NBER WORKING PAPER SERIES

DEEP RESEARCH ON A LOOP: USING AI AGENTS TO CONSTRUCT ECONOMIC DATASETS

Santiago Afonso
Sebastian Galiani
Ramiro H. Gálvez
Raul A. Sosa

Working Paper 35188
http://www.nber.org/papers/w35188

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
May 2026

Deep Research on a Loop: Using AI Agents to Construct Economic Datasets
Santiago Afonso, Sebastian Galiani, Ramiro H. Gálvez, and Raul A. Sosa
NBER Working Paper No. 35188
May 2026
JEL No. C0

ABSTRACT

Constructing datasets from primary sources is one of the costliest tasks in empirical economics. We propose Deep Research on a Loop (DRIL), a methodology that uses AI agents to assemble datasets from publicly available sources. DRIL applies a fixed research instrument across a mapped unit space (e.g., countries by years), with a two-stage architecture separating design from implementation. The instrument specifies variables and coding rules, an evidence policy governs sources and citations, and data quality mechanisms track gaps and uncertainty explicitly. We exercise DRIL on a 2025 update of the Global Tax Expenditures Database for eight Latin American and Caribbean countries. The run produces 129 sources and 136 evidence records, covering 22 qualitative fields fully and 6 quantitative estimate types with documented gaps, at the cost of a standard LLM subscription comparable to a few hours of research-assistant work. We argue that even partial automation of dataset construction can shift the production function of empirical economics.

Santiago Afonso
University of Buenos Aires
santiago.afonso@gmail.com

Ramiro H. Gálvez
Universidad Torcuato Di Tella
rgalvez@utdt.edu

Sebastian Galiani
University of Maryland, College Park
Department of Economics
and NBER
sgaliani@umd.edu

Raul A. Sosa
Universidad de San Andrés
Department of Economics
rsosa@udesa.edu.ar

1 Introduction

The questions that empirical economists can answer are bounded, in practice, by the datasets they can afford to build (Gentzkow et al., 2019; Donaldson and Storeygard, 2016; Einav and Levin, 2014). For large classes of problems in comparative economics and political economy, the relevant data do not come pre-assembled. They must be constructed from primary sources: legal texts (La Porta et al., 1998; Botero et al., 2004), regulatory filings (Hoberg and Phillips, 2016), government gazettes (Atkin et al., 2018), and policy documents (Romer and Romer, 2010; Cloyne, 2013), each read and coded across jurisdictions, years, and languages. Recent advances in large language models (LLMs) are changing this. LLMs have shown strong performance on text annotation and classification tasks, with growing evidence that they can match or exceed trained human coders across a range of domains (see Section 2 for a review). Most of this work has naturally focused on annotation, the task that LLMs were first able to perform reliably. The upstream step, finding and assembling the source material, is only now becoming tractable as AI systems capable of conducting autonomous, multi-step web research have emerged.

For empirical economists building institutional, legal, or policy datasets at the country, municipality, firm, or other unit level, each unit poses a complete research problem. Producing a coded answer for a single unit requires searching for relevant sources, often across jurisdictions, languages, or formats, evaluating their authority, and reading and interpreting complex texts. It also requires resolving conflicts when sources disagree, coding the result according to predefined rules, and documenting the evidentiary basis for every value. Current approaches automate fragments of this process. As these AI capabilities advance, automating the full cycle is becoming feasible.

AI *agents*, systems that autonomously plan, use tools, and reason across multiple steps, make this possible. An LLM takes text in and returns text. A research agent can take a research question in and return sources, coded answers, and a record of what it could and could not find.

This paper contributes to the growing effort to use AI agents for full dataset construction. Alongside advances in agent capabilities, we need a structured methodology that makes agent-produced data comparable, auditable, and reproducible across units. We present *Deep Research on a Loop* (DRIL), a methodology that applies an AI research agent iteratively across a mapped unit space to produce structured, evidence-backed data records. The "loop" applies the same research instrument (a formal specification of what to measure and how) to each unit in the target population. DRIL is not a general-purpose AI tool. It targets a specific class of problems: constructing structured datasets where the underlying information exists in publicly accessible sources, where the variables of interest can be formally defined in a codebook[1], and where the resulting data must

[1] A codebook is the technical document that defines each variable in a dataset: what it measures, the values it can take, and the rules for coding source material into those values.

be auditable and traceable to primary evidence. DRIL does not eliminate the need for human judgment in dataset construction. It redirects it, from the labor of collecting data to the design of the instruments that govern collection and the review of the evidence the agent produces.

We make two contributions. First, we propose DRIL, a methodology with formally defined building blocks (a research instrument, a mapped unit space, an evidence policy, and data quality mechanisms) and a two-stage architecture that separates research design from research implementation. Second, we exercise DRIL on a real research extension: a 2025 update of the Global Tax Expenditures Database (GTED) for eight Latin American and Caribbean (LAC) countries. We report the records produced, the resources used, and what the run reveals about where the methodology delivers and where it does not yet.

The remainder of the paper is organized as follows. Section 2 reviews related work on LLMs in empirical research and AI agents for data tasks, and introduces a three-tier taxonomy of AI involvement in dataset construction. Section 3 presents the DRIL methodology. Section 4 describes the application: the GTED update for eight LAC countries. Section 5 discusses risks and limitations. Section 6 concludes.

2 Related Work

We review three connected literatures relevant to DRIL: how empirical economists are using LLMs as research tools, the text-as-data and LLM methods that anchor annotation, and the recent literature on AI agents for data tasks. A final subsection positions DRIL through a three-tier taxonomy of AI involvement in dataset construction.

2.1 LLMs in Empirical Research

A growing body of work in empirical economics studies LLMs as research tools. Korinek (2023) provides the canonical taxonomy of LLM use cases for economists, ranging from ideation and writing to data analysis and mathematical derivation, grading each by the model's reliability at the time. Cowen and Tabarrok (2023) make the parallel case for graduate teaching and learning. Two studies move beyond use cases toward formal methods. Manning et al. (2024) demonstrate "automated social science," using LLMs to propose and test causal hypotheses in silico through structural causal models. Ludwig et al. (2025) develop an applied econometric framework for using LLMs as instruments, with attention to training-data leakage and the validation samples needed when LLM-generated variables enter downstream regressions.

A separate cluster studies LLMs as objects of empirical analysis, measuring how access to a model affects worker output. Noy and Zhang (2023) run a preregistered experiment on 444 college-

educated professionals and find that ChatGPT raises productivity on writing tasks by 40% in time and 18% in quality. Cruces et al. (2026) extend this design to 1,174 adults aged 25–45 with heterogeneous educational backgrounds and find that AI access closes three-quarters of the education-based productivity gap, from 0.548 to 0.139 standard deviations. Brynjolfsson et al. (2025) exploit the staggered rollout of an AI conversational assistant across 5,172 customer-support agents and find a 15% average productivity gain. Eloundou et al. (2024) construct occupation-level LLM-exposure measures and argue that LLMs are general-purpose technologies with broad task-level reach. Labaschin et al. (2025) extend the framework to the firm.

DRIL is part of a broader shift in empirical economic research. As LLMs have grown more capable, economists have been working out where in the research process they can be put to use, and what new kinds of work become possible when they are. DRIL is one such answer: a methodology that treats the construction of a research dataset as a single, integrated process.

