AI and Research in Economics

Sebastian Galiani

Tulane University and NBER

Universidad Torcuato Di Tella

2 September 2026

This session

1. **AI Agents and Prompt Engineering in Econometric Coding**
   Galiani, Lopez, and Sosa
2. **Measuring Efficiency and Equity Framing in Frontier Economics Research: LLM-Based Evidence from 1950–2021**
   Galiani, Gálvez, Mettola La Giglia, and Sosa
3. **Divergence in Climate Change Communication: LLM-Based Evidence from the IPCC and the Press**
   Galiani, Mettola La Giglia, and Sosa
4. **Deep Research on a Loop: Agent-Driven Construction of Auditable Structured Datasets**
   Afonso, Galiani, Gálvez, and Sosa

AI Agents and Prompt Engineering in Econometric Coding

Sebastian Galiani
Tulane University and NBER

Federico A. Lopez
Universidad de San Andrés

Raul A. Sosa
Universidad de San Andrés

Motivation

- The researcher chooses the question, the method, and the identification assumptions, then writes code in Stata, R, or Python.
- Then the task is clear. Coding it is mechanical and is sometimes delegated to research assistants... now to a large language model (LLM) too.
- Coding errors can appear in that mechanical step.

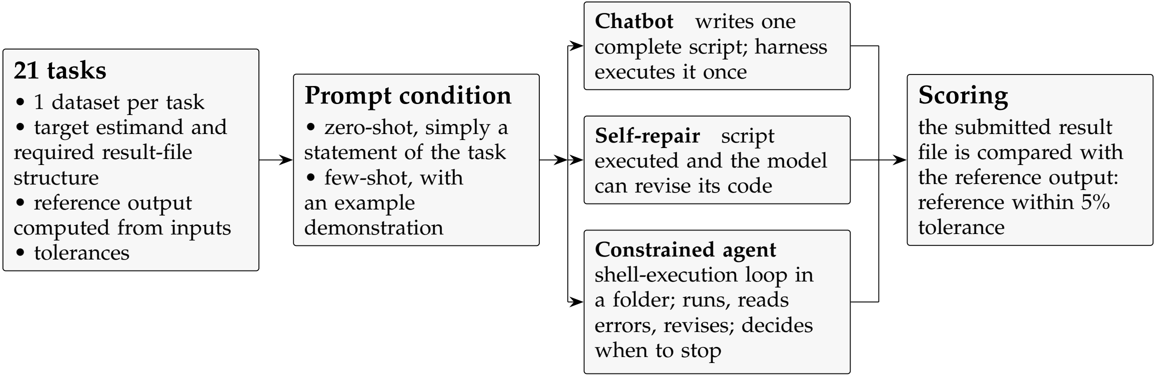
Question. How well do LLMs write the code the researcher asks for,
and how does that depend on the **prompt** (zero-shot: the bare task; few-shot: task plus a worked example),
the **agency** the harness allows, i.e., how far the model may act on its own (chatbot, self-repair, agent),
and the **software** (Stata, R, Python)?
And what are the different costs?

Where this sits in the literature

- **Code in empirical economics.** Gentzkow–Shapiro (2014) propose engineering standards for research code; Galiani–Gertler–Romero (2017) and Brodeur et al. (2026) find frequent coding errors.
- **Executable code benchmarks.** HumanEval, DS-1000, and SWE-bench execute and score generated code; Eltokhy (2026) runs a Stata benchmark; Chen et al. (2025) evaluate an econometrics agent.
- **Agency and harnesses.** ReAct, SWE-agent, and self-repair let the model act, observe, and revise; Meng et al. (2026) call that layer the *agent harness*.

Gap. No design evaluates different degrees of agency, statistical software, and prompts.

Design



21 tasks $\times$ 3 software environments $\times$ 2 prompt conditions $\times$ 3 agency conditions $\times$ 12 repetitions $=$ 4,536 runs for each of the two models, 9,072 runs in all

Three agency conditions

Coding has five stages: translate, write, execute, inspect, revise. Agency is how far past "write" the harness lets the model go.

- **Chatbot.** The model writes one script and that is it.
- **Self-repair.** The model writes code, sees the error message log, and tries again.
- **Constrained agent.** The model runs code, reads output, revises, and decides when to submit.

