IS LLM-CURATED POLICY EVIDENCE TRUSTWORTHY ENOUGH FOR CAUSAL ANALYSIS? AN AUDIT-THEN-PROPAGATE FRAMEWORK APPLIED TO GLOBAL DIGITAL-NOMAD VISA DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

Large Language Models are increasingly used to curate structured policy datasets from unstructured legal text, raising an unresolved question for researchers who build on such datasets: *is LLM-curated evidence trustworthy enough to support downstream causal and policy analysis, and how would one know?* A field-level extraction-accuracy number alone cannot answer this, since it says nothing about how a specific error moves a specific downstream conclusion. We propose and demonstrate an *audit-then-propagate* framework for staggered-adoption settings: (i) independent human verification of program-existence and adoption-timing claims against primary legal sources, rather than LLM-as-judge approximation, and (ii) a semi-synthetic experiment tracing one observed, field-specific extraction error to its measured effect on a downstream causal estimator. We instantiate this on a global panel linked to a 33-jurisdiction digital-nomad-visa (DNV) policy benchmark. On a small, fully cross-checked subset, adoption timing matches in 5 of 6 cases; tax-treatment and financial-threshold fields have no independently verified reference value and are reported as unverified. Because staggered-adoption difference-in-differences (DiD) estimators are highly sensitive to *how* treatment timing is coded, we run a semi-synthetic experiment isolating one failure mode: coding treatment to an LLM-plausible announcement date rather than the verified effective date. This error induces a consistent attenuation-toward-zero bias ($+0.25$ to $+0.28$) and cuts 95% CI coverage from 72–90% to 50.0–67.5%, for both a two-way fixed-effects (TWFE) and a Callaway–Sant'Anna (CS)-style estimator. Applying verified effective dates to real national-unemployment data yields event-time estimates that are directionally negative (lower unemployment) but statistically indistinguishable from zero, with no pre-treatment estimates available to check for differential pre-trends. We read this as evidence that curation-stage treatment-date error is a substantial, testable threat to staggered-adoption DiD conclusions, and that trustworthiness of AI-curated policy evidence is a property of the whole evidence-to-conclusion pipeline, not the extraction step alone. We offer audit-then-propagate as a reusable template for other policy domains and release the verified benchmark, audit annotations, and replication code.

1 INTRODUCTION

1.1 IS LLM-CURATED EVIDENCE TRUSTWORTHY ENOUGH FOR CAUSAL ANALYSIS?

Large Language Models now extract structured indicators from unstructured legal gazettes, policy portals, and news archives (Gilardi et al., 2023; Ziems et al., 2024; Törnberg, 2025; Tan et al., 2024), sometimes matching or exceeding human-coder agreement on well-defined classification tasks (Törnberg, 2025; 2024), though value accuracy for numeric and unit-sensitive fields lags well behind schema compliance (Geng et al., 2025). This creates a practical question for any researcher who wants to build on such curated evidence: a schema-compliant, plausible-looking extraction is not the same thing as a *trustworthy* one, since an aggregate agreement rate does not say whether the

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

errors that remain are consequential for what the data will be used for. We chose full population-level human verification as our audit design, because language-model-derived artifacts should be evaluated before they inform higher-stakes use. Rather than using an LLM-as-judge to approximate ground truth (Zheng et al., 2023), we independently verify program-existence claims, and cite a primary source, for all 190 jurisdictions; adoption-timing fields are independently re-derived for the confirmed adopters, but tax-treatment, income, and fee fields are re-derived only by a second LLM pass and are not independently human-verified in this audit (see the pipeline description and Limitations below). We then ask a further question the audit alone cannot answer: when a curated field is wrong, does it change the conclusion a downstream analysis would draw?

Agreement on a classification task does not reveal how a field-specific extraction error propagates through a downstream econometric estimator. The concern is acute for treatment-*timing* fields in staggered-adoption settings. A substantial econometrics literature shows that conventional two-way fixed-effects (TWFE) estimators can be biased under staggered timing when treatment effects are heterogeneous, motivating heterogeneity-robust alternatives such as the Callaway–Sant'Anna group-time estimator and the Sun–Abraham interaction-weighted estimator (Sun & Abraham, 2021; Callaway & Sant'Anna, 2021; Goodman-Bacon, 2021; de Chaisemartin & D'Haultfœuille, 2020), with applied guidance synthesized in Roth et al. (2023). That literature addresses bias from *how effects are aggregated* across correctly measured cohorts, which is orthogonal by construction to whether treatment dates themselves are correct. A separate strand studies bias from *mismeasured treatment status or timing*: misclassification in DiD designs can bias estimates even under an otherwise correctly specified estimator (Denteh & Kedagni, 2022; Negi, 2025; Botosaru & Gutiérrez, 2018). An LLM pipeline may extract the right adoption *year* but the wrong *month*, or conflate an announcement date with an operational launch date. The mismeasurement literature characterizes this kind of curation-stage timing error abstractly, but its magnitude and consequence for a concrete heterogeneity-robust estimator have not, to our knowledge, been directly measured. We use the digital-nomad-visa (DNV) policy wave only as the setting for this measurement. It reflects a real institutional realignment targeting location-independent workers (Ryazantsev & Ryazantsev, 2024; Bednorz, 2024; KC & Triandafyllidou, 2025). Existing accounts of it rely largely on isolated qualitative case studies or unverified aggregators (Sánchez-Vergara et al., 2023), and primary sources exist for every jurisdiction, so we can check an LLM extraction pipeline against them end to end.

1.2 Study Design & Contribution

We answer the question posed above with an audit-then-propagate framework that combines three areas not usually addressed together. *Trustworthy AI and evaluation* supplies the audit design: full population-level human verification against primary sources, with explicit reporting of uncertainty, provenance, and failure modes (Zheng et al., 2023), rather than an LLM-as-judge approximation. *Global public policy and mobility research* supplies the domain: a harmonized, primary-source-grounded digital-nomad-visa benchmark of a kind existing accounts of this policy wave have generally lacked (Sánchez-Vergara et al., 2023). *Causal econometrics* supplies the stakes: a demonstration of how treatment-timing errors the audit surfaces propagate into staggered-adoption DiD bias and coverage. Together these let us trace one curation error, at one field, through to its effect on one causal conclusion. We argue that any trustworthiness evaluation of AI-curated policy evidence needs to follow this pattern. One result from our own audit, discussed in the Limitations, qualifies claims of "global" coverage: 144 of 190 jurisdictions remain Unresolved by the LLM pipeline's own evidentiary standard after both extraction passes. Claims of global coverage in this literature, including ours, must be read against substantial, uneven gaps in primary-source availability.