2.2 Text-as-Data and LLM Methods

Text has become a major source of data in empirical economics. Gentzkow et al. (2019) provide the foundational text-as-data framework, and Ash and Hansen (2023) extend it for the deep-learning era. Dell (2025) surveys the framework's practical reach, from document digitization to record linkage to large-scale classification. Within this lineage, LLMs are the latest and most general-purpose tools.

Recent evaluations show that LLMs match or exceed trained human coders on text annotation tasks. Gilardi et al. (2023) find that ChatGPT achieves 59 to 83 percent zero-shot accuracy on annotation tasks, outperforming crowd-workers by about 25 percentage points on average. Törnberg (2024) reports that GPT-4 outperforms both expert coders and crowd-workers on political text classification. Bermejo et al. (2025) show that frontier LLMs outperform outsourced human coders on complex annotation tasks in Spanish, with higher accuracy and far greater internal consistency. Bojic et al. (2025) find that LLMs match or exceed human inter-rater reliability on sentiment, political leaning, and emotional intensity. Yet they systematically underrate emotional intensity relative to humans, and both groups struggle equally with sarcasm detection. Ziems et al. (2024) evaluate LLMs across 25 computational social science benchmarks and find at least fair agreement with human coders on most classification tasks. Performance varies, with weaker results on about 30 percent of tasks.

Applications scale these methods across corpora. Fang et al. (2025) use LLMs over three million Chinese government documents to extract structured information on industrial-policy objectives, targeted industries, tones, and tools. Bäcker-Peral et al. (2025) validate a multimodal-LLM digitization pipeline on county-level historical vehicle registrations, achieve 98.6% fidelity, and

report about 100× cost savings against manual outsourcing. Within economics, two recent papers apply LLM pipelines to the literature itself. Garg and Fetzer (2025) extract claim graphs from 44,852 NBER and CEPR working papers. Galiani et al. (2026) build and validate an LLM-based measurement pipeline on 27,464 full-text journal articles to track efficiency and equity framing in economics from 1950 to 2021.

LLMs can reliably annotate the documents they are given. Bail (2024) calls content analysis "perhaps the most promising task that could be outsourced to Generative AI," while warning that hallucination and reproducibility concerns demand rigorous human validation. As capabilities have grown, the natural next frontier is the upstream task: assembling the right documents in the first place.

2.3 AI Agents for Data Tasks

AI agents can search the web, read documents, and reason about intermediate results to refine their approach toward a research objective. The ReAct framework (Yao et al., 2023) interleaves reasoning and acting to solve tasks that require information gathering, while Toolformer (Schick et al., 2023) teaches language models to invoke external tools autonomously. Wu et al. (2024) introduce AutoGen, a framework for orchestrating multi-agent conversations. Three recent surveys catalogue the literature (Xi et al., 2025; Wang et al., 2024; Yehudai et al., 2025).

In benchmarks, Xu et al. (2025) find that the best current agent solves only 30% of consequential office tasks autonomously, a sober ceiling on what unsupervised agents deliver. Visible practical realizations include "deep research" products from major AI providers (OpenAI, 2025; Perplexity, 2025) and engineering write-ups of multi-agent research systems (Hadfield et al., 2025).

From these foundations, a nascent literature has begun applying agent capabilities to research work. Three contributions sit closest to DRIL, each on a different dimension. Ma et al. (2025) present a multi-agent system that, given a natural-language instruction, assembles tabular datasets from structured online sources: conference proceedings, financial APIs, and sports databases. Liu and Quan (2025) benchmark agents on the retrieval of economic data points from 82 authoritative websites across 360 tasks, finding that the best current agent solves 46.9% against 93.3% for human experts. Both treat the data as something already on the web in structured form. What they study is how well an agent can fetch it. DRIL targets heterogeneous, often unstructured sources such as legal texts, regulatory documents, and government reports, where each observation requires interpretive reading and coding against a formal instrument. Dawid et al. (2025) propose agentic workflows for the full research lifecycle, with specialized agents for ideation, literature review, modeling, empirical analysis, interpretation, and replication. Their methodology is built for adaptive inquiry on a single project, where mid-run revision of question and methodology is the

design intent. DRIL focuses on a single phase of that lifecycle: dataset construction. It replicates that phase consistently across many units, with the codebook fixed before the run begins.

2.4 Positioning DRIL

We distinguish three tiers of AI involvement in dataset creation, ordered by increasing autonomy (Figure 1).

Tier I: LLMs as classifiers. The researcher supplies the documents and the LLM returns structured annotations. The approach has proven effective for tasks such as sentiment analysis, political stance classification, and entity extraction from text corpora. The input is a document. The output is a label or score. The LLM does not search for information.

Tier II: LLMs in multi-step pipelines. The researcher designs a data processing pipeline and the LLM handles specific extraction or transformation steps. The LLM may process multiple documents per unit, but the researcher specifies the pipeline logic, including what to read, in what order, and how to combine information.

Tier III: Agents as autonomous researchers. DRIL operates at this tier. The AI agent receives a research instrument and a unit assignment, then autonomously decides what to search for, which sources to consult, how to evaluate conflicting information, and how to code the result. The agent performs *discovery*, finding relevant sources alongside extraction and coding. The input is a research question. The output is an auditable research record with full citations.

The key difference across tiers is the locus of research judgment. In Tier I, the researcher does all the finding and the LLM does the reading. In Tier II, the researcher designs the workflow and the LLM executes prescribed steps. In Tier III, the agent exercises judgment about where to look, what to trust, and how to handle the unexpected, all subject to the constraints encoded in the research instrument and evidence policy. These tiers are cumulative, not competing. A Tier III agent performs classification and extraction (Tier I and II work) as intermediate steps within a broader research cycle. The distinction is one of scope. Tier I and II apply when the researcher has already assembled the relevant documents; Tier III applies when document assembly is part of the task.

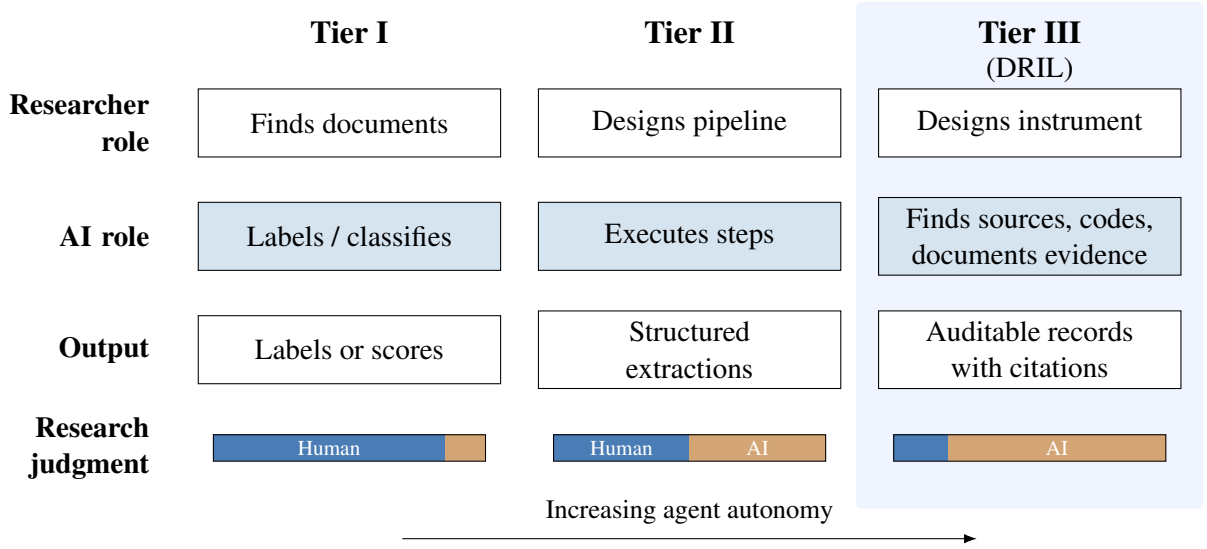

Figure 1: Three tiers of AI involvement in dataset construction. The proportion bars show the approximate share of research judgment held by the human researcher versus the AI system at each tier. DRIL operates at Tier III, where the agent autonomously discovers sources, codes variables, and documents evidence.

