A Researcher's Guide to Understanding Chinese Overseas Lending Data

Navigating the data landscape and the AI frontier

Teal Emery · Founder, Teal Insights · lte@tealinsights.com

Princeton Sovereign Finance Lab — June 2, 2026

Hi, I'm Teal.

MARKETS *Then*

EM sovereign research, MSIM

Covered China, and the countries China lent to: Angola, Zambia, Kenya, and others. Watched the cycle from the buy-side.

DATA *Then*

Research consultant, AidData

Two years operationalizing the loan-performance and restructuring dataset on Chinese overseas lending. I know the data well.

TOOLING *Now*

Founder, Teal Insights

Building SovTech: open-source sovereign-debt tools for finance ministries, the IMF, the World Bank, debt investors, and rating agencies. Philanthropic mandate, so the tools stay open.

Making Sense of Chinese Lending Data

It's complex. Here's a working map — and three realities that help you avoid simple mistakes.

1

The data

What's out there.
Which one to pick up tonight.

2

Three realities

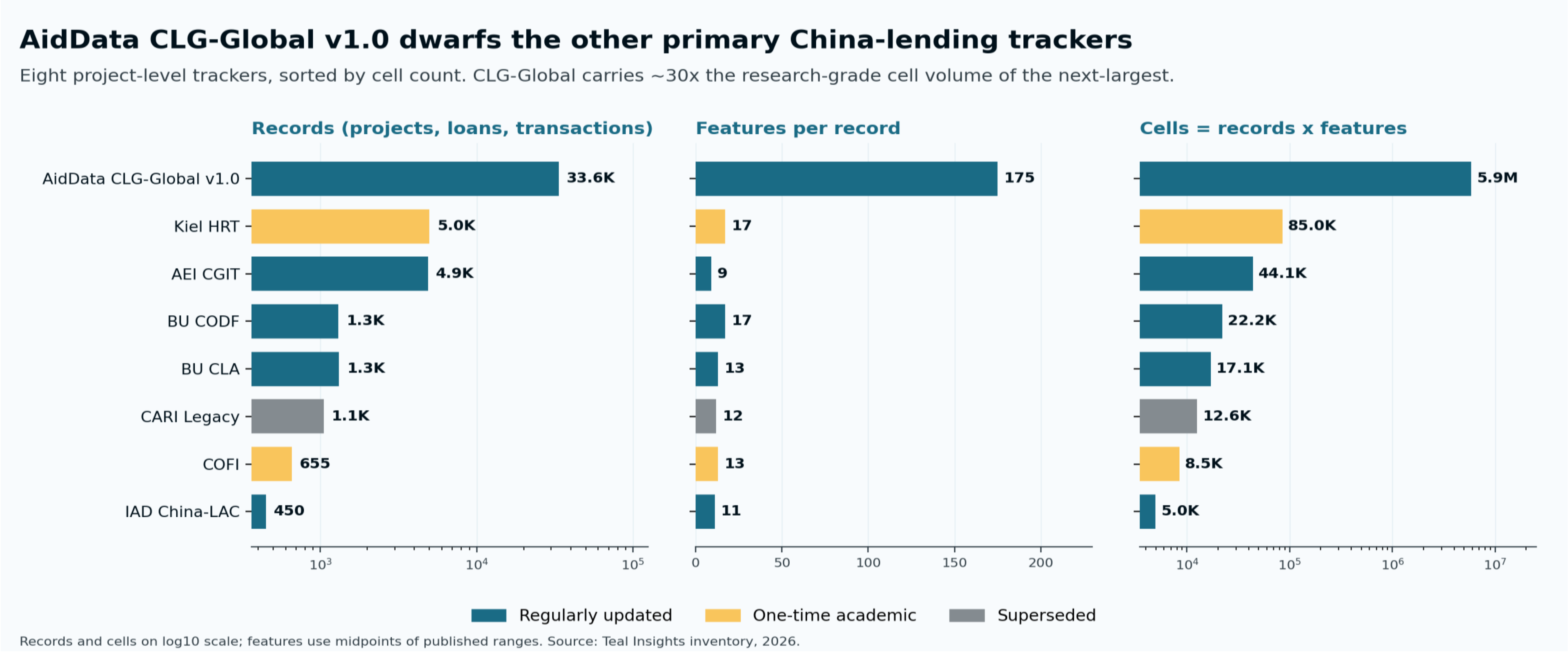
How the data actually behaves.
Avoid simple mistakes.