Our contribution is an audited AI-assisted policy-curation pipeline for staggered-adoption settings: an LLM extraction pass, complete independent human verification against primary sources, and a semi-synthetic experiment isolating how the observed errors propagate into a DiD conclusion. We study this via DNV policy adoption, using a verified benchmark of 33 adopting jurisdictions in which a curated treatment-timing variable feeds identically into a TWFE baseline and a CS-style group-time estimator, so we can check whether curation error moves the conclusion under both. We chose this setting because primary sources exist for every jurisdiction, complete human verification is tractable at this scale, and the curated fields span several data types, which lets us compare error rates across field types instead of reporting one pooled number. We do not claim the setting generalizes; the audit-then-propagate pipeline, not the DNV findings, is the reusable contribution: a template

for asking whether AI-curated evidence is trustworthy enough for its intended use in other policy domains.

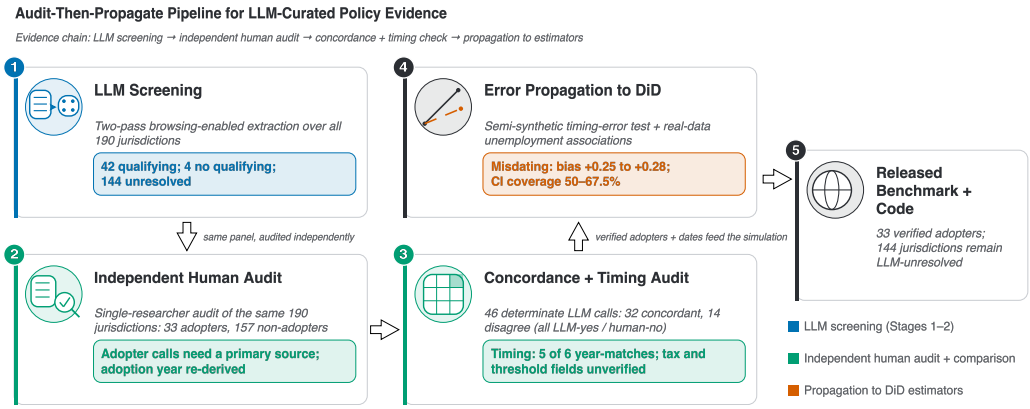


Figure 1: An audit-then-propagate framework for asking whether LLM-curated policy evidence is trustworthy enough for staggered-adoption inference. (1) A two-pass LLM screening of 190 jurisdictions yields 42 Qualifying, 4 No Qualifying, and 144 Unresolved calls. (2) A single-researcher human audit of the same panel against primary sources, run independently of the LLM output, confirms 33 adopters. (3) On the 46 determinate LLM calls, the two processes agree on 32 and disagree on 14 (all LLM-Qualifying, human-no; Cohen's $\kappa = 0.258$); adoption-year timing matches in 5 of 6 cross-checked jurisdictions. No adjudication pass has yet resolved the disagreements, and tax, income, and fee fields are not human-verified. (4) A semi-synthetic experiment that substitutes announcement dates for verified effective dates attenuates TWFE and CS-style estimates by 0.25–0.28 and lowers 95% CI coverage to 50.0–67.5%; a real-data unemployment application uses the verified dates. (5) The released benchmark contains 33 verified adopters; 144 jurisdictions remain Unresolved by the LLM pipeline, so coverage is not global.

Our contributions, organized around the three areas above:

Trustworthy AI and evaluation. A full-population audit of LLM program-existence claims against primary sources across all 190 jurisdictions, reporting Cohen's κ for program-existence concordance and field-specific match rates with Wilson 95% confidence intervals where an independently verified reference value exists; tax treatment and financial-threshold fields have no independent reference value in this audit and are reported as unverified rather than scored. We also report an error taxonomy separating internal data-quality gaps from substantive cross-coder disagreements.

Global public policy and mobility research. A harmonized, openly documented benchmark of the macroeconomic, tourism, and policy panel assembled from UN Tourism and World Bank sources, with explicit missingness and attrition accounting down to the causal-estimation sample.

Causal econometrics. A semi-synthetic treatment-date sensitivity experiment quantifying, via Monte Carlo simulation, how much bias and coverage loss a plausible LLM curation error induces in TWFE and CS-style estimators; and a real-data staggered-adoption application estimating the DNV–unemployment association using verified effective dates, reporting the cohort-support limitations that bound its conclusions. We keep the two distinct: the former is a controlled sensitivity analysis with a known, imposed effect, the latter observational and cohort-limited; neither substitutes for the other.

Relation to SDG 8. This work relates to UN Sustainable Development Goal 8 (Decent Work and Economic Growth). DNV policy is adopted partly on labor-market claims, and our real-data application measures a national-unemployment association using the same labor-market indicators tracked under SDG 8. Auditing that data and measuring how curation error propagates into a downstream labor-market conclusion checks the trustworthiness of the kind of AI-curated evidence that could otherwise inform SDG 8-relevant decisions uncritically.

2 THE DIGITALNOMADDATASET

2.1 OVERVIEW AND SCOPE

The benchmark integrates macroeconomic, tourism, and policy indicators across 190 countries and territories spanning 2008–2024 ($N = 2{,}464$ country-year observations), linked to a human-verified digital-nomad policy panel of 33 confirmed adopting jurisdictions. It comprises three tables joined by ISO3 country code: a country-year macroeconomic-and-tourism panel, a cross-sectional policy table for verified adopters, and a curated non-adopter comparison table for illustrative jurisdiction-level comparisons.

2.2 DATA SOURCES AND HARMONIZATION

Tourism fields were extracted and consolidated from UN Tourism database series (International Recommendations for Tourism Statistics 2008 framework); an ongoing transition to an upgraded UN Tourism questionnaire accounts for a meaningful share of the missingness reported below. Macroeconomic and structural controls come from World Bank Open Data, which aggregates series from the ILO, ITU, IMF, and others; we did not re-derive these from primary national sources. Tables were linked by country name and ISO3 code, and continuous policy variables (income thresholds, visa duration, fees) were captured in their source units. Table 1 summarizes the resulting field groups; a complete data dictionary with units and harmonization rules is released alongside the dataset.