- Two models, each through its own production harness: Anthropic's **Claude Sonnet 4.6** (Claude Code) and OpenAI's **GPT-5.4** (Codex).
- The two endpoints match how researchers actually use these models: chatbots (most common) and coding agents (a fast-growing share).

Tasks and evaluation

21 applied econometric tasks

- Robust and clustered OLS, IV, staggered DiD, sharp RD, dynamic panel GMM, VAR, duration, discrete choice, diagnostics.
- Each task: a dataset, a target estimand, and a reference output computed by hand in all three languages.

Scoring

- A run succeeds when its reported estimates match the reference within 5% tolerance.
- The same automated rule for every run of both models.

$21 \times 3 \times 2 \times 3 = 378$ cells per model, 12 runs each: 4,536 runs per model, **9,072 in all**. Regression model with task fixed effects; wild-cluster-bootstrap inference (21 task clusters).

Result 1: agency raises task success

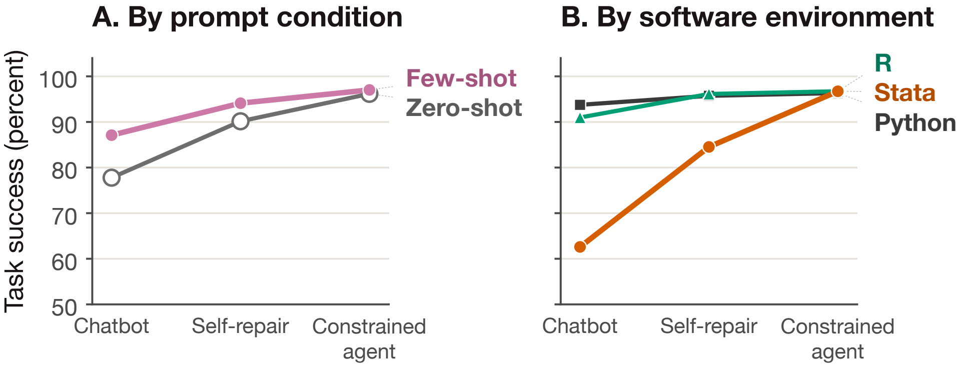
Task success (%)	Chatbot	Self-repair	Constrained agent
Claude Sonnet 4.6	83.1	93.4	**95.9**
GPT-5.4	81.8	90.9	**97.3**

- Monotone in agency for both models, in both prompt conditions.

Result 2: the gap with Python closes under the agent

- At the zero-shot chatbot, Stata sits 23 points below Python for Sonnet and 53 points below for GPT-5.4. The R gap with Python is small.
- From chatbot to constrained agent, Sonnet's zero-shot Stata success rate rises from 65.1% to 98.0%, and Stata's gap with Python is no longer distinguishable from zero for either model.
- Stata even becomes Sonnet's best environment under the agent.

Result 3: prompts and agency are not additive



Costs

	Chatbot	Self-repair	Constrained agent
Median cost per run, Sonnet	\$0.177	\$0.188	\$0.243
Median cost per run, GPT-5.4	\$0.025	\$0.026	\$0.070
Median runtime, Sonnet (seconds)	10.9	11.8	24.8
Median runtime, GPT-5.4 (seconds)	20.2	20.6	35.1

Conclusion

- Letting the model execute, inspect, and revise raises task success from about 82% to 96–97%, for less than ten additional cents per run.
- The few-shot example is valuable where the harness gives the model no other help, and adds almost nothing once the model runs as an agent.
- Under the agent, software choice largely stops mattering: the zero-shot Stata penalty disappears.