3

The frontier

Where LLMs change what
a small team can do.

So, You're a Large-N Political Scientist Wanting to Explore Chinese Lending Data?



AidData CLG-Global v1.0 alone ≈ 30× the combined cell volume of all other primary trackers. Use it for project-level work; the others fill specific niches.

What's Actually in AidData CLG-Global v1.0

175 fields across four families. A starting map, not a complete inventory.

BORROWER

Who's on the hook

- Country, region, ISO3
- DRA incorporation jurisdiction
- Offshore-financial-center flag
- Public-liability level (PPG, sovereign-guaranteed, private)
- Borrower ownership tab → ultimate parent

LENDER

Who's writing the check

- Funding agency + parent (CDB, Exim, ICBC, …)
- Lender type (policy bank, commercial, SOE, fund)
- Overseas branch / subsidiary flag
- Co-financiers + co-financing type
- Syndicated-loan tranche IDs

LOAN

What the deal looks like

- Amount (constant 2023 USD)
- Interest rate, maturity, grace period
- Credit instrument (29 categories)
- Flow class (ODA-like / OOF-like / OFIC)
- Concessionality + grant element

PROJECT

What it's actually for

- OECD sector code + name
- Status (Pipeline → Implementation → …)
- Infrastructure / COVID / emergency-rescue flags
- Narrative_Description (free text)
- Source URLs + Staff_Comments

Supplementary Sources Worth Knowing

When you need country-level macro context, debt stocks, or cross-border banking flows.

WORLD BANK

International Debt Statistics (IDS)

WHAT YOU GET

Public + publicly-guaranteed external debt stocks and flows. Bilateral creditor breakdowns including China.

CAVEAT

PPG only. No private-sector or SOE debt unless guaranteed.

BIS

International Banking Statistics

WHAT YOU GET

Cross-border claims by reporting bank country. Catches commercial bank lending where IDS misses it.

CAVEAT

Chinese banks under-report; OFC channels obscure ultimate exposure.

IMF

BoP, IIP, DSAs, Article IVs

WHAT YOU GET

Country-level macro: balance of payments, FX reserves, debt sustainability narratives.

CAVEAT

Macro-level. Not project-level. Some DSA detail is in PDFs only.