Missingness is low for core macroeconomic controls but substantially higher for several tourism series (full breakdown in the appendix). Because tourism-GDP share is missing for over 40% of treated observations, our primary sensitivity analysis targets national unemployment, retaining log GDP as a lagged covariate and consistency check rather than as a separate regression outcome. As an external consistency check, arrivals growth and expenditure growth are strongly aligned both by country and pooled (full breakdown in the appendix).

The full benchmark is released at `https://anonymous.4open.science/r/Digital-Nomad-Research-A953`, including the data dictionary and the replication code for the LLM-assisted extraction workflow, the human verification pipeline, and the inferential analyses.

Table 1: Abbreviated data dictionary: the five field groups that make up the released panel, their originating source, and their intended role in the analysis. Macroeconomic and tourism fields are defined by their originating agency's standard series; policy fields are cross-sectional and verified via the two-stage audit pipeline described below.

Field group	Description	Source
`country_name`, `iso3`, `year`	Country-year identifiers	Panel key; 0% missing
GDP, unemployment, inflation, exchange rate, price-level index, internet usage	Macroeconomic and structural controls	World Bank Open Data
Tourism arrivals, expenditures, GDP share, employment	Tourism-sector indicators	UN Tourism
Policy fields (adoption year, tax treatment, income threshold, visa duration, application fee)	Cross-sectional, one row per verified adopting jurisdiction	Two-stage LLM + human audit
Non-adopter comparison fields	Qualitative alternative mobility pathways, curated non-adopter set	Two-stage LLM + human audit

3 A TWO-STAGE LLM EXTRACTION PIPELINE AND VERIFICATION

Digital nomad visa adoption dates and program attributes were extracted using a two-stage large-language-model pipeline: a program-discovery pass across the full 190-jurisdiction panel, followed

by a field-level audit of confirmed adopters, both checked independently against a single-researcher human audit of the same 190-jurisdiction panel. The pipeline defined a digital nomad visa as a legally distinct immigration status or permit whose eligibility criteria explicitly permit remote work for a non-domestic employer or clients, independent of a domestic job offer, with seven boundary cases (freelancer visas, superseded programs, multi-tiered policies, etc.) resolved via consistent rules and a source hierarchy prioritizing statutes and gazettes over ministry guidance and consular pages; the full boundary-case rules, source hierarchy, model identifiers, and execution protocol are in Appendix C.

3.1 Stage 1: Program-Discovery Extraction (190-Country Universe)

Stage 1 used browsing-enabled LLM-assisted research over the full 190-jurisdiction panel, run in two rounds (protocol in Appendix C.1). Round 1, a full-panel primary-source adjudication, classified 33 jurisdictions Qualifying Policy, 1 No Qualifying Policy, and left 156 Unresolved (Table 2). Round 2 restricted a structured discovery/trace pass to those 156 cases, upgrading 9 to Qualifying Policy and 3 to No Qualifying Policy, leaving 144 Unresolved.

Table 2: Stage-1 classification outcomes by pass, out of the full 190-jurisdiction discovery universe. ✓ Verified calls agree with the independent human-audit determination; • Partial Qualifying Policy calls require the Stage-2 field-level audit in Table 3; – Unresolved cases have no adjudicated determination. The 42 Stage-2 candidates are the LLM pipeline's output alone; the independently derived, 33-jurisdiction human-verified benchmark agrees with these 42 on 28 joint adopter calls (Section 5.2).

Stage-1 pass	Classification outcome	n	% of 190
Pass 1: full-panel primary-source adjudication (n = 190)			
	Qualifying Policy • Partial	33	17.4%
	No Qualifying Policy ✓ Verified	1	0.5%
	Unresolved Status (carried to Pass 2) – Unresolved	156	82.1%
Pass 2: structured secondary-source discovery/trace (n = 156)			
	Upgraded to Qualifying Policy • Partial	9	4.7%
	Upgraded to No Qualifying Policy ✓ Verified	3	1.6%
	Remained Unresolved Status – Unresolved	144	75.8%
Final Stage-1 disposition (n = 190)			
	Qualifying Policy • Partial	42	22.1%
	No Qualifying Policy ✓ Verified	4	2.1%
	Unresolved Status – Unresolved	144	75.8%

All 144 Unresolved cases carried Low confidence by construction; Qualifying Policy calls were predominantly High confidence (38/42), as were No Qualifying Policy calls (3/4). Every No Qualifying Policy call matched the human-audit determination exactly, making exclusionary calls on a positively identified primary source the pipeline's most reliable category.

3.2 Stage 2: Field-Level Extraction

For each of the 42 adopters (Table 3), the model re-derived adoption timing, tax treatment, income threshold, fee, and duration from primary government sources (logging protocol in Appendix C.2). LLM re-derivation was completed for 15 of 42 jurisdictions (35.7%, Medium-High confidence); the remaining 27 (64.3%) carry forward their Stage-1 lead as incomplete, Low/pending confidence.

Table 3: Stage-2 field-level status for the 42 Stage-1 Qualifying Policy candidates. • Partial rows completed a second, independent LLM re-derivation pass (*not* human verification). – Unresolved rows carry forward their Stage-1 lead without re-derivation. The 33-jurisdiction human-verified benchmark (drawn from the separate human audit below) is not identical to either the 15 LLM-re-derived rows or the 42 candidates alone; tax, income, and fee values have no human-audited reference value anywhere in this study.

Stage-2 status (n = 42)	**n (%)**	**Confidence**
LLM re-derivation completed (second pass, not human-verified) • Partial	15 (35.7%)	Medium-High
Field-level re-derivation incomplete, Stage-1 lead retained – Unresolved	27 (64.3%)	Low / pending

3.3 Documenting the Human Audit

A single domain-expert researcher independently audited the full 190-jurisdiction panel against primary sources, with Stage-2 scope limited to adoption year (protocol in Appendix C.2). All 33 coded adopters (17.4% of the panel) carry a primary source; no "yes" was assigned without one. Non-adopter coding defaulted to "no" from either an affirmative portal check (26 of 157, 16.6%) or absence of any primary source (131, 83.4%).