Measuring Efficiency and Equity Framing in Frontier Economics Research

LLM-Based Evidence from 1950–2021

Sebastian Galiani
Tulane University and NBER

Ramiro H. Gálvez
Universidad Torcuato Di Tella

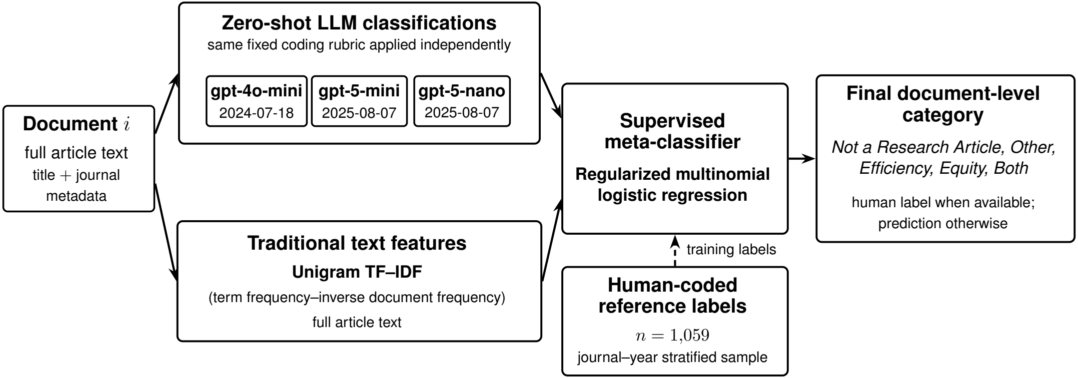
Franco Mettola La Giglia
Universidad de San Andrés

Raul A. Sosa
Universidad de San Andrés

Motivation

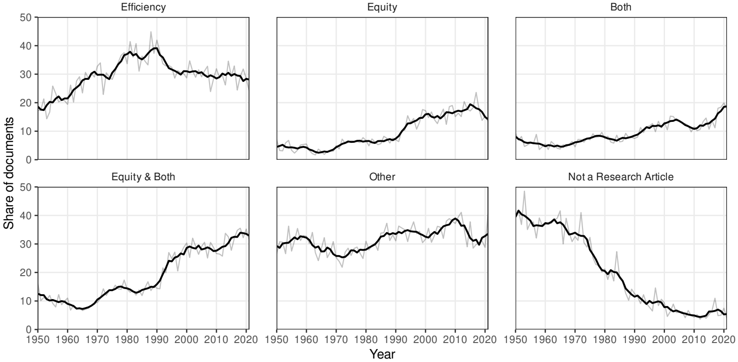
- Economists analyze policy through two often-competing normative lenses, efficiency and equity: Okun's leaky bucket.
- Whether the relative emphasis has shifted over time is an empirical question, and framing is data: every article signals which objective, if any, it treats as being at stake.
- We classify 27,464 full-text top-five articles (AER, Econometrica, JPE, QJE, ReStud), 1950–2021, into five categories: *Not a Research Article*, *Other*, *Efficiency*, *Equity*, *Both*.

Measurement pipeline



Out of sample, the ensemble classifies 81% of hand-coded articles correctly, against 66% for TF-IDF alone.

Result: two phases of normative framing



Efficiency framing falls from about 40% to about 30% after the late 1980s; papers with an equity component reach a third of the corpus by 2021.

Policy evaluation and the missing welfare accounting

- The post-1990 rise is concentrated in applied research, where policy evaluations grow to about 45% of applied publications and become substantially more equity-oriented.
- Equity adds to efficiency rather than displacing it: within applied work, the efficiency share stays comparatively stable as equity grows.
- The greater emphasis on equity rarely comes with formal welfare accounting: only 25 of 953 hand-coded policy evaluations (2.6%) report a formal cost–benefit analysis.
- The shift extends beyond the journals: *Economic Report of the President* letters show a parallel post-1990 move toward equity.

Divergence in Climate Change Communication

LLM-Based Evidence from the IPCC and the Press

Sebastian Galiani
Tulane University and NBER

Franco Mettola La Giglia
Universidad de San Andrés

Raul A. Sosa
Universidad de San Andrés

Motivation

- Technical evidence rarely reaches the public in its original form. An expert body, here the Intergovernmental Panel on Climate Change (IPCC), writes a long assessment; a smaller group condenses it into a Summary for Policymakers (SPM); the press condenses that for readers.
- Every summarizer chooses what to foreground, and every summarizer faces losses that are asymmetric in the same direction.
- Each handoff is a chance for emphasis to move, without any claim ever becoming false.

Question. How far does a claim move at each institutional handoff, which step does the moving, and can we measure it without ever needing a verdict on what the truth is?

Where this sits in the literature

- **Strategic communication.** Crawford–Sobel (1982) show that a sender with divergent preferences transmits only coarse information. Kamenica–Gentzkow (2011) add commitment: the sender is truthful but chooses what to make salient. Gentzkow–Shapiro (2010) measure slant against an external reference text.
- **Large language models (LLMs) as measurement instruments.** Grimmer–Stewart (2013) and Gentzkow et al. (2019) synthesize text-as-data. The closest methodological precedent is Galiani, Gálvez, Mettola La Giglia and Sosa (2026), who apply LLM-based structured extraction to 27,464 economics articles.
- **The IPCC relay.** Mach et al. (2016) track how revision and approval reshape the SPM across stages; De Pryck (2021) documents the negotiation itself.