PART 2 OF 3

Three realities

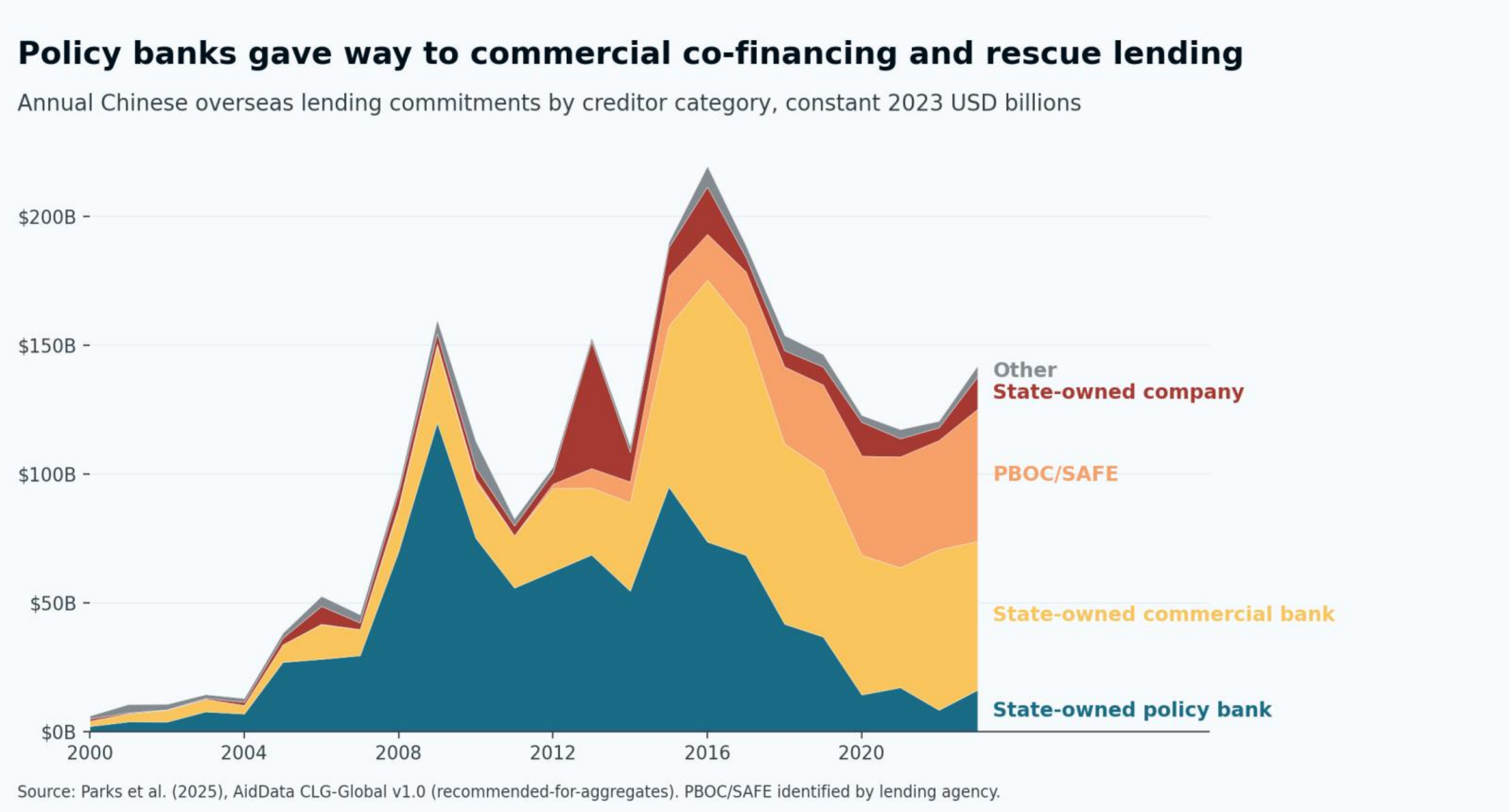
1 The data
✓ COVERED

2 Three realities

3 The frontier
UPCOMING

Reality 1 of 3

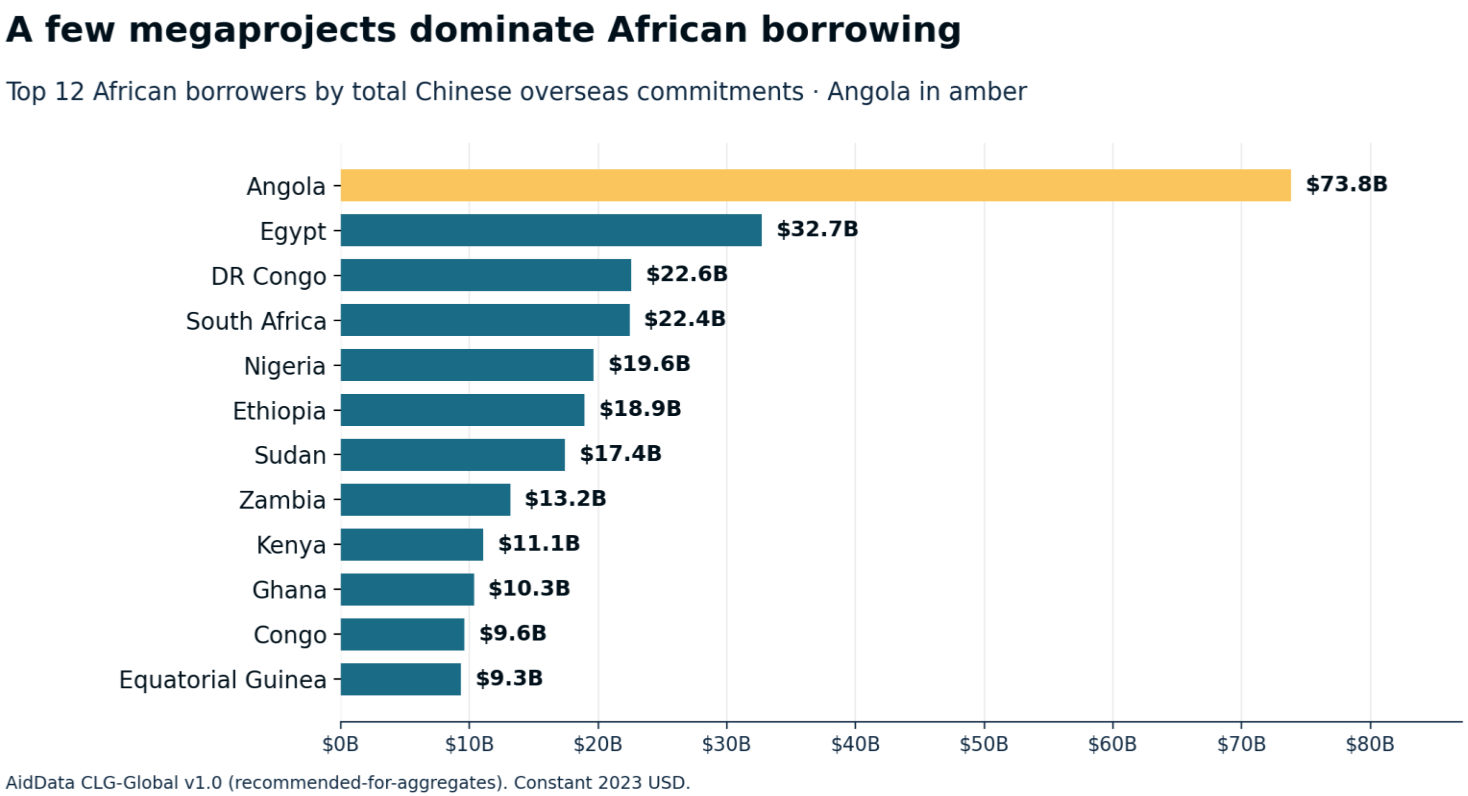
The Mix Has Shifted



KEY TAKEAWAYS

- Policy banks front-loaded commodity-backed loans to oil exporters. Peaked ~2009; faded after the 2014-15 oil crash.
- Commercial banks replaced them: co-financed, syndicated, commercial terms – majority to high-income countries (US, UK, Australia).
- PBOC/SAFE is rising, but most of it (by $) is rescue lending to distressed borrowers like Argentina and Pakistan.