3.4 Error Taxonomy

Restricting to the 46 jurisdictions where the LLM Stage-1 pipeline reached a determinate call, concordance with the independent human audit is 32/46 (69.6%; Cohen's $\kappa = 0.258$, Wilson 95% CI 55.2–80.9%); all 14 disagreements run the same direction (LLM Qualifying, human no). Of the 144 LLM-Unresolved cases (not scored as agreement or disagreement, since Unresolved is not a determinate claim), the human audit separately found 5 adopters the LLM pipeline missed entirely. Combining these gives 19 jurisdictions where the two processes reach a different determination about adoption status (Table 6, error type 8). This figure spans two different comparison frames (46 determinate calls, and all 190) and should not be divided by 190 to produce an implied agreement rate. Surfaced label/citation issues in the human-audit file itself decompose into further, mechanistically distinct error types (the two largest: 15 rows conflating discovery with verification, 11 carrying an unsupported "primary" label) rather than undifferentiated noise; the full nine-category taxonomy is in Appendix A.6.

4 Digital Nomad Visa Adoption and Unemployment: A Staggered-Adoption Sensitivity Analysis

4.1 Estimand and Sample Flow

The panel spans 190 countries; 33 jurisdictions were confirmed by the independent human audit to have a qualifying DNV program, consistent with the heterogeneous policy rationales (tourism promotion, wealth attraction, immigration management) documented in comparative DNV reviews (Koskela & Beckers, 2024). Of these, 28 carry an active-equivalent status as of the audit's access date (allowing for minor label variants in the audit file), 2 (Nepal, the Philippines) are recorded as Announced but not yet active, and 3 have no recorded status value. Of the 33, 25 have a precisely dated (month/day) effective date, 8 have ambiguous or missing dates, and 15 have at least one full observed post-treatment year, which leaves 3 treatment cohorts (2021–2023) to support the main group-time estimator. This attrition foreshadows the power limitations below.

4.2 Treatment Date and Covariates

The model includes one-year lagged GDP, inflation, and internet-penetration controls, and maps treatment using exact adoption months rather than calendar-year exposure to resolve partial-year ambiguity. This matters because staggered-adoption DiD estimators are highly sensitive to how treat-

ment timing is coded: heterogeneity-robust estimators (Callaway & Sant'Anna, 2021; Goodman-Bacon, 2021; de Chaisemartin & D'Haultfœuille, 2020) and applied guidance (Roth et al., 2023) note that mistimed treatment dates can reintroduce the contamination bias these methods were designed to eliminate. Anchoring treatment to the announcement date rather than the operational launch date is one such misspecification; the semi-synthetic exercise below quantifies its cost directly in this dataset.

4.3 Semi-Synthetic Error-Propagation Experiment

A Monte Carlo simulation (200 iterations, true ATT parameterized at -1.0, seed 12345, 15-jurisdiction/3-cohort subset) shows date misspecification degrades both estimators, unequally. The misdating regime is a deterministic substitution, not a sampled or probabilistic error: for every qualifying jurisdiction that has both an announcement date and a precisely dated effective-date cohort, the announcement year is substituted for the true effective-exposure year in every simulation replicate; jurisdictions without both dates keep their true cohort. The number of jurisdictions this substitution actually reaches in a given run is data-dependent and was not independently recomputed, since the underlying audit file was not available to us (see Limitations).

Table 4: Monte Carlo simulation results for treatment-date specification (200 iterations; true ATT $= -1.0$; seed 12345; 15-jurisdiction, 3-cohort subset). Coding treatment to the verified effective date (✓ Verified) leaves both estimators nearly unbiased; substituting the real announcement date (! Error-prone), for every jurisdiction where both dates are on record, attenuates both toward zero by roughly 0.25–0.28 and cuts 95% CI coverage by 20–25 percentage points. Bias is defined as (mean estimate − true ATT), so a positive value indicates attenuation toward zero (the true ATT is negative); coverage is the share of simulation replicates whose 95% CI contains the true ATT of -1.0. Estimation details (clustering, noise process, aggregation, treatment of non-finite replicates) are in the appendix.

Estimator	Date definition	Mean estimate	Bias	95% CI coverage
CS-style	Effective date ✓ Verified	−0.984	0.016	**72.0%**
CS-style	Announcement (misdated) ! Error-prone	−0.746	0.254	**50.0%**
TWFE	Effective date ✓ Verified	−0.999	0.001	**90.0%**
TWFE	Announcement (misdated) ! Error-prone	−0.725	0.275	**67.5%**

Under correct effective-date coding, both estimators are nearly unbiased, with TWFE near-nominal coverage (90.0%) and CS somewhat under-covering (72.0%), plausibly reflecting wider standard errors under a thin, unbalanced cohort structure. Misdating to the announcement date induces a consistent attenuation-toward-zero bias of roughly $+0.25$ to $+0.28$ for both estimators and collapses coverage to 50.0–67.5%. This shows that the specific timing error tested here is, on its own, large enough to more than offset the coverage gains of using a heterogeneity-robust estimator over TWFE; we do not compare it against other candidate sources of bias (e.g., control-group contamination, functional-form misspecification), so we do not claim it is the single dominant threat to inference in this setting generally, only that it is a substantial and testable one.

4.4 Descriptive Real-Data Dynamics

Adopting jurisdictions were not experiencing acute labor-market distress at intervention: mean unemployment peaked above 10% near event time $t = -10$ and declined steadily to near 7% at $t = 0$, consistent with a decade-long macroeconomic recovery rather than crisis-driven adoption. It is a useful complement to formal pre-trend testing in the event-study estimates, though not a substitute for it (descriptive trajectory in Appendix A.5, Figure 6). This descriptive series and the event-time estimators below share the same outcome, national unemployment; the descriptive trajectory is shown for context and is not itself an event-study estimate.

