ICPSR Policy on the Use of Large Language Models (LLMs) and other forms of AI

Approved: December 11, 2024

General statement: *LLMs may only be used to manage, process, or analyze data distributed by ICPSR if the LLM meets specific criteria regarding retention of user-supplied data and/or placement within a secure network.*

Under ICPSR's Bylaws, researchers are forbidden to distribute data or other materials we supply (apart from study-level metadata and related publications described below) to other members, organizations, or individuals without ICPSR's written permission. This includes openICPSR (self-published) data sets that were brought in under the ICPSR Terms of Use (check the study's home page for licensing information). This document explains how this policy affects the ability of researchers to use large language models (LLMs) to analyze data distributed by ICPSR. This policy does not apply to ICPSR metadata.

For purposes of this policy, LLMs are classified into three categories:

- **Type 1**: LLMs that retain user-provided data for any purpose, including training the LLM (e.g., GPT, Llama)

- **Type 2**: LLMs that are licensed by an institution and have conditions of use that do not permit the retention of user-provided data (e.g., the University of Michigan's Maizey)

- **Type 3**: Type 2 LLMs that are isolated within a secure network with no access to the Internet

LLM Type	ICPSR Data that May be Shared with the LLM, with permission from ICPSR	Reason
Type 1	None	Type 1 LLMs ingest the data and make use of it. This counts as redistributing the data to the company operating the LLM, so it is not permitted.
Type 2	Public-Use datasets	Type 2 LLMs do not retain the data or make use of it, so this does not count as redistribution. However, they are not isolated from broader networks or the Internet, so they would not comply with data security plans for Restricted-Use data.
Type 3	Public-Use datasets Restricted-Use datasets	Type 3 LLMs do not retain or make use of the data, and they are also isolated on individual machines or within secure networks, so they may comply with the requirements for Restricted-Use data.

Public-Use Datasets

Users are not allowed to redistribute ICPSR datasets without permission. If you give ICPSR data to a Type 1 LLM, it ingests the data and makes use of it. This counts as redistributing the data to the company operating the LLM, so it is not permitted.

With a Type 2 LLM, the contract between the owning institution and the company operating the LLM forbids retention of user-provided data – it processes the data but is not allowed to keep a copy. Type 3 LLMs have the same contractual requirements but are also isolated so they are even safer. ICPSR may permit you to use ICPSR data with a Type 2 or 3 LLM as long as the LLM is not used to make connections in the data that would increase the risk of identifying individuals or organizations represented in the data. Please contact ICPSR prior to use.

Restricted-Use Datasets via Secure Download

In addition to the prohibition on redistribution, access to restricted-use datasets via secure download requires that the user specify and adhere to a data security plan. While there are several different possible plans, all of them require the data set to be isolated from the internet to protect it from intrusion. Therefore, the only type of LLM that can be used with restricted data is Type 3, and then only if the use is consistent with the security plan and approved by ICPSR. For clarification, or to discuss possible exceptions in unusual circumstances, contact ICPSR.

Restricted-Use Datasets via the ICPSR Virtual or Physical Data Enclaves (VDE or PDE)

At present, no LLMs are available within the VDE or PDE, so this is not an option.

Study-level Metadata and Data-Related Publications

ICPSR shares its study-level metadata records to promote wider awareness and use of ICPSR's social science data resources. ICPSR metadata records are licensed under a Creative Commons Attribution-Noncommercial 3.0 United States License. Be aware that the license requires that any user of the records gives credit to ICPSR, which can be a challenge with respect to LLMs.

(ICPSR would like to credit Sebastian Karcher at Syracuse University for originating this taxonomy of LLMs)