Outliers Drive Aggregates

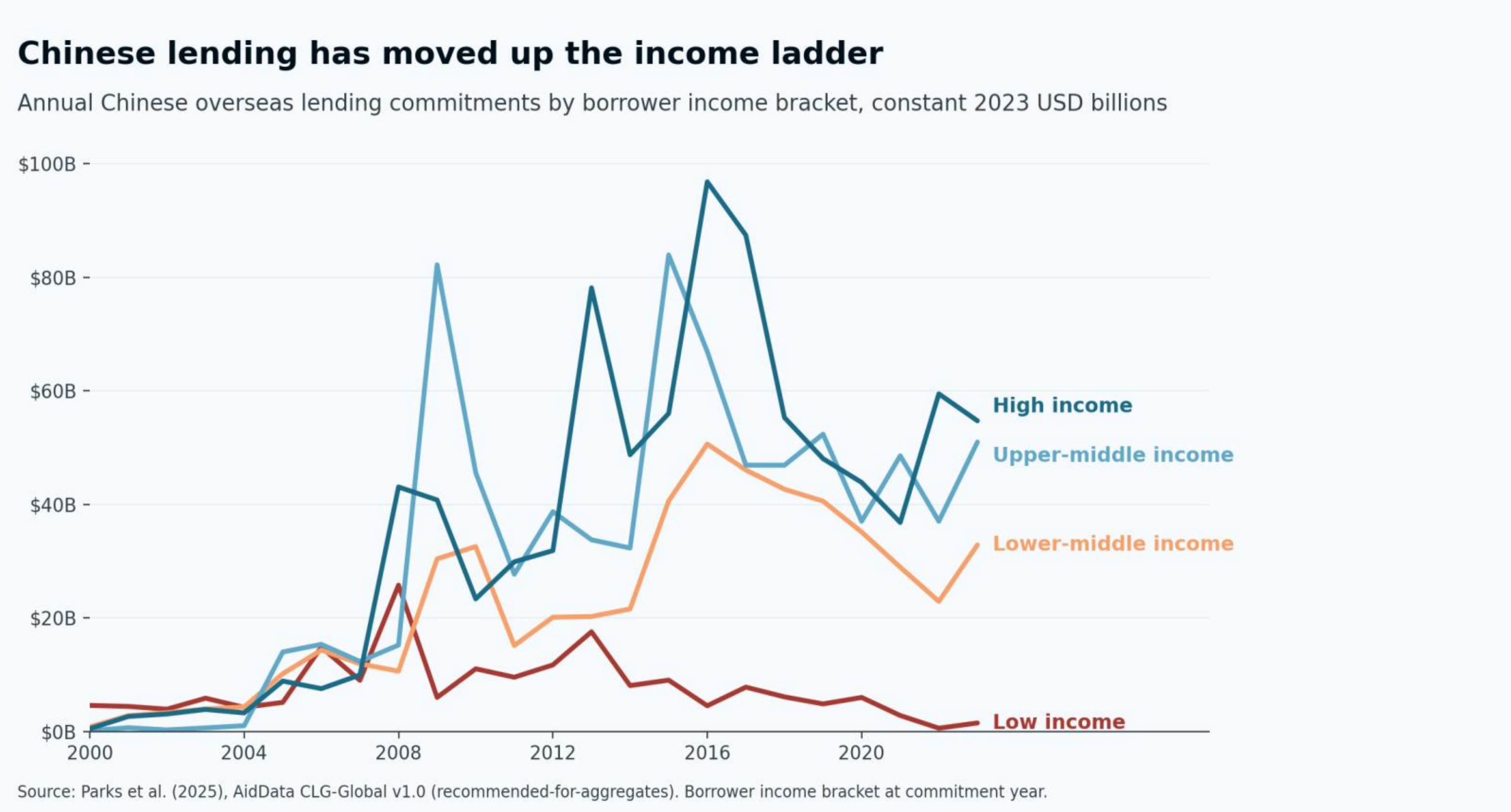


KEY TAKEAWAYS

- Beware of fat tails: a handful of mega-deals drive the totals, while most deals are far smaller. Angola in Africa is just one example.
- The same pattern recurs at every cut: rescue lending leans on Argentina and Pakistan; even within Angola, a few giant deals dominate.
- Before generalizing from an aggregate, see what's driving it - by country, sector, and deal size. Watch the median, not just the sum.

AidData CLG-Global v1.0 · Africa region · Constant 2023 USD

Lending Moved Up the Income Ladder



KEY TAKEAWAYS

- By dollars, most Chinese lending now goes to upper-middle and high-income borrowers.
- Lending to low-income countries has dwindled to almost nothing. Even lower-middle income is well off its peak.
- Part of this is graduation. Borrowers like Russia and Saudi Arabia moved up income brackets. But the tilt toward richer countries is real.

PART 3 OF 3

The frontier

1 The data
✓ COVERED

2 Three realities
✓ COVERED

3 The frontier

LLMs Unlock Ambitious Research

Individuals and small teams can now ask questions that used to need million-dollar grants and an army of RAs.

THEN.

Country-year aggregates

Big, blunt, low resolution. Think IDS.

BETTER.

Loan commitments from policy banks

Bilateral, project-level. CDB / Exim tracking starts here.

NOW.

AidData CLG-Global v1.0

EM + DM. 175 fields per project. Contracts, collateral, parties, sectors, geotagging. Oh my!

NEXT.

LLM-extracted structured data

Validated against human-labeled samples. Publishable. Scalable.

From Paragraph to Rectangle

Political scientists love rectangles. LLMs make rectangles from prose.

Project description (free text)

"In October 2017, China Development Bank committed a USD 5.0 billion 12-year loan to Mexico's state oil company Pemex. The facility is co-financed with HSBC (USD 2.0B) and ICBC (USD 1.5B), and is collateralized against Pemex's projected oil-export revenues from its Pacific Rim contracts."