A static TWFE model gives an aggregated association of -1.24 percentage points (SE $= 0.76$, $p \approx 0.100$, 95% CI: $[-2.73, 0.24]$), not distinguishable from zero at conventional levels. Throughout this section the outcome is the national unemployment rate, so a negative coefficient indicates lower unemployment among adopting jurisdictions, not lower economic output. Decomposing via the

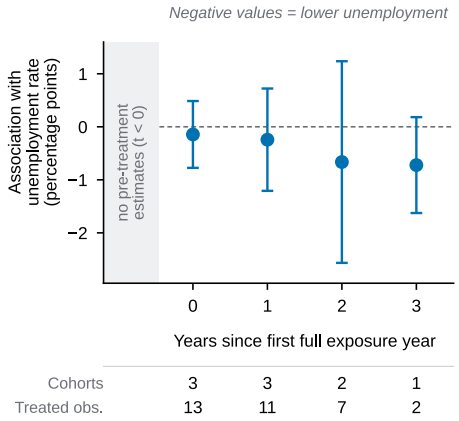


Figure 2: Event-time associations between digital nomad visa adoption and national unemployment (percentage points; negative values indicate lower unemployment). CS group-time estimates with 95% CIs; $t = 0$ is the first full exposure year. No pre-treatment estimates are available. The table shows contributing cohorts and treated-country observations at each horizon.

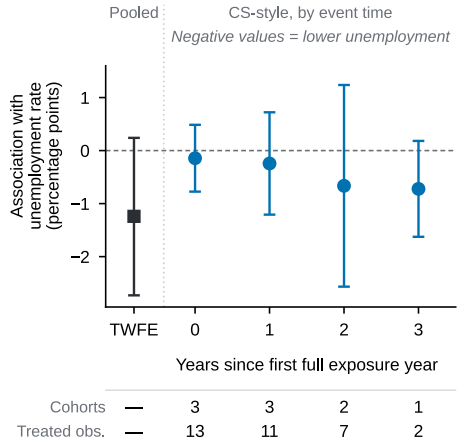


Figure 3: Static TWFE estimate (1,743 observations, 145 countries) versus the CS-style event-time profile, in percentage points of national unemployment (95% CIs; negative values indicate lower unemployment). The TWFE estimate (-1.24) pools post-treatment effects, masking a profile near zero at $t = 0$ (-0.14) and larger but thinly supported estimates at $t = 2$ (-0.66; 2 cohorts, 7 observations) and $t = 3$ (-0.72; 1 cohort, 2 observations).

Callaway and Sant'Anna (Callaway & Sant'Anna, 2021) group-time estimator instead gives event-time associations of -0.14 at $t = 0$ (SE $= 0.32$), -0.24 at $t = 1$, -0.66 at $t = 2$, and -0.72 at $t = 3$ (SE $= 0.46$), all directionally negative but none individually distinguishable from zero at conventional levels; the confidence intervals widen sharply at longer horizons: at $t = 2$ the estimate rests on only 2 cohorts and 7 treated-country observations, and at $t = 3$ on a single cohort (2021) and just 2 treated-country observations (Figure 2). This specification has no pre-treatment ($t < 0$) estimates available, so we do not interpret the growing point estimates at longer horizons as evidence of a delayed causal response, nor treat the absence of a visible pre-trend as a validated parallel-trends check; both would require estimates this design does not produce.

This is the "forbidden comparisons" and effective-sample erosion problem that heterogeneity-robust DiD estimators were built to expose (Goodman-Bacon, 2021; de Chaisemartin & D'Haultfœuille, 2020). A single static TWFE coefficient of -1.24 (Figure 3) masks a dynamic that starts near zero and is estimated with rapidly diminishing precision as the treated cohort composition thins.

Consequently, while the point estimates are directionally negative (lower unemployment) and consistent with a growing association at longer event-time horizons, definitive causal inference about long-run effects on unemployment remains limited both by the absence of pre-treatment estimates in this specification and by cohort support, which is a property of the DNV rollout itself (15 precisely dated adopters, concentrated in three cohorts) and not of the estimator. No individual event-time estimate, nor the pooled TWFE estimate, is statistically distinguishable from zero at conventional levels.

5 Conclusion and Limitations

5.1 Summary of Findings

Returning to the question posed in the introduction: our answer, at least for this setting, is that trustworthiness cannot be established by a curation-accuracy number alone, but requires tracing a specific, field-level error through to its effect on a specific downstream estimator. On the curation

side, program existence is independently human-verified for all 190 jurisdictions, with 69.6% concordance between the LLM pipeline's determinate calls and the human audit (Cohen's $\kappa = 0.258$); on a small, fully cross-checked subset of 6 jurisdictions, adoption-year timing matches in 5 of 6 cases (83.3%, Wilson 95% CI 43.6–97.0%), a figure whose sample size is too small to support a precise headline accuracy claim. Tax treatment and financial-threshold fields have no independently verified reference value in this audit and are reported as unverified rather than scored. On the inference side, the semi-synthetic experiment shows that the curation error we are most exposed to here, substituting an announcement date for a verified effective date, attenuates both a TWFE and a CS-style estimator toward zero by roughly 0.25–0.28 and cuts 95% CI coverage by 20–25 points. Applying the verified effective dates to real national-unemployment data, we obtain event-time estimates that are directionally negative (lower unemployment) and grow at longer horizons, but none is individually distinguishable from zero, the pooled TWFE estimate is not significant at conventional levels either, and the specification has no pre-treatment estimates available to check for differential pre-trends; precision at $t = 2, 3$ is further bounded by the small number of adopting cohorts. We consider these findings useful but modest, and we state their limits below.

5.2 Limitations

Dataset coverage. Of 190 jurisdictions, the LLM pipeline flagged 42 candidates, while the human audit verified 33 adopters, with 28 joint calls. All 14 disagreements were LLM-only positives, while the audit identified 5 adopters missed by the LLM. The remaining 144 were Unresolved under the LLM's evidentiary standard. This limits "global" coverage: Unresolved cases are concentrated where primary legal sources are harder to access, so the benchmark may under-represent less-digitized or non-English jurisdictions. Further open items are documented in Appendix A.4; we do not claim greater certainty than the underlying records support.

Measurement. Tax treatment and financial-threshold (income, fee) fields have *no independently verified human reference value at all* in this audit: the human audit covers only program existence and adoption timing, while tax, income, and fee values exist only as a second LLM pass's self-rederivation. We therefore report no agreement rate, accuracy figure, or Cohen's κ for these fields. Program existence (69.6% concordance, $\kappa = 0.258$) and adoption timing (5/6 year-matches on a very small, fully cross-checked subset) are the only fields we can speak to with independent evidence, and a field-level agreement rate is in any case not the same thing as downstream validity: we show adoption-timing error to be consequential for staggered-adoption DiD conclusions, but make no corresponding claim about tax-treatment or threshold errors, which remain untested.