3 Deep Research on a Loop

3.1 Overview

Dataset construction is expensive because each unit[2] in the final dataset represents a distinct research task. DRIL is a methodology for converting open-ended web research into standardized, auditable dataset construction. It assigns each unit to an AI agent that executes a fixed research instrument and returns a fully cited research record. Figure 2 diagrams the methodology. The key idea is that DRIL imposes the same formal discipline on an AI agent that a well-managed research team imposes on its assistants.

Why does this matter? Why not simply submit individual queries to a deep research tool (such as those offered by major AI providers) and compile the results? Consider a researcher who wants subnational borrowing rules for 30 countries over 20 years. Ad hoc querying produces 600 narrative memos, each structured differently, each drawing on different sources, each resolving ambiguity by its own lights. Three problems follow. First, **consistency**: without a fixed instrument, country A's memo might code "borrowing authority" by examining the constitution while country B's memo relies on a statute, producing incomparable data. Second, **structured output**: narrative answers must themselves be coded into variables, doubling the work. Third, **auditability**: a prose summary that concludes "subnational governments may borrow subject to central

[2] A unit is the entity for which one row of the target dataset is built: a country, a firm, a country-year, or whatever the research design picks out.

approval" gives no way to verify the claim short of redoing the research from scratch. These requirements, familiar to any researcher who has managed a team of research assistants, are what separate a dataset from a collection of research memos.

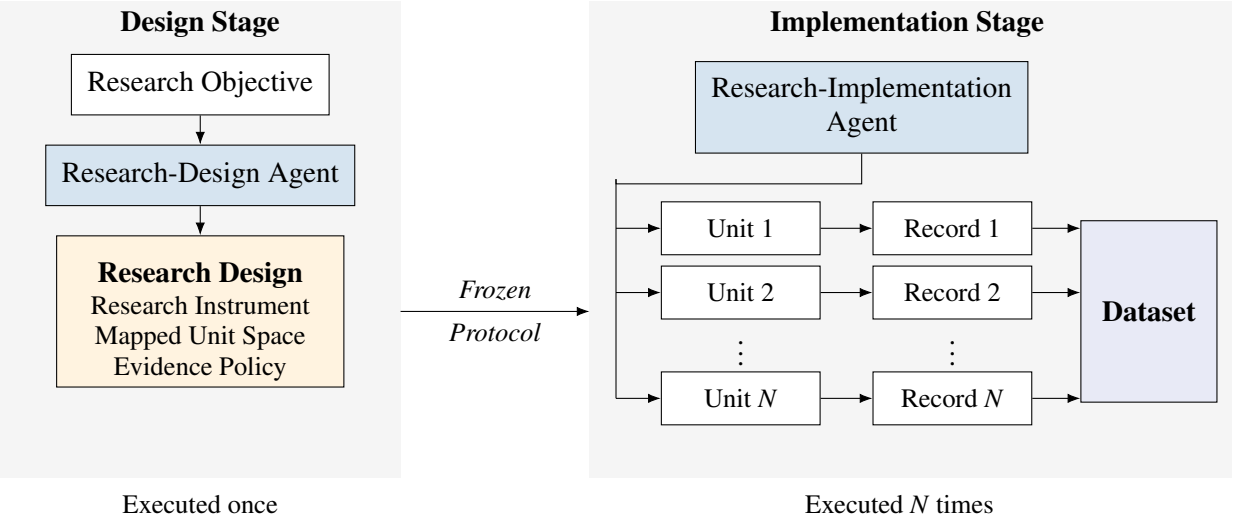

Figure 2: The DRIL Architecture. The design stage produces a frozen research design that bundles the research instrument, the mapped unit space, and the evidence policy. The implementation stage applies it independently to each unit in the mapped space, producing documented records that are assembled into the final dataset.

3.2 Building Blocks

The DRIL methodology rests on four conceptual building blocks: (i) a formal research instrument, (ii) a mapped unit space, (iii) an evidence policy with source hierarchies, and (iv) explicit data quality mechanisms. The first three are project-specific and not prerequisites the researcher must arrive with. DRIL's design stage (Section 3.3) guides their construction: a research-design agent turns a natural-language research objective into a complete specification. The fourth is methodology-wide and inherited from DRIL itself.

Research Instrument. A codebook-grade specification of what the dataset should contain. For each variable in the target dataset, it defines:

- A precise **question** (e.g., "Does the subnational government have independent tax-setting authority?")
- A **value type** and any constraints (e.g., categorical with values {full, limited, none}, or a continuous measure with defined units)

- **Coding rules** that specify how to translate source material into the variable's value domain
- A **field kind** indicating whether the variable produces a coded scalar, a coded multivalue, a narrative summary, an extractive note, an entity list, or a tabular extract, recognizing that research-grade datasets often require qualitative context alongside quantitative coding
- Optional **per-variable source overrides** when the default source hierarchy (defined in the evidence policy) is not appropriate for a particular variable

Beyond individual variables, the instrument specifies cross-cutting policies governing common situations, including what to do when a literal match is absent, how to document missing data, when to add interpretive notes, under what conditions to flag items for follow-up, and how much source detail to retain.

The research instrument is the *authoritative specification* for the entire process. This instrument drives all downstream operations: the structure of the output data, the instructions given to the research agent, and the validation of results. Changes to the instrument's substantive content (as opposed to cosmetic adjustments) trigger explicit review, ensuring that the dataset's conceptual foundations do not drift during execution.

Mapped Unit Space. The set of entities for which the research instrument will be applied. The space is defined by one or more *dimensions*, each with a specified set of values. A dimension might be a set of countries (identified by ISO 3166-1 codes), a range of years, a list of policy domains, or any other enumerable classification relevant to the research question.

The mapped unit space is the Cartesian product (or a defined subset) of these dimensions. Each element in this space, called a *mapped unit*, represents a single research assignment. For a two-dimensional space of 30 countries and 20 years, there are 600 mapped units. The research agent will execute the instrument once for each.

The formulation extends beyond the country-year panels common in economics. The mapped unit space can represent firm-quarter observations, municipality-year panels, court-case classifications, or any other structure where the same set of questions applies across a defined population of entities.

Evidence Policy. Governs how the research agent should find, evaluate, and document its sources. It has three components:

- A **source-role hierarchy** assigns priority to different categories of sources. The appropriate ranking depends on the research domain. For institutional and legal research, for instance, the hierarchy might rank constitutional provisions above statutes, statutes above regulations,

regulations above official guidance, and official guidance above academic syntheses. This hierarchy can be configured globally or overridden for specific variables where a different source ordering is appropriate.