Gap. No one measures the displacement against each summary's own source, claim by claim, at every step including the press.

Design: two summarization steps, measured the same way

- Each summary is scored against **its own immediate source**, never against contested ground truth. That differences out the unknown true severity.
- Three independently built LLMs judge every pair: GPT-5-mini, Claude Haiku 4.5, Gemini 2.5 Flash. Agreement across families is the measurement-error check.
- Six Assessment Report (AR) cycles, AR1 (1990) through AR6 (2021–23); press coverage spans AR3 onward.

Method, step 1: from documents to comparable claims

- **Extraction.** Each document is parsed into numbered claims carrying the IPCC's calibrated language, its fixed uncertainty vocabulary mapped to probability bands: confidence level, likelihood terms, scenario references, quantitative ranges.
- **Internal linking.** From AR4 onward every SPM claim cites its supporting Technical Summary sections in braces, so we read the link directly from the document. AR1 through AR3 predate that convention and are linked by topic instead.
- **Media links are retrieved, then chosen.** For each media claim we retrieve the five most similar SPM claims by term frequency–inverse document frequency (TF-IDF) cosine within the release window. The model then reads the article headline, the media claim, and all five candidates, and names the best-matching SPM claim, or reports that none of them match.

Method, step 2: the reporting gate

- Before any score exists, each model answers one question: *is this claim genuinely reporting that SPM finding?* Sharing a broad topic is not reporting. A claim that reports none of its candidates is dropped.
- **Two of three must agree.** A pair enters the sample only when at least two of the three models judge it genuine reporting, so no single model can set the sample.
- The gate is strict: of about 149,000 topic-bearing media claims, **25,305** survive, roughly one in six. This is the sample's central validity safeguard.

Method, step 3: scoring the displacement

- Every kept pair is scored on three dimensions, each from -2 to $+2$:
  - **Severity shift.** Does the summary pick the more severe end of the source's range?
  - **Scenario salience.** Does it foreground the high-emission scenario?
  - **Uncertainty compression.** Does it drop the source's calibrated hedges?
- Each score is halved onto $[-1, +1]$ and the three are averaged into one composite. Positive means the summary sits nearer the severe end than its source; zero means faithful transmission.
- The media composite is the mean of the reporting models; every internal pair is scored by all three and averaged.

A worked example, 1: the press claim

- **The article.** The Guardian, 23 April 2007, two and a half weeks after an AR4 SPM was approved.
- **The sentence.**

> "*Whole areas, and indeed whole island nations, will be submerged.*"

- **One claim among many.** We break the article into the separate statements it makes about the climate, and then note what each one is about. This one is about sea level, and it uses none of the IPCC's careful wording.

A worked example, 2: its source in the SPM

- **The SPM.** The SPM yields multiple claims, at least one per finding, labeled by section and position: B.2.6 is the sixth claim of section B. To find which one the article is talking about, we use TF-IDF matching to shortlist five candidates.

Rank	Cosine	Candidate SPM claim
1	0.087	[B.2.3] some human activities in the Arctic (e.g., hunting and travel . . .
2	0.079	[B.2.2] some aspects of human health, such as heat-related mortality . . .
3	**0.078**	[B.2.6] Sea-level rise and human development are together contributing to losses of coastal wetlands and mangroves and increasing damage from coastal flooding in many areas.
4, 5	0.000	non-related claims

Verdict. Each model reads the outlet, the headline, the claim, and the five candidates, and decides whether the claim is genuinely reporting one of them. All three choose B.2.6, ranked only third by word matching: the top three share a single word with the press sentence, "areas."

A worked example, 3: the two texts side by side

The SPM, section B: observed impacts

“However, based on the published literature, the impacts have not yet become established trends. Examples include: [. . .] **Sea-level rise and human development are together contributing to losses of coastal wetlands and mangroves and increasing damage from coastal flooding in many areas.**”

The Guardian

“This would eventually add another 5m to global sea levels – again, the timescale is uncertain, but as sea level rise accelerates coastlines will be in a constant state of flux. **Whole areas, and indeed whole island nations, will be submerged.**”

The models read the outlet, headline, press sentence, and five candidates, and score.