Causal inference and generalizability. The analysis is observational, and the semi-synthetic result depends on the observed adoption-date errors and simulated outcome process in Appendix D; it may differ under other misclassification or outcome processes (Denteh & Kedagni, 2022; Negi, 2025; Botosaru & Gutiérrez, 2018). The imposed -1.0 effect is not calibrated to the real-data unemployment scale. We therefore treat the analysis as a sensitivity check, not identification of the causal effect: neither real-data event-time nor pooled TWFE estimates are statistically distinguishable from zero, and no pre-treatment estimates are available to assess pre-trends. Generalizability is also limited by the small adopter sample (33 treated units; 30 in the causal sample). Future work should expand human verification, conduct powered heterogeneity analysis, and test calibrated misclassification corrections, potentially with double/debiased machine learning (Chernozhukov et al., 2018). Appendix A.4 documents unresolved simulation, sample-exclusion, and generation-parameter items.

AI Use Statement

Generative AI is the object of study in this paper, not only a writing aid, so we disclose its use at both levels. **As research method:** both the Stage-1 program-discovery pass and the Stage-2 field-level extraction pass were LLM-assisted (OpenAI GPT-5.6 Luna via the ChatGPT interface; source cutoff/access date 2026-08-09). We recorded Temperature $= 1$, Top-P $= 1$, Seed $=$ null for these sessions; these specific parameter values are as recorded by the authors but were not independently corroborated against the audit-protocol worksheets checked in this revision, which record an access date and methodological rules but not model or decoding settings, and should be confirmed against

the authors' own session records before submission. Program-existence claims and adoption-timing fields for confirmed adopters were independently re-derived by a human researcher against a primary source before being used in any analysis, and no LLM output was accepted as ground truth on its own for those fields. Tax-treatment, income-threshold, and fee fields were re-derived only by the Stage-2 LLM pass itself and were *not* independently verified by a human researcher against a primary source; this manuscript does not claim otherwise, and reports these fields as unverified rather than as agreement or accuracy figures (Limitations). **As a writing and coding aid:** we used generative AI tools for light copy-editing and LaTeX formatting assistance during manuscript preparation, and for boilerplate/scaffolding code in the data-processing and estimation pipeline (e.g., standard plotting and table-generation snippets). We did not use generative AI tools to generate research ideas. All AI-assisted code was reviewed, tested against the reported summary statistics, and verified for correctness by the authors before use. AI was not used for methodological feedback, framework development, or interpretation of results.

Reproducibility Statement

We report the two-stage audit protocol (policy definitions, boundary-case adjudication rules, source hierarchy, and per-stage execution details) in the pipeline description above and in the appendix, so that the curation pipeline can be independently re-run or re-audited. The access date is recorded, but the exact model and generation parameters could not be corroborated against the supplied protocol records (AI Use Statement above). The semi-synthetic error-propagation experiment is parameterized in the text and in Table 4 (iteration count, true ATT, jurisdiction/cohort subset), but its random seed, clustering procedure, exact noise-generating process, and treatment of failed replicates are not available to us and are not asserted here; the calibration sample of ten announcement/effective-date discrepancies is, at best, a post-hoc reconstruction from the audit records rather than a confirmed match to the original simulation code (Limitations). The verified benchmark, the complete audit annotations underlying Tables 2–6, the data dictionary in the appendix, and replication code for the extraction workflow, human verification pipeline, and all inferential analyses are released at our anonymous repository, `https://anonymous.4open.science/r/Digital-Nomad-Research-A953`.

References

J. Bednorz. Working from anywhere? work from here! approaches to attract digital nomads. *Annals of Tourism Research*, 105:103715, 2024.

Irene Botosaru and Federico H. Gutiérrez. Difference-in-differences when the treatment status is observed in only one period. *Journal of Applied Econometrics*, 33(1):73–90, 2018.

Brantly Callaway and Pedro H. C. Sant'Anna. Difference-in-differences with multiple time periods. *Journal of Econometrics*, 225(2):200–230, 2021.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018.

Clément de Chaisemartin and Xavier D'Haultfœuille. Two-way fixed effects estimators with heterogeneous treatment effects. *American Economic Review*, 110(9):2964–2996, 2020.

Augustine Denteh and Desire Kedagni. Misclassification in difference-in-differences models. *arXiv preprint arXiv:2207.11890*, 2022.

Saibo Geng, Hudson Cooper, Michal Moskal, Samuel Jenkins, Josh Berman, Neel Ranchin, Robert West, Eric Horvitz, and Harsha Nori. JSONSchemaBench: A rigorous benchmark of structured outputs for language models. *arXiv preprint arXiv:2501.10868*, 2025.

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

Andrew Goodman-Bacon. Difference-in-differences with variation in treatment timing. *Journal of Econometrics*, 225(2):254–277, 2021.

Him Kabindra KC and Anna Triandafyllidou. Digital nomadism and the emergence of digital nomad visas: What policy objectives do states aim to achieve? *International Migration Review*, 60(1): 71–92, 2025.

Kaisa Koskela and Pascal Beckers. Typologizing digital nomad visas: Comparing policy rationales from tourism to wealth accumulation to immigration. *Journal of Comparative Policy Analysis: Research and Practice*, 26(5):407–425, 2024.

Nikhil Simha Negi. Difference-in-differences with a misclassified treatment. *Journal of Applied Econometrics*, 2025.

Jonathan Roth, Pedro H. C. Sant'Anna, Alyssa Bilinski, and John Poe. What's trending in difference-in-differences? a synthesis of the recent econometrics literature. *Journal of Econometrics*, 235(2):2218–2244, 2023.

Sergey V. Ryazantsev and Nikita S. Ryazantsev. Migration of "digital nomads": Global trends and states' approach. *World Economy and International Relations*, 68(2):116–126, 2024.

José Ignacio Sánchez-Vergara, Marko Orel, and Ignasi Capdevila. Home office is the here and now: Digital nomad visa systems and remote work-focused leisure policies. *World Leisure Journal*, 65 (2):236–255, 2023.

Liyang Sun and Sarah Abraham. Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics*, 225(2):175–199, 2021.

Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. Large language models for data annotation: A survey. *arXiv preprint arXiv:2402.13446*, 2024.

Petter Törnberg. *How to Use Large-Language Models for Text Analysis*. Sage Publications, 2024.

Petter Törnberg. Large language models outperform expert coders and supervised classifiers at annotating political social media messages. *Social Science Computer Review*, 43(6):1181–1195, 2025.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Yifan Zhuang, Zhanghao Wu, Yonghao Zhuang, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 46595–46623, 2023.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. Can large language models transform computational social science? *Computational Linguistics*, 50(1): 237–291, 2024.

A Supplementary Dataset Diagnostics

A.1 Data Dictionary

Table 5 lists every variable in the three linked benchmark tables. The longitudinal panel (2008–2024) is keyed by official country name and ISO code. It contains GDP, unemployment, inflation, internet penetration, and exchange rates as economic controls, and inbound visitor volumes, expenditures, sectoral employment, and tourism's direct GDP share for the tourism sector. The cross-sectional table covers jurisdictions with formal digital nomad visas. It records each visa's launch year, maximum initial duration, and financial thresholds (minimum monthly income and application fees), plus a qualitative assessment of the jurisdiction's tax treatment of remote income. For jurisdictions without a formal digital nomad visa, the qualitative non-adopter table identifies the de facto mechanisms remote workers use, such as specialized talent, investor, or working holiday visas. It also records local hub incentives, regional remote infrastructure programs, and non-resident corporate registration rules.

Table 5: Complete field-by-field data dictionary for the three linked benchmark tables introduced in Table 1. The first block follows World Bank Open Data conventions for macroeconomic and structural fields and UN Tourism series compiled under the International Recommendations for Tourism Statistics 2008 framework for tourism fields; the second and third blocks are cross-sectional policy fields constructed and verified by the two-stage audit described in the main text, not by an external statistical agency. Table names in parentheses match the released files.

Variable	Description
Tourism and Macroeconomic Data (`tourism_and_macroeconomic_data`), sourced from World Bank / UN Tourism	
`country_name` / `iso3`	Official name and ISO 3166-1 alpha-3 code.
`year`	Calendar year of observation spanning 2008–2024.
`gdp` / `unemployment_rate`	Gross domestic product (current USD) and % unemployment of labor force.
`expenditures`	International tourism expenditures from balance of payments.
`tourism_gdp_share` / `employment`	Tourism direct GDP share (%) and sectoral headcounts.
`arrivals_business` / `total`	Inbound business and total tourist arrivals (in thousands).
`internet_usage_pct`	Individuals using the Internet as a % of the population.
`inflation_annual_pct`	Consumer prices measured as an annual percentage change.
`exchange_rate_lcu_per_usd`	Official period average exchange rate (Local Currency Units per USD).
`price_level_index_gdp`	Purchasing power parity price level index of GDP (US = 100).
Cross-Sectional Digital Nomad Policy Data (`policy_data`), sourced from the two-stage audit	
`visa_adoption_year`	Calendar year the formal digital nomad visa program launched.
`coarse_tax_treatment`	Qualitative characterization of the remote income fiscal regime.
`min_income_to_apply_per_month_usd`	Minimum monthly income required to qualify for eligibility (USD).
`visa_duration_months` / `fee_usd`	Maximum initial stay duration (months) and minimum application fee (USD).
Curated Comparison Sample Data (`non_adopters`), sourced from the two-stage audit	
`primary_remote_work_visa`	De facto or primary remote entry mechanism used by nomads.
`supplementary_alternative_visas`	Specialized talent, investor, or working holiday visa frameworks.
`opc_corporate_setup_policy`	Single-member corporate registration rules for non-residents.
`digital_nomad_community_policies`	Localized hub incentives and regional remote infrastructure programs.

A.2 Variable-Level Missingness

Structural macroeconomic controls from the World Bank show near-complete coverage (0.77% missingness for GDP, 9.70% for unemployment, 8.16% for inflation). UN Tourism series are considerably less complete (55.52% for tourism employment, 50.28% for tourism direct GDP share), consistent with the questionnaire-transition pattern described in the main text (Figure 4).

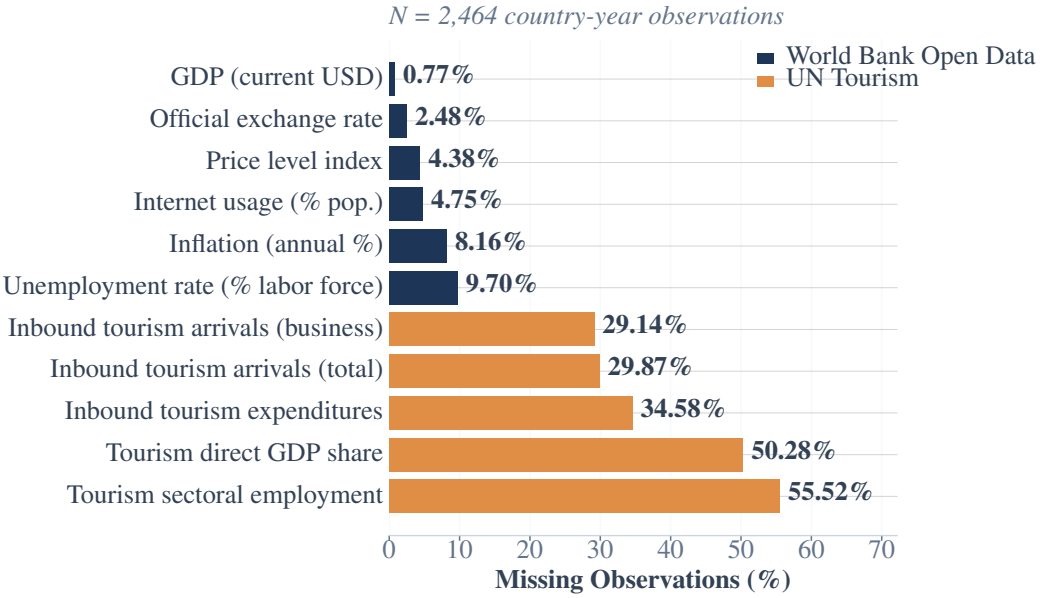


Figure 4: Variable-level missingness across the global macroeconomic and tourism panel ($N = 2{,}464$), by field.

