Human-Validated AI-Assisted Document Scraping

To systematically identify the policy capacities associated with the implementation of Artificial Intelligence (AI) in the health and justice sectors, we employed a **Human-Validated AI-Assisted Document Scraping (HVAI-ADS)** protocol. This approach combines human-led document collection with large language model (LLM)-assisted qualitative analysis and subsequent human validation, allowing for transparent, replicable, and evidence-based analysis.

The unit of analysis in this study is the **country–sector pair**, resulting in ten analytical units: Argentina–Health, Argentina–Justice, Brazil–Health, Brazil–Justice, Chile–Health, Chile–Justice, Colombia–Health, Colombia–Justice, Mexico–Health, and Mexico–Justice. Each unit comprises all AI applications (AIA) identified for that specific country and sector in the AISPAL database. Across the five countries, we identified 23 AI applications in the justice sector and 17 in the health sector, each constituting empirical evidence used to characterize the policy capacities of its corresponding country–sector unit.

The document collection process began with the project's AI applications AISPALL database, which contains descriptive information for every application, including sector, responsible public organization, and other identifying characteristics. Using this information, human research assistants conducted a targeted search for documentary evidence relating to each individual application. Evidence was collected from four complementary sources: (1) the project's application database, (2) official sources such as government websites, institutional reports, technical documentation, or legal instruments, (3) research produced by academic institutions, think tanks, or specialized civil society organizations, and (4) journalistic sources documenting implementation processes, public debates, or reported impacts. This multi-source strategy was intended to maximize both the breadth and reliability of the available evidence.

Rather than analyzing websites directly, all collected documents were downloaded and consolidated into PDF portfolios (see appendices). This decision addressed an important methodological limitation encountered during early pilot analyses, where direct references to websites proved unstable because online content frequently changed or became unavailable. By preserving all documentary evidence in static PDF format, the research created a permanent and reproducible corpus for analysis. One portfolio was created for each country–sector unit, resulting in ten PDF portfolios, each containing all available documentation for the AI applications belonging to that analytical unit.

Each PDF portfolio was analyzed independently using a dedicated LLM prompt. Separate prompts were developed for the health and justice sectors facing small nuances. Analysis was conducted sequentially, processing one country–sector

portfolio at a time (e.g., Argentina–Justice, Brazil–Justice, Chile–Justice, etc.), thereby preventing contextual contamination across cases.

The analytical prompts incorporated several methodological safeguards. First, each prompt included the Hypotheses Portfolio document, which operationalized the theoretical expectations regarding the nine policy capacities derived from the analytical framework. These hypotheses provided the coding criteria used by the LLM to identify relevant evidence within the documents. Second, the prompts explicitly defined the types of documentary evidence considered valid for supporting or rejecting each hypothesis. Third, the prompt included the verified list of AI applications corresponding to the specific country–sector unit under analysis. This restriction was necessary because many documents referred to additional AI initiatives outside the scope of the study; the LLM was instructed to consider only the information provided in the set of documents designed for the purpose.

For every country–sector portfolio, the LLM generated two structured outputs. The first was an evidence matrix, organizing documentary evidence according to the nine policy capacity hypotheses across analytical, operational, and political capacities and their respective individual, organizational, and systemic levels. The second was an application-level evidence table, documenting for each AI application the specific supporting evidence together with the exact PDF page from which the evidence had been extracted. These outputs constitute the document scraping appendix accompanying the study and provide a transparent audit trail linking every analytical claim to its original documentary source.

Importantly, the LLM outputs were not treated as final research findings. Instead, they served as structured evidence inventories that were subsequently reviewed in detail by human researchers. The research team examined all extracted evidence, verified its accuracy and contextual interpretation and synthesized the validated evidence into the country–sector diagnoses reported in the Findings section. Likewise, the comparative results presented in Tables 3 and 4 were produced through this process of human interpretation rather than through automated AI classification alone.