- **Severity shift.** The SPM has flood damage. The press has nations under water.
- **Scenario salience.** The press keeps one extreme outcome.
- **Uncertainty compression.** “Contributing to . . . in many areas” becomes “will be.”

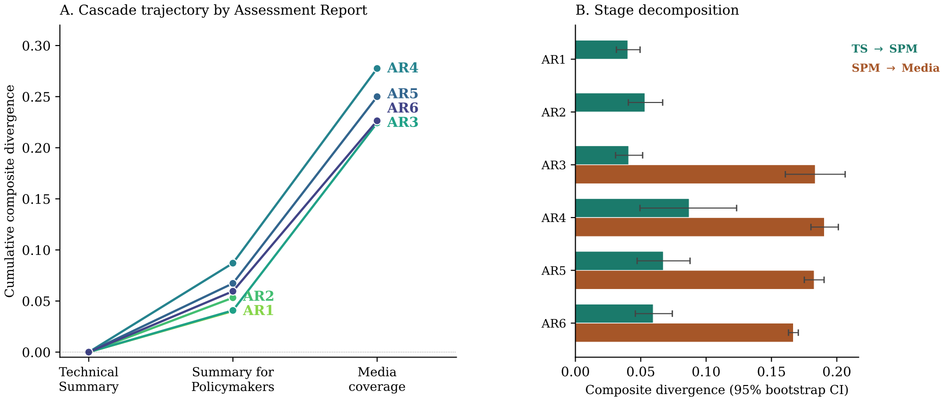
A worked example, 4: the scores and the composite

- **A relative measure.** Every score compares the press sentence with its SPM source: zero means the same, positive sharper, negative softer. The nine scores on a pair average into the composite, so its smallest move is one model moving one question by one point, about 0.06. The IPCC's own summary averages one such move per pair, the press three, a full rewrite eighteen.
- **The scale.** Integers from -2 to $+2$: $+1$ is somewhat sharper than the source, $+2$ much sharper, in alarm, worst-case focus, or dropped hedges.

Model	Severity	Scenario	Uncertainty	Reasoning, as returned
GPT	$+2$	$+1$	$+2$	Exaggerates coastal flood impacts
Claude	$+2$	$+2$	$+1$	Media predicts island nation submersion; SPM discusses coastal flooding damage generally.
Gemini	$+2$	$+2$	$+1$	Exaggerates submergence, drops hedging

- **The composite.** Each integer is halved and averaged over questions, then models: GPT $(1.0 + 0.5 + 1.0)/3 = 0.83$, the other two likewise, composite $+0.83$.

Result 1: the press step is three times the IPCC's own



The internal step pools to $+0.05$, the press step to $+0.17$.

Result 2: the shift is one-sided, not noise

- A score of $+1$ would be a wholesale rewrite, turning the SPM's calibrated "*extremely likely* ... more than half" into a flat "100 percent." Zero is faithful transmission. The average press pair is at $+0.17$.
- The distribution behind that average:

Share of pairs	Amplify	Exactly faithful	Soften
SPM $\rightarrow$ press	**74.4%**	19.4%	6.2%
Technical Summary $\rightarrow$ SPM	55.3%	29.0%	15.7%

- In the press, **twelve** claims are sharpened for every one softened, against about three and a half to one inside the IPCC.
- What moves is emphasis, not the ranges themselves.

Result 3: the measurement survives outside the models

- **Against an external coder.** An external coder judged 296 pairs blind. Coder and pipeline agree on the gate decision in **71 percent** of pairs. On confirmed pairs the coder's scores correlate **0.68** with the ensemble, reading $+0.15$ where the models read $+0.22$ on those same audited pairs: the same direction, at about two-thirds the magnitude.
- **Across model families.** On the press stage, pairwise score correlations at the claim level run 0.37 to 0.51: three families built by different firms on different data, agreeing on direction and rough magnitude.
- **Across the gate and the sample.** Tightening the gate to unanimity moves the pooled estimate from $+0.17$ to about $+0.19$; loosening it to any single model gives about $+0.15$. Five outlet-balance restrictions leave it between $+0.14$ and $+0.17$.