- A **citation contract** specifies the minimum documentation required for any factual claim. In DRIL, the citation contract requires: (a) a verbatim excerpt from the source, (b) a working URL and access date for web-based sources, or a document name and page number for document-based sources, and (c) a precise locator within the source (e.g., "Article 73, paragraph 2" or "Table 3.1, row for Argentina"). This contract ensures that every cell in the final dataset can be traced back to a specific passage in a specific source.

- A set of **conflict resolution rules** specifies what the agent should do when sources disagree: prefer the more authoritative source, prefer the more recent source, report both values, or flag the conflict for human review.

Data Quality Mechanisms. Three mechanisms govern data quality:

- **Explicit data gaps.** When the research agent cannot find information for a variable in a particular unit, it does not silently produce a null value. Instead, it creates an explicit *data gap record* documenting what was searched for, what was found (if anything), and why the search was unsuccessful. Data gaps are classified by reason, such as *not found after search*, *not applicable*, *unclear definition*, *conflict unresolved*, or *out of scope*, and tracked as integral components of the dataset. This distinction between "we looked and found nothing" and "we did not look" is essential for research-grade data.

- **Structured uncertainty.** Complementing the data gap record, each observation in the dataset carries a status indicator drawn from a defined taxonomy: *answered* (the variable was coded with evidence), *not found after search* (no qualifying source was located), *not applicable* (the variable does not apply to this unit), *proxy used* (the answer relies on a substitute measure), *inferred* (the answer was derived rather than directly observed), or *conflict unresolved* (sources disagree and the conflict was not adjudicated). Where an indicator overlaps with a gap reason, the indicator serves as the cell-level summary and the gap record holds the detail. This taxonomy makes the dataset's uncertainty structure visible and queryable.

- **Append-only observation ledger.** When results are updated, whether because new evidence was found, an error was corrected, or a design change triggered re-coding, the original observation is not overwritten. Instead, a new observation row is appended that *supersedes* the prior one. The current view of the dataset reflects the latest observation for each variable-unit

combination, but the full history is preserved for audit. This append-only discipline ensures that the provenance of every value is traceable.

The difference between giving these building blocks to a human research assistant and giving them to an agent is significant. A human RA exercises implicit judgment that is difficult to audit. They may silently resolve an ambiguity, skip a source they deem irrelevant, or adjust a coding rule to fit an unexpected case. These micro-decisions are invisible unless the RA happens to document them. An agent's judgment, by contrast, is constrained by the explicit instrument. The agent's decisions are governed by the rules in the codebook and the evidence policy, and its search logs provide a traceable record of how it reached each answer. This makes errors *findable*, but it also makes the instrument's completeness more critical: what the instrument fails to specify, the agent cannot fill in with common sense.

3.3 Two-Stage Agent Architecture

DRIL separates design from execution, assigning each stage to a distinct AI agent.

The first stage is *research design*. The research-design agent receives the researcher's objective (stated in natural language) and, optionally, reference materials such as published tables, codebooks, or methodology sections from existing papers. Its task is to analyze these inputs and produce three of the four building blocks: the research instrument, the mapped unit space, and the evidence policy. The fourth block, data quality mechanisms, is methodology-wide and inherited from DRIL itself.

The design agent performs the interpretive work that, in traditional dataset construction, a principal investigator does before sending research assistants into the field. It identifies the latent analytical questions implied by the research objective, structures them as formal variable definitions, specifies coding rules and applicability conditions, recommends a source strategy, and flags design ambiguities that require human resolution.

The research-design agent *does the interpretive work once*, so the implementation agent does not need to rediscover the meaning of the research objective for each unit. The design agent produces a complete codebook, not a vague summary. The resulting instrument should be specific enough that a competent research assistant, whether human or artificial, could apply it consistently to any unit in the mapped space without further guidance.

The second stage is *research implementation*. The research-implementation agent receives the settled research instrument, the mapped unit space, and the evidence policy, and executes the instrument for each mapped unit.

For each unit, the implementation agent conducts autonomous web research: it formulates search queries, reads and evaluates sources, applies the coding rules from the research instrument,

and produces a structured research record. This record contains:

- **Coded answers** for each variable in the instrument, following the specified value types and coding rules
- **Evidence items** linked to each answer, each meeting the citation contract (verbatim quote, source locator, URL or page reference)
- **A source inventory** documenting all sources consulted, with their roles in the source hierarchy
- **Data gaps** for any variables that could not be answered, with documented reasons
- **Narrative notes** where the instrument specifies that qualitative context is needed (e.g., describing how a subnational fiscal arrangement deviates from the typical pattern)
- **Search logs** recording what queries were executed and what results were obtained

The implementation agent operates under strict operational constraints. It must stay within the scope of the approved design. It cannot add variables, broaden the unit space, or reinterpret the coding rules. It must use the source-role hierarchy specified in the evidence policy. It must record uncertainty honestly rather than guessing. And it must produce output that conforms exactly to the structure derived from the research instrument.

Execution begins from a *frozen protocol* that locks the research instrument version, the evidence policy, the unit roster, and the operational parameters. This ensures that the conditions under which data were collected are fully documented and reproducible. In practice, execution may reveal problems in the research design that were invisible at the design stage: a coding rule that does not account for a common legal structure, or a source hierarchy that omits the most informative source type for a particular region. DRIL supports protocol versioning and re-execution, so that design-stage corrections can be applied and their effects tracked without discarding the original run.

The two execution stages do not exhaust the methodology. DRIL also defines a verification stage that operates as an audit on top of what the implementation agent produces: a separate agent re-reads the cited sources for each coded answer and judges whether the recorded evidence supports the value. The verification stage is conceptually part of DRIL, but not a third execution stage in the construction loop. It can run at any time after implementation completes, including after an initial release of the dataset. We return to its role in the GTED-update application of Section 4.

DRIL is a methodology built to compose with any sufficiently capable AI agent host. It runs on Claude Code, OpenAI Codex, or comparable systems, leveraging whatever search, reading, and reasoning capabilities the host provides. As host capabilities advance, DRIL advances with them.

3.4 Classes of Problems

DRIL targets a specific class of dataset construction problems. It applies when:

1. **The information exists in publicly accessible sources.** The research agent conducts web research and reads documents. It cannot access proprietary databases, conduct interviews, or run experiments. DRIL is suited for institutional, legal, regulatory, and policy data that are published, even if scattered across many sources in many languages.

2. **A codebook can be defined.** The research question must be expressible as a set of formal variables with coding rules. DRIL is not suited for exploratory research where the variables of interest are not yet defined.

3. **The unit space is enumerable.** There must be a defined set of entities across which the instrument will be applied. Open-ended research ("find all countries that have this policy") is less natural for DRIL than closed-set research ("for each of these 50 countries, determine whether this policy exists").

4. **Auditability matters.** DRIL's citation contract, data gap documentation, and observation history add overhead. For datasets where the primary concern is volume rather than traceability, such as sentiment scores for millions of social media posts, simpler approaches suffice.

Within this class, DRIL is particularly well suited to problems that have historically been addressed through large-scale research assistant efforts: cross-country institutional databases, policy diffusion panels, historical compilations of legal or regulatory provisions, and any dataset where the construction methodology involves a research assistant reading sources, applying a codebook, and recording answers with citations.

DRIL is less suited to problems that require real-time data access, confidential information, or forms of human judgment that cannot be codified into a research instrument (e.g., ethnographic assessments or qualitative case comparisons that resist standardization).

