft-of-local-news>
- “Why I’ll Quit Publishing the *McKinley Park News*”
McKinley Park News, May 24, 2023
<https://mckinleypark.news/info/letter-from-the-editor/4135-why-i-ll-quit-publishing-the-mckinley-park-news>
- “AI’s Impact on Local Journalism”
WGN Radio – Outside the Box, June 3, 2023, broadcast
<https://outsidetheloopradio.com/2023/06/03/otl-868-nuclear-power-in-illinois-ais-impact-on-local-journalism-born-in-chicago-documents-blues-history/>
- “Community Journalism and the Threat of AI”
The Picayune Sentinel (Eric Zorn), June 1, 2023
<https://ericzorn.substack.com/p/my-grandfather-was-fired-and-we-couldnt>
- “AI’s ‘going to destroy local news’”
Chicago Public Square, May 26, 2023
<https://www.chicagopublicsquare.com/>