Conclusion

- Information is reweighted as it passes through institutions, and the changes compound along the chain.
- Inside the IPCC, the internal step is small and anchored. The press step is about three times larger and one-sided at twelve to one.
- The pattern is consistent with the rational incentives of the institutions involved; it is not a verdict on the science.
- **The method is the portable part.** A claim-level, multi-model, gated comparison of any document with its summary can ask the same question of central-bank communications, budget-office reports, or court rulings as the press relays them.

Deep Research on a Loop (DRIL)

Agent-Driven Construction of Auditable Structured Datasets

Santiago Afonso
World Bank

Sebastian Galiani
Tulane University and NBER

Ramiro H. Gálvez
Universidad Torcuato Di Tella

Raul A. Sosa
Universidad de San Andrés

Background

- Much of empirical research needs structured, comparable data across many units: country-year policy panels, firm-quarter regulatory features, judge-case classifications.
- The binding cost is often not coding a known source but locating and evaluating the evidence behind each observation.
- Deep-research agents — the “deep research” mode in ChatGPT, Gemini, and similar tools — now do that work. They search the web and synthesize what they read.
- But the researcher has little control. The agent picks the sources, applies its own rules, and returns a narrative memo.

Related work

- **LLMs as coders.** Frontier models match or exceed human coders on many tasks when the researcher supplies the source (Gilardi et al. 2023; Ziems et al. 2024).
- **Dataset assembly.** AutoData builds tabular datasets from already structured online sources (Ma et al. 2025).
- **Retrieval benchmarks.** On EconWebArena, the best agent solves 46.9% of tasks; human experts solve 93.3% (Liu and Quan 2025).
- **Research workflows.** Full-lifecycle agentic designs treat mid-run revision as a feature (Dawid et al. 2025); DRIL freezes the design first.

DRIL is complementary. It codes dispersed, partly unstructured evidence against a research instrument (a codebook), and every output can be checked.

Three tiers of AI involvement

	Tier I	Tier II	Tier III (DRIL)
Researcher	Finds documents	Designs pipeline	Designs instrument
AI	Labels, classifies	Executes steps	Finds sources, codes, documents evidence
Output	Labels or scores	Structured extractions	Auditable records with citations
Judgment	Human	Human / AI	AI

Increasing agent autonomy →

DRIL constrains that autonomy with a frozen protocol.

Deep Research on a Loop (DRIL)

- DRIL turns open-ended agent research into standardized, auditable datasets.
- **Design stage, run once.** An agent helps the researcher write a formal research design (instrument, unit space, evidence policy); the researcher freezes it into a protocol.
- **Implementation stage, run per unit.** Agents apply the frozen protocol independently. Every value carries verbatim evidence, and gaps are documented, not guessed.
- **One deployment.** A 2025 partial update of the Global Tax Expenditures Database (GTED): 86 jurisdictions, USD 22.50 in equivalent API cost, what the run would cost at metered pay-per-use prices.

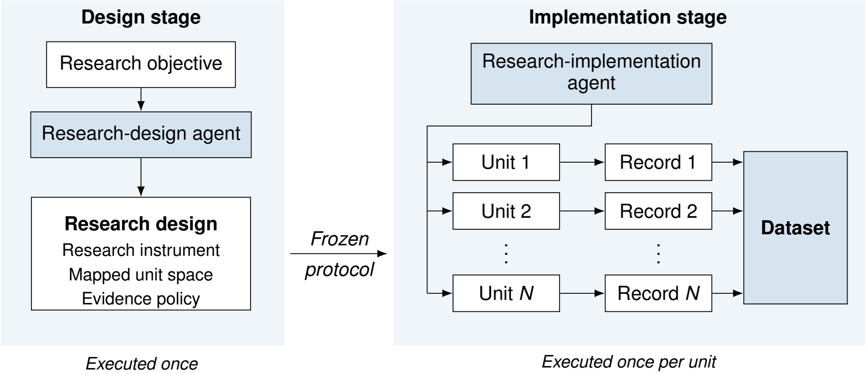
Memos are not data

Ask an agent the same question for 30 countries: 30 memos, each with its own structure, sources, and judgment calls. A dataset requires three properties that memos lack.

- **Consistency.** The same question must follow the same rules in every unit.
- **Structure.** Output must arrive variable-ready; coding prose after the fact doubles the work and revives coder discretion.
- **Auditability.** Each value must be traceable to the evidence behind it, so readers can verify without redoing the research.