4 Application: The GTED Update

We exercise DRIL on a real research extension: a 2025 update of the GTED (Redonda et al., 2021). GTED documents tax expenditures (revenue forgone through deductions, exemptions, credits, and similar provisions) and is jointly maintained by the Council on Economic Policies (CEP) and the German Institute of Development and Sustainability (IDOS). Its codebook organizes the legal, institutional, and reporting features of each country's tax-expenditure regime alongside quantitative estimates of revenue forgone.

The update reported here covers eight LAC countries: Argentina, Bolivia, Brazil, Barbados, the Dominican Republic, Ecuador, Jamaica, and Trinidad and Tobago. We chose GTED as a test case because the questions it asks of each country are well defined, the documentation is dispersed across heterogeneous primary sources, and a published version of the dataset already exists and could serve as a future benchmark.

The application reported here is an exercise of DRIL, not a finished GTED dataset. The run covers eight of the 218 GTED jurisdictions, and we queued the verification stage described in Section 3 but did not execute it. Our aim is to show how the methodology behaves in a real economic-research setting, what kinds of artifacts it produces, and where its weaknesses become legible.

4.1 The Research Instrument

The original GTED codebook is a list of 18 prompts that a human research assistant fills in by reading official reports, ministry websites, and tax-administration publications: whether a country publishes regular tax-expenditure reports, the definition of the benchmark tax system, the cost methodology, the legal grounding, the reporting frequency, the level of disaggregation by tax instrument, and so on. It works well as a guide for a careful coder who can resolve ambiguities through judgment. It is too informal, however, to drive an autonomous agent.

The research-design stage of DRIL took this list and produced a formal research instrument adapted to the implementation agent's mode of operation. The research-design agent rewrote each prompt as a structured question with an explicit answer space (typically a small set of categorical labels, sometimes a numeric estimate), an evidence requirement specifying what kind of source would and would not satisfy the question, and a default behavior when no qualifying source could be found. It also added new fields to capture provenance: the legal instrument from which each answer was drawn, the date of the report consulted, and a free-text note documenting any ambiguity the implementation agent encountered.

The resulting instrument separates qualitative variables (legal, institutional, reporting) from quantitative ones (estimates of revenue forgone, by year and by tax instrument), since the two have

Table 1: The GTED codebook before and after the design stage.

	Original GTED guide	**DRIL research instrument**
Qualitative coverage	18 narrative prompts	22 fields with explicit categorical answer spaces
Quantitative coverage	Aggregate and instrument-level figures, format varies by country	6 estimate types: total, VAT, PIT, CIT, customs duties, tax holidays
Provenance	Implicit; coder records sources at the report level	12 companion-metadata fields per country, every cell carrying source ID and quote
Evidence requirement	Coder's judgment	Verbatim quote with URL or page reference; standardized evidence-role tagging
Behavior when no source found	Cell left blank	Explicit gap record with negative-search documentation

different evidentiary requirements. It retains six quantitative estimate types, matching the breakdown commonly used in tax-expenditure reporting: total revenue forgone, plus the breakdowns by VAT, personal income tax, corporate income tax, customs duties, and tax holidays.

Table 1 summarizes the result: an instrument with 22 qualitative fields, 6 quantitative estimate types, and 12 fields of companion metadata that document evidence and gaps. Two countries received minor extensions: Barbados received three and Jamaica two additional qualitative fields, to capture institutional features of small-island Caribbean tax systems that the design agent flagged as not fitting cleanly into the standard categories.

4.2 Implementation

With the instrument frozen, the implementation agent ran independently for each of the eight countries. The agent searched, read, and coded one country at a time, drawing on a four-tier source hierarchy adapted to this domain: international organizations and benchmark databases (IMF Article IV reports, OECD reviews, CIAT studies, World Bank documents), professional firms and legal commentaries, official government sources (ministry reports, tax-administration publications, primary legislation), and the specialized economic press. For each question and each country, the agent recorded an answer, a citation trail with verbatim quotes, and either a structured estimate or a documented gap with notes on what the agent had searched for.

The run produced 129 sources distributed across the eight countries, shown in Figure 3. Argentina contributed 21 sources, the most of any country in the sample, and Brazil 13, the fewest, reflecting differences in how tax-expenditure information is published in each jurisdiction rather than in the agent's effort. Across countries, evidence records (full citations with verbatim quotes)

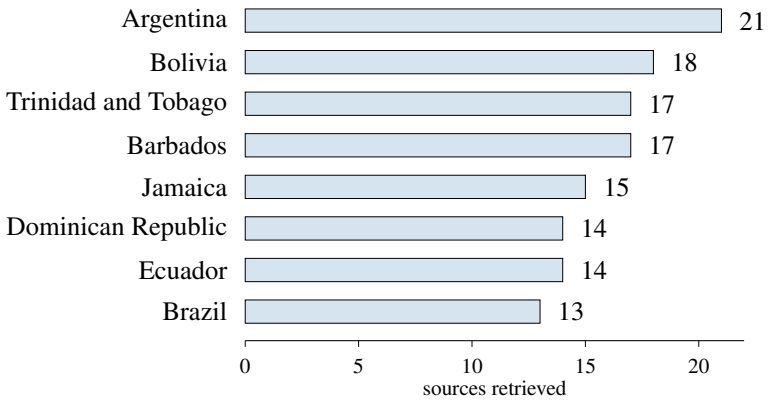

Figure 3: Sources retrieved per country in the DRIL implementation run for the GTED update; 129 in total. The four-tier hierarchy described above weights heavily toward the top in every country: international organizations (IMF, OECD, CIAT, World Bank) and official government primary sources account for the bulk of the count, with only isolated records from professional or commentary sources and the specialized economic press essentially absent.

totaled 136, with Bolivia accounting for the largest share at 37 and Argentina providing structured extracts for the value-added tax. The agent extracted 19 quantitative estimates in structured form, each with year, publishing institution, and a cited URL. Search logs recorded 110 entries, capturing the queries the agent issued and what each returned.

The full run amounted to approximately USD 21 in equivalent API cost and 260 model calls under a single underlying language model.[3] We treat this as the marginal cost of one full pass on this particular instrument across the eight countries, not as a forecast for general use, since cost varies with instrument complexity, country documentation density, and the verbosity of intermediate reasoning.

4.3 Results and a Diagnostic Finding

The clearest pattern in the output is a coverage asymmetry between the qualitative and quantitative parts of the instrument. All 181 qualitative answer cells (the 22 core fields plus the country-specific extensions, applied across the eight countries) were filled with field-specific categorical values. None defaulted to the system's standardized data-gap placeholder. The agent was able to find a source supporting a categorical answer for every legal, institutional, and reporting question,

[3]We executed the agent through the OpenAI Codex CLI under a ChatGPT subscription rather than via direct API access, so the USD 21 figure is the equivalent API cost the Codex telemetry computes from posted gpt-5.5 rates, not money charged for this run. Across the 35 pass records the run consumed 1.6M new input tokens, produced 0.24M output tokens, and re-read 11.4M cached-input tokens from prompt cache; the reported cost is the sum of the per-token rates applied to each component. We report the equivalent figure so it remains comparable to runs that incur direct API fees.

in every country.

The quantitative side tells a different story. Only 16 of the 48 country-by-instrument cells received a numeric estimate of revenue forgone as a share of GDP (one of the 16 is a combined PIT-plus-CIT figure for the Dominican Republic, drawn from a source that did not separate the two instruments). The remaining 32 cells were not silent failures. The agent left documentation for each: a partial record pointing to a source where a usable estimate could not be extracted, an explicit gap entry with negative-search notes, or, when an entire country had no usable quantitative source, a country-level gap covering the quant section as a whole.