LLM extraction

Typed columns (structured data)

Lender	China Development Bank
Borrower	Pemex
Country	Mexico
Year	2017
Amount (USD bn)	5.0
Tenor (yrs)	12
Sector	Energy / Oil
Co-financiers	HSBC, ICBC
Collateral	Oil export receivables

Same loan, two formats. The right one is what your regression actually reads. (Illustrative example, not a real loan record.)

What We Were Trying to Answer

THE MANDATE

Understand what China's shift to co-financing means for the energy transition in developing countries.

THE QUESTION

What does the evidence actually show about Chinese lending's role in funding energy transition?

THE PROBLEM

Lots of rumor. Existing datasets gave only partial answers — country-level aggregates or sectoral splits, not project-level structure. We needed project-level climate classification at scale.

Chen & Emery (ODI 2025), "Greener on the Other Side? Mapping China's overseas co-financing and financial innovation."

TEAL INSIGHTS

How We Validated

Two stages. Pragmatic standard — defensible, policy-useful, not Reviewer-#2-proof.

STAGE 1

Inter-model agreement

Tested 4 LLMs on the same 300 projects. Frontier models (Claude, DeepSeek, GPT-4o mini) agreed 85-89% with each other. Llama 3.3 was an outlier.

STAGE 2

Human gold-standard benchmark

Yunnan Chen and I hand-coded 300 projects to establish ground truth. Then ran each LLM against our gold standard and measured agreement.

Model	Overall agreement	Green projects	Cost (full dataset)	Time
DeepSeek v3	**91.8%**	**95.5%**	**$1.58**	**15 hours**
Claude Sonnet 3.5	85.9%	90.9%	~$4,700	16 hours
GPT-4o mini	87.3%	88.4%	~$54	11 hours
Llama 3.3 (local)	70.1%	76.2%	$0	338 hours

Findings as of January 2025 · Method walkthrough: teal-insights.github.io/soas_llm_training · Methodological annex + GitHub linked from the ODI publication.

But What About Reviewer #2?

Replicability is the elephant in the room. Spirling et al. are right. Three failure modes. The field is starting to answer them.

THE CRITIQUE (Barrie, Palmer & Spirling)

1. Closed APIs.

You can't audit what model produced the output.

2. Shifting weights.

"GPT-4" today is not "GPT-4" six months ago.

3. Non-deterministic outputs.

Same input, different result. Replication breaks.

POSSIBLE ANSWERS

1. Versioned local models

Run the model on your machine, freeze weights, ship the Docker image. (I tried this in early 2025 on a $3,500 Mac Mini. Output was terrible. Might work better now with newer open models.)

2. Multi-model + statistical validation

Egami-Hinck-Stewart-Wei DSL. Bisbee-Spirling on humans-as-no-longer-the-gold-standard. R packages already exist.

Above my pay grade. This is the conversation I'm hoping we start in Q&A.

Let's Build This Together.

We just covered: the data, three realities, and the frontier. Now the ask.

I get paid to build open-source sovereign-debt tooling.

I'd happily volunteer time to build out the validation tooling that lets us trust LLM outputs in published research.

Practitioners need pragmatism. Academics need rigor. We need each other.

Big problems. Let's do it.

Teal Emery · lte@tealinsights.com

Resources

Chen & Emery (ODI 2025): "Greener on the Other Side?"

odi.org/en/publications/greener-on-the-other-side

AidData CLG-Global v1.0 dataset

aiddata.org/data/chinas-global-loans-and-grants-dataset-1-0

The case-study walkthrough: "AI for the Skeptical Scholar"

teal-insights.github.io/soas_llm_training

Econdataverse: R + Python packages for sovereign data

econdataverse.org

Egami, Hinck, Stewart & Wei: Design-based Supervised Learning (DSL)

https://naokiegami.com/dsl/

Barrie, Palmer & Spirling 2025: Replication for Language Models

https://arthurspirling.org/documents/BarriePalmerSpirling_TrustMeBro.pdf

Teal Insights: contact

lte@tealinsights.com