These properties do not follow from model capability alone; DRIL is the harness that enforces them.

Architecture



The design stage concentrates the judgment calls; implementation then runs uniformly across units. A verification agent can later re-read cited sources and check any value.

Protocol: instrument and unit space

- **Research instrument.** The dataset codebook: for each variable, the question, the permitted value space, and the coding rules that translate sources into values.

  > “Does the subnational government have independent tax-setting authority?” $\rightarrow$ categorical $\in$ {full, limited, none}, with rules for edge cases and missing information.

- **Mapped unit space.** The enumerable set of observations the instrument covers. A design with 30 countries and 20 years defines 600 country-year units. The same logic fits firm-quarters and judge-cases.

- The design agent drafts these from the researcher’s objective and reference material; the researcher settles ambiguities and freezes the result.

Protocol: evidence policy and data quality

- **Evidence policy.** A source hierarchy (constitution $>$ statute $>$ regulation $>$ guidance $>$ academic syntheses), a citation contract, and fixed conflict-resolution rules.

  The citation contract: every recorded value carries a verbatim quote, a precise locator (“Article 73, paragraph 2”), and a working URL or page reference.

- **Data-quality mechanisms.** When no answer qualifies, the agent writes an explicit gap record with a reason from a fixed taxonomy, e.g., *not found after search*, *not applicable*, *unclear definition*, *conflict unresolved*, *out of scope*. Records are append-only, so updates never overwrite entries and the full history stays auditable.

- The implementation agent cannot add variables, expand the unit space, or reinterpret the coding rules.

What a record looks like

Field. `quantitative_estimates.vat` (Rwanda, reference year 2025).

Source. *Tax Expenditure Report FY 2023/24*, Ministry of Finance and Economic Planning (MINECOFIN), Republic of Rwanda; published 2025.

Verbatim evidence (Summary of Findings, p. 3). "Total tax expenditure (TE) in 2023/24 is Rwf 643.0 bn, representing 3.6% of GDP. VAT TE is Rwf 323.4 bn (1.8% of GDP) . . . "

Coded value. 1.8% of GDP (`source_original_value`: Rwf 323.4 bn; `estimate_year`: 2024; `source_report_year`: 2025).

Corroborating evidence (p. 6). The report's explicit benchmark-VAT definition confirms the figure measures deviations from the benchmark base, not VAT revenue.

Trust becomes an operational check: does this quote support this value? An invented quote fails it.

GTED deployment

- GTED, the Global Tax Expenditures Database, records revenue forgone through exemptions, credits, and deductions in 218 jurisdictions.
- The 2025 update ran under one fixed design: seed pilot (4), LAC (29), Africa (54). One exclusion leaves 86 analytic units, about 40% of the universe.

Sources retrieved	1,040
Evidence records	1,113
Qualitative cells	1,892 / 1,892
Quantitative estimates	103
Quantitative gaps	261
Units with estimates	66 of 86
Equivalent API cost	USD 22.50

Diagnostic finding. The 261 quantitative gaps decompose: 71% absent public reporting, 16% parsing limits, 6% unresolved conflicts. Only 16 of the 261 trace to design choices.

Noise that does not average out

An unconstrained agent codes each country with its own judgment calls: the recorded value is the truth plus noise, $x_i = x_i^* + u_i$.

- **In an average, noise cancels** by the law of large numbers.
- **In a regressor, it does not.** Measurement error attenuates the estimate toward zero, and the bias survives large N:

$$\operatorname{plim} \hat{\beta} \;=\; \beta \, \frac{\sigma^2_{x^*}}{\sigma^2_{x^*} + \sigma^2_u}.$$

- **DRIL attacks σ^2_u.** The frozen protocol removes unit-to-unit discretion, and the evidence record lets anyone check any x_i.

Standardization keeps downstream estimates consistent.

Conclusion

- Agents become infrastructure for dataset construction only inside a frozen protocol: a fixed instrument, fixed rules, and outputs bound to evidence.
- The audit trail replaces trust in the model with verification of the record. Each value is backed by a quote, a locator, and a source.
- Gap structure is information. Missingness decomposes into design choices, infrastructure limits, and true absence in the public record, each with a different remedy.

Thank you!

Sebastian Galiani

Tulane University and NBER

sgaliani@tulane.edu