This asymmetry is methodologically informative. It is not a hallucination problem and not a search-effort problem. It traces directly to a preference encoded in the research instrument at design time: as a rule, all quantitative estimates for a country should come from the same year and the same authoritative report so that they can be compared across instruments, with documented exceptions allowed. Where a country's most recent tax-expenditure report broke down revenue forgone by some instruments but not others (a frequent pattern in LAC reports, which often disaggregate VAT and income tax but bundle minor instruments), the constraint bound, and the agent recorded a documented gap rather than splice estimates from different sources or years. The same constraint can be relaxed in a future run by amending the instrument; the point of the exercise is that the agent honored the design choice and made its consequences legible in the output.

Box 1 reproduces a representative evidence record drawn from the Argentine VAT extraction, with the verbatim Spanish quote from the original report and an English translation. It illustrates the structure of the audit trail every coded cell carries: the legal or statistical claim, the source, the verbatim quote, and the contextual note the agent attached when the source raised an interpretive question. In this case, the aggregate VAT row in the report omitted the peso amount, and the agent reconstructed the total from the two component rows of the same table.

Box 1. Annotated evidence record for the Argentine VAT estimate.

Field. `quantitative_estimates.vat.source_original_value` (Argentina, reference year 2025).

Source. *Informe sobre Gastos Tributarios: Estimación para los años 2024–2026*, Ministerio de Economía, Dirección Nacional de Investigaciones y Análisis Fiscal, September 2025; Cuadro Nro. 1 and Cuadro Nro. 2, VAT rows.

Verbatim quote (Spanish). "IMPUESTO AL VALOR AGREGADO … 10.632.064,4 1,23 … En regímenes de promoción económica … 1.311.227,6 0,15."

Translation. Value-added tax: 10,632,064.4 million pesos, 1.23 percent of GDP; in economic-promotion regimes: 1,311,227.6 million pesos, 0.15 percent of GDP.

Coded value. 11,943,292.0 million pesos; 1.38 percent of GDP.

Note. The aggregate VAT row in the report omitted the peso amount; the agent reconstructed the total from the two component rows of the same table (VAT in tax norms plus VAT in economic-promotion regimes), keeping source and reference year unchanged.

A note on what was not done. We queued the verification stage described in Section 3, in which a separate agent re-reads the cited sources and judges whether the recorded evidence supports the recorded value, but did not execute it for this run. The numbers above therefore reflect what the implementation agent produced, not what a verification pass would confirm. The omission matters for any downstream use of these data: documented gaps and citation trails make the run auditable, but a published GTED update would require running the verification stage and resolving its findings before release.

5 Risks and Limitations

A methodology that produces datasets from primary sources is only as credible as the practices used to verify it. Two related strands in economics over the past decade (the credibility revolution in empirical work and the data-and-code transparency agenda) have reset what readers, editors, and replicators are entitled to expect from an empirical paper (Christensen and Miguel, 2018; Vilhuber et al., 2020; Hamermesh, 2007; Maniadis et al., 2017). The same expectations apply, with sharper teeth, to a dataset that an AI agent constructed: every cell carries an implicit claim about what was found, where it was found, and what was done with it. We organize the principal risks of DRIL around how the methodology fares against those expectations and where it does not, with reference to the GTED-update run reported in Section 4.

Hallucination and fabrication. Language models can generate plausible but false text, including invented quotes and citations to documents that do not say what the model claims they say. Hallucination is one of the most discussed failure modes of LLM-based research, and it would be fatal if it propagated silently into a dataset. DRIL's response is the citation contract: every

recorded answer must carry a verbatim quote, a locator, and a working URL or page reference. An invented quote will not match the cited source upon verification, which converts a silent data-quality risk into an auditable one. The GTED run produced exactly this kind of audit trail: 136 evidence records, each with a verbatim quote and source identifier. But the run also illustrates the limit of the contract. We queued the verification stage that re-reads the cited sources to confirm the recorded evidence, but did not execute it for this run. The numbers in Section 4 therefore reflect what the implementation agent produced under the contract, not what an audit pass would confirm. For any downstream use, the verification stage is not optional.

Reproducibility over time. Web sources move, change, and disappear. A run conducted today may not reproduce verbatim tomorrow, and the gap widens as the unit space and the time horizon grow. DRIL addresses this through a frozen protocol that records the research instrument version, the evidence policy, the unit roster, the operational parameters, the model version, and a search log of every query the agent issued and what it returned. The GTED run's 110 search-log entries make it possible to audit not only the records that were produced but also the records that were not. A researcher revisiting the run a year from now can read the queries the agent issued, see which reports it consulted and which it failed to find, and decide whether the gaps reflect a stable absence of evidence or an artifact of the moment of execution. A frozen protocol does not eliminate web drift, but it converts an opaque "re-run" into a documented comparison. For agent-constructed datasets, the frozen protocol is the natural analogue of the archives required by current data-and-code availability policy (American Economic Association, 2026). The search log additionally records what was looked for and not found.

Confident but incorrect coding. Beyond outright fabrication, a subtler risk is that the agent finds a real source, quotes it accurately, but interprets it incorrectly. It might cite a constitutional provision granting subnational borrowing authority without noticing a later statutory amendment that restricts it; it might pull a number from a table whose footnote rewrites the denominator. The GTED run contains a transparent example: the Argentine VAT extraction reproduced in Box 1. The aggregate VAT row in the source report omitted the peso amount, and the agent reconstructed the total from the two component rows of the same table. The evidence note documents the reconstruction, which is internally consistent but also a judgment call that a verification pass would adjudicate. Errors that escape ordinary verification are a long-running concern in empirical research, documented by the credibility-revolution literature on p-hacking and non-replication (Brodeur et al., 2016, 2020; Camerer et al., 2016, 2018); agent-coded datasets inherit the same risk. DRIL's contribution is not to eliminate them but to make them legible: every coded value carries the source, the quote, and a free-text note that records the agent's interpretive moves, and a separate verification agent can revisit the same trail.

Systematic bias from the model. Language models inherit the distribution of their training

data. They are more familiar with English-language sources, with countries and policy areas that are over-represented in the open web, and with framings that recur in their training corpora. The GTED-update run shows this signature in its source mix: international organizations (IMF, OECD, CIAT, World Bank) and official government reports dominate every country's source list, while professional and commentary sources contribute only isolated records and the specialized economic press is essentially absent. Some of this is appropriate (tier-one sources are the right anchor for a tax-expenditure dataset), but the same pattern would mask a serious gap in domains where the specialized press is the primary source of recent information. Unlike the idiosyncratic errors of human coders, which average out across a team, model bias propagates uniformly across all units. The mitigations are diagnostic rather than corrective: report the source-tier composition of the run, examine whether coverage and gap-recording rates vary systematically across geographies and source languages, and treat domains where the specialized press is essential as out of scope until the source policy can be adjusted.

Design-stage bias. The research instrument is the authoritative specification from which everything else follows; biases that enter at the design stage carry through the entire dataset. The research-design agent may emphasize variables that are commonly studied, frame coding rules that favor common-law over civil-law systems, or impose constraints whose consequences become visible only at execution time. The GTED run's clearest design-stage decision was to require all quantitative estimates for a country to come from the same year and the same authoritative report so that the breakdowns by tax instrument would be comparable. The constraint is defensible (it is the kind of choice a careful human coder would make), but it is also the binding reason that the agent left 32 of the 48 quantitative cells without numeric estimates. A different design choice would have produced different results; the design-stage decision is therefore a research output in its own right, not a hidden parameter. Pre-registering the instrument and the evidence policy before execution, in the spirit of pre-analysis plans (Olken, 2015) and the AEA RCT registry (American Economic Association and Abdul Latif Jameel Poverty Action Lab, 2013), addresses this risk by making the design stage visible and contestable.

These risks do not disqualify the methodology; they specify the conditions under which an agent-constructed dataset deserves to be read alongside a hand-coded one. The construction loop produces an audit trail (structured evidence, documented gaps, search logs, frozen protocol), and the burden the methodology places on the researcher is to use that audit trail rather than treat the agent's output as black-box data. A DRIL dataset published without verification, without source-tier reporting, and without the frozen-protocol artifact would meet none of the standards above; one published with all three would meet most of what current transparency and reproducibility expectations ask of any empirical artifact.

6 Conclusion

This paper makes two contributions. First, a formal methodology (DRIL) with defined building blocks and a two-stage architecture that separates research design from research implementation. Second, an exercise of DRIL on a real research extension (the 2025 update of the GTED for eight LAC countries) that produces a fully documented audit trail and surfaces a methodologically informative coverage pattern within the run itself.

DRIL automates the data-collection work that currently absorbs most researcher and RA time, but it does not eliminate the need for human judgment. It redirects it. Tasks that were previously informal become demanding formal requirements: designing machine-readable research instruments that must anticipate edge cases exhaustively, constructing evidence policies that make implicit judgments explicit, and auditing citation trails against source hierarchies. A researcher who once spent months directing RAs to read constitutions across dozens of countries could instead spend weeks designing the instrument that will govern the agent's research, and days reviewing the agent's output with full citation trails in hand.

The GTED update reported here is a single data point (eight countries, one underlying language model, no verification pass), but it is enough to support two conclusions. First, the methodology produces artifacts that meet current data-and-code transparency standards: every coded answer carries a verbatim quote, every gap carries a documented reason, and the frozen protocol fixes the conditions of the run. Second, the binding constraint on coverage was a design-stage decision, not an agent failure, which is the kind of finding the audit trail exists to make legible. Extending the unit space to the rest of the GTED panel, executing the verification stage, and benchmarking the result against the published GTED data are the natural next steps; we view them as continuations of the same methodological program rather than separate exercises.

If agent-based dataset construction proves viable at scale, it could change the cost structure of empirical research. At posted gpt-5.5 API rates, the run would have cost approximately USD 21, a figure that scales with workload. Instead, we paid through a single ChatGPT subscription, whose flat monthly fee is roughly the price of a few hours of research-assistant work. The same subscription would cover runs many times this size at no additional cost. A proper cost-benefit analysis at scale is the focus of follow-up work. The numbers here are one observation, not a forecast for general use; they indicate why the approach merits serious investigation.

If an agent can do the work in hours rather than months, cite every claim, and never tire, all guided by a formal instrument that ensures consistency and auditability across units, the binding constraint shifts from "can we build this dataset?" to "what dataset should we build?" That would be a different discipline, and a more productive one.

References

American Economic Association (2026), 'Data and code availability policy', AEA institutional policy. February 2026 revision; supersedes prior data policies. Accessed 2026-05-03.

American Economic Association and Abdul Latif Jameel Poverty Action Lab (2013), 'AEA RCT registry', Online registry, founded 2013. As of 2026, hosts ca. 12,000 randomized controlled trials across 170 countries. Accessed 2026-05-03.

Ash, E. and Hansen, S. (2023), 'Text algorithms in economics', *Annual Review of Economics* **15**, 659–688.

Atkin, D., Faber, B. and Gonzalez-Navarro, M. (2018), 'Retail globalization and household welfare: Evidence from mexico', *Journal of Political Economy* **126**(1), 1–73.

Bäcker-Peral, V., Meursault, V. and Severen, C. (2025), Can LLMs credibly transform the creation of panel data from diverse historical tables?, Federal Reserve Bank of Philadelphia Working Paper 25-28, Federal Reserve Bank of Philadelphia. arXiv:2505.11599.

Bail, C. A. (2024), 'Can generative artificial intelligence improve social science?', *Proceedings of the National Academy of Sciences* **121**(21), e2314021121.

Bermejo, V. J., Gago, A., Gálvez, R. H. and Harari, N. (2025), 'LLMs outperform outsourced human coders on complex textual analysis', *Scientific Reports* **15**, 40122.

Bojic, L., Zagovora, O., Zelenkauskaite, A., Vukovic, V., Cabarkapa, M., Veseljevic Jerkovic, S. and Jovancevic, A. (2025), 'Comparing large language models and human annotators in latent content analysis of sentiment, political leaning, emotional intensity and sarcasm', *Scientific Reports* **15**, 11477.

Botero, J. C., Djankov, S., La Porta, R., Lopez-de Silanes, F. and Shleifer, A. (2004), 'The regulation of labor', *The Quarterly Journal of Economics* **119**(4), 1339–1382.

Brodeur, A., Cook, N. and Heyes, A. (2020), 'Methods matter: P-hacking and publication bias in causal analysis in economics', *American Economic Review* **110**(11), 3634–3660.

Brodeur, A., Lé, M., Sangnier, M. and Zylberberg, Y. (2016), 'Star wars: The empirics strike back', *American Economic Journal: Applied Economics* **8**(1), 1–32.

Brynjolfsson, E., Li, D. and Raymond, L. (2025), 'Generative AI at work', *Quarterly Journal of Economics* **140**(2), 889–942.

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., Heikensten, E., Holzmeister, F., Imai, T., Isaksson, S., Nave, G., Pfeiffer, T., Razen, M. and Wu, H. (2016), 'Evaluating replicability of laboratory experiments in economics', *Science* **351**(6280), 1433–1436.

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B. A., Pfeiffer, T., Altmejd, A., Buttrick, N., Chan, T., Chen, Y., Forsell, E., Gampa, A., Heikensten, E., Hummer, L., Imai, T., Isaksson, S., Manfredi, D., Rose, J., Wagenmakers, E.-J. and Wu, H. (2018), 'Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015', *Nature Human Behaviour* **2**(9), 637–644.

Christensen, G. and Miguel, E. (2018), 'Transparency, reproducibility, and the credibility of economics research', *Journal of Economic Literature* **56**(3), 920–980.

Cloyne, J. (2013), 'Discretionary tax changes and the macroeconomy: New narrative evidence from the united kingdom', *American Economic Review* **103**(4), 1507–1528.

Cowen, T. and Tabarrok, A. T. (2023), How to learn and teach economics with large language models, including GPT, GMU Working Paper in Economics 23-18, George Mason University Department of Economics.

Cruces, G., Fernández Meijide, D., Galiani, S., Gálvez, R. H. and Lombardi, M. (2026), Does generative AI narrow education-based productivity gaps? Evidence from a randomized experiment, NBER Working Paper 34851, National Bureau of Economic Research.

Dawid, H., Harting, P., Wang, H., Wang, Z. and Yi, J. (2025), 'Agentic workflows for economic research: Design and implementation', arXiv preprint. arXiv:2504.09736.

Dell, M. (2025), 'Deep learning for economists', *Journal of Economic Literature* **63**(1), 5–58.

Donaldson, D. and Storeygard, A. (2016), 'The view from above: Applications of satellite data in economics', *Journal of Economic Perspectives* **30**(4), 171–198.

Einav, L. and Levin, J. (2014), 'Economics in the age of big data', *Science* **346**(6210), 1243089.

Eloundou, T., Manning, S., Mishkin, P. and Rock, D. (2024), 'GPTs are GPTs: Labor market impact potential of LLMs', *Science* **384**(6702), 1306–1308. Earlier version: arXiv:2303.10130, March 2023.

Fang, H., Li, M. and Lu, G. (2025), Decoding China's industrial policies, NBER Working Paper 33814, National Bureau of Economic Research.

Galiani, S., Gálvez, R. H., Mettola La Giglia, F. and Sosa, R. A. (2026), Measuring efficiency and equity framing in economics research: LLM-based evidence from 1950 to 2021, NBER Working Paper 34714, National Bureau of Economic Research.

Garg, P. and Fetzer, T. (2025), 'Causal claims in economics', arXiv preprint. arXiv:2501.06873. Submitted Jan 2025; revised Feb 2026.

Gentzkow, M., Kelly, B. and Taddy, M. (2019), 'Text as data', *Journal of Economic Literature* **57**(3), 535–574.

Gilardi, F., Alizadeh, M. and Kubli, M. (2023), 'ChatGPT outperforms crowd-workers for text-annotation tasks', *Proceedings of the National Academy of Sciences* **120**(30), e2305016120.

Hadfield, J., Zhang, B., Lien, K., Scholz, F., Fox, J. and Ford, D. (2025), 'How we built our multi-agent research system', Anthropic Engineering blog. Published 2025-06-13; accessed 2026-05-03.

Hamermesh, D. S. (2007), 'Viewpoint: Replication in economics', *Canadian Journal of Economics* **40**(3), 715–733.

Hoberg, G. and Phillips, G. (2016), 'Text-based network industries and endogenous product differentiation', *Journal of Political Economy* **124**(5), 1423–1465.

Korinek, A. (2023), 'Generative AI for economic research: Use cases and implications for economists', *Journal of Economic Literature* **61**(4), 1281–1317.

La Porta, R., Lopez-de Silanes, F., Shleifer, A. and Vishny, R. W. (1998), 'Law and finance', *Journal of Political Economy* **106**(6), 1113–1155.

Labaschin, B., Eloundou, T., Manning, S., Mishkin, P. and Rock, D. (2025), 'Extending "GPTs are GPTs" to firms', *AEA Papers and Proceedings* **115**, 51–55.

Liu, Z. and Quan, Y. (2025), EconWebArena: Benchmarking autonomous agents on economic tasks in realistic web environments. arXiv preprint arXiv:2506.08136.

Ludwig, J., Mullainathan, S. and Rambachan, A. (2025), Large language models: An applied econometric framework, NBER Working Paper 33344, National Bureau of Economic Research.

Ma, T., Qian, Y., Zhang, Z., Wang, Z., Qian, X., Bai, F., Ding, Y., Luo, X., Zhang, S., Murugesan, K., Zhang, C. and Ye, Y. (2025), AutoData: A multi-agent system for open web data collection, *in* 'Advances in Neural Information Processing Systems (NeurIPS 2025)'.

Maniadis, Z., Tufano, F. and List, J. A. (2017), 'To replicate or not to replicate? exploring reproducibility in economics through the lens of a model and a pilot study', *Economic Journal* **127**(605), F209–F235.

Manning, B. S., Zhu, K. and Horton, J. J. (2024), Automated social science: Language models as scientist and subjects, NBER Working Paper 32381, National Bureau of Economic Research. arXiv:2404.11794.

Noy, S. and Zhang, W. (2023), 'Experimental evidence on the productivity effects of generative artificial intelligence', *Science* **381**(6654), 187–192.

Olken, B. A. (2015), 'Promises and perils of pre-analysis plans', *Journal of Economic Perspectives* **29**(3), 61–80.

OpenAI (2025), 'Introducing deep research', OpenAI announcement. Released 2025-02-02; accessed 2026-05-03.

Perplexity (2025), 'Introducing Perplexity deep research', Perplexity announcement. Released 2025-02-14; accessed 2026-05-03.

Redonda, A., von Haldenwang, C. and Aliu, F. (2021), The global tax expenditures database (GTED): Companion paper, Technical report, German Institute of Development and Sustainability (IDOS), Bonn.

Romer, C. D. and Romer, D. H. (2010), 'The macroeconomic effects of tax changes: Estimates based on a new measure of fiscal shocks', *American Economic Review* **100**(3), 763–801.

Schick, T., Dwivedi-Yu, J., Dessì, R., Raileanu, R., Lomeli, M., Hambro, E., Zettlemoyer, L., Cancedda, N. and Scialom, T. (2023), Toolformer: Language models can teach themselves to use tools, *in* 'Advances in Neural Information Processing Systems (NeurIPS 2023)'.

Törnberg, P. (2024), 'ChatGPT-4 outperforms experts and crowd workers in annotating political Twitter messages with zero-shot learning', *Social Science Computer Review* **42**(6), 1829–1841.

Vilhuber, L., Turrito, J. and Welch, K. (2020), 'Report by the AEA data editor', *AEA Papers and Proceedings* **110**, 764–775.

Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., Chen, X., Lin, Y., Zhao, W. X., Wei, Z. and Wen, J. (2024), 'A survey on large language model based autonomous agents', *Frontiers of Computer Science* **18**(6), 186345. arXiv:2308.11432.

Wu, Q., Bansal, G., Zhang, J., Wu, Y., Li, B., Zhu, E., Jiang, L., Zhang, X., Zhang, S., Liu, J., Awadallah, A. H., White, R. W., Burger, D. and Wang, C. (2024), AutoGen: Enabling next-gen LLM applications via multi-agent conversation, *in* 'Proceedings of the First Conference on Language Modeling (COLM)'.

Xi, Z., Chen, W., Guo, X., He, W., Ding, Y., Hong, B., Zhang, M., Wang, J., Jin, S., Zhou, E., Zheng, R., Fan, X., Wang, X., Xiong, L., Zhou, Y., Wang, W., Jiang, C., Zou, Y., Liu, X., Yin, Z., Dou, S., Weng, R., Cheng, W., Zhang, Q., Qin, W., Zheng, Y., Qiu, X., Huang, X. and Gui, T. (2025), 'The rise and potential of large language model based agents: A survey', *Science China Information Sciences* **68**(12), 121101.

Xu, F. F., Song, Y., Li, B., Tang, Y., Jain, K., Bao, M., Wang, Z. Z., Zhou, X., Guo, Z., Cao, M., Yang, M., Lu, H. Y., Martin, A., Su, Z., Maben, L., Mehta, R., Chi, W., Jang, L., Xie, Y., Zhou, S. and Neubig, G. (2025), TheAgentCompany: Benchmarking LLM agents on consequential real world tasks, *in* 'Advances in Neural Information Processing Systems 38: Datasets and Benchmarks Track'.

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K. and Cao, Y. (2023), ReAct: Synergizing reasoning and acting in language models, *in* 'International Conference on Learning Representations (ICLR 2023)'.

Yehudai, A., Eden, L., Li, A., Uziel, G., Zhao, Y., Bar-Haim, R., Cohan, A. and Shmueli-Scheuer, M. (2025), 'A survey on evaluation of LLM-based agents', arXiv preprint. arXiv:2503.16416.

Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z. and Yang, D. (2024), 'Can large language models transform computational social science?', *Computational Linguistics* **50**(1), 237–291.