A.3 Tourism Series Consistency Check

Restricting the evaluation sample to countries meeting the minimum-coverage threshold of at least five valid year-over-year observations yields a total of $N = 117$ jurisdictions. Across these countries, the country-level Pearson correlation coefficient distribution between arrivals growth and expenditure growth demonstrates strong positive alignment: the mean country-level correlation is $\bar{r} = 0.553$, the median is $\tilde{r} = 0.670$, and 85.5% of countries (100 out of 117) show a positive correlation. Pooling across all available country-year growth-rate pairs ($n = 1{,}083$) yields a consistent overall pattern, with a pooled correlation of $r = 0.430$ and an ordinary least squares (OLS) regression slope of $\beta = 0.392$, shown alongside the country-level distribution in Figure 5.

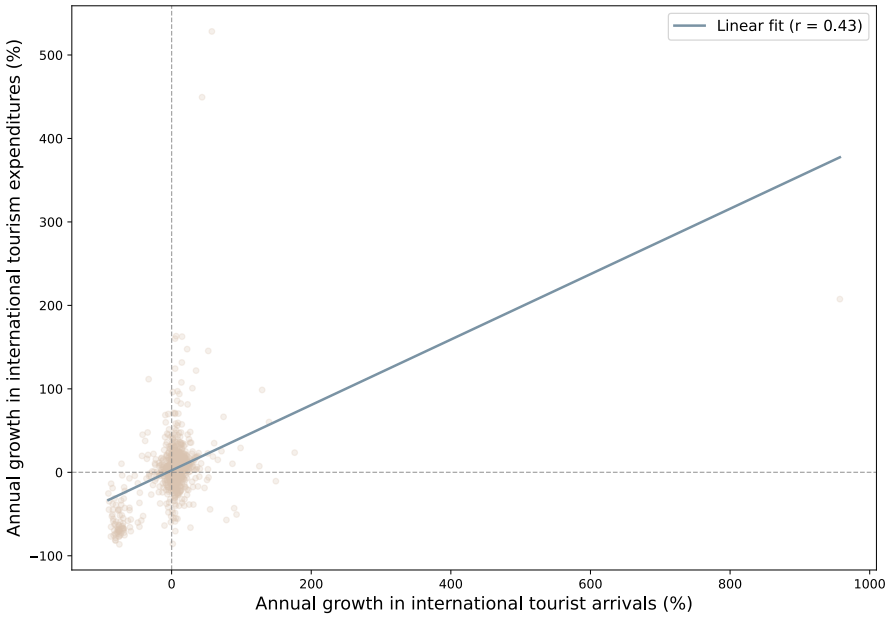


Figure 5: Pooled Country-Year Growth Rates: Tourism Arrivals vs. Expenditures.

A.4 Extended Limitations and Open Items

This subsection gives the finer-grained detail summarized in Section 5.2.

Announced-but-inactive adopters. Of the 33 human-confirmed adopters, 2 (Nepal, the Philippines) carry an "announced but not yet active" status as of the audit's access date rather than an operative program. We retain them in the reported adopter count for transparency, but whether they were included in or excluded from the 15-jurisdiction/3-cohort causal-estimation subset could not be confirmed from the estimation output alone and is flagged as an open item.

Causal-estimation sample exclusions. The causal-estimation sample (30 of the 33 verified adopters) is smaller than the full verified-adopter count because a small number of treated jurisdictions are absent from the macroeconomic panel entirely. We could not independently re-verify the exact excluded set against a supplied macroeconomic panel file and flag this as unconfirmed rather than re-asserting it with full confidence. Similarly, the specific 45 of the 190 panel countries excluded from the 145-country TWFE estimation sample are not individually itemized here.

Measurement interpretation difficulty. Policy coding remains difficult because governments define eligibility differently (a monthly income flow, an annual income figure, or a savings balance, as in Thailand's case; see Appendix B), requiring interpretative harmonization even after full human verification.

Broader stakes. Overstating certainty in the real-data application could mislead governments considering similar programs, researchers building on the released benchmark, and the migrants and remote workers whose eligibility these policies govern. This is part of the motivation for reporting the sample-size and pre-trend caveats in Section 5.2 explicitly instead of omitting them.

Items we could not resolve. The following remain open rather than assumed: the exact number of jurisdictions the semi-synthetic simulation's deterministic misdating rule reaches in a given run, and the exact identities of the treated and control jurisdictions it draws on, since these depend on the underlying audit file, which was not available to us (the simulation mechanism itself is documented in Appendix D); and the model identifier and decoding parameters recorded in the AI Use Statement, which were not corroborated by the audit-protocol records available for this check.

A.5 Descriptive Unemployment Trajectories Around Adoption

Figure 6 shows the descriptive event-time trajectory referenced in Section 4.4: mean national unemployment among adopting jurisdictions by event time relative to first full DNV exposure. This series is shown for context only and is not itself an event-study estimate; the corresponding event-time estimates are in Section 4.4 and Figure 2.

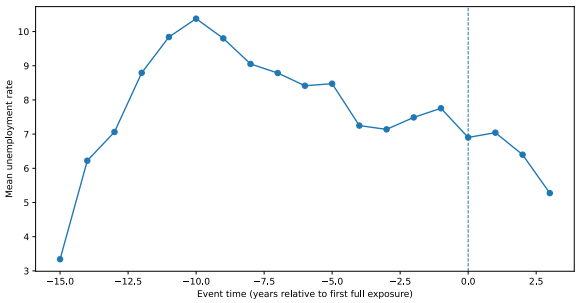


Figure 6: Descriptive unemployment trajectories around digital nomad visa adoption. Mean unemployment rate by event time (years relative to first full exposure); vertical dashed line marks $t = 0$.

A.6 Full Error Taxonomy

Table 6 reports the complete nine-category error taxonomy underlying the summary in the main text, with a description and representative cases for each error type. Rows are grouped into four blocks that reflect where responsibility for each error type lies, following the qualifications discussed in the main text: internal data-quality gaps within the human-audit file, evidentiary heterogeneity that is not itself a defect, a panel-construction issue, and cases that remain unresolved pending further work.

Table 6: Full nine-category error taxonomy underlying the 19 cross-coder disagreements and surfaced label/citation issues discussed in the main text; n is a row count within the relevant audit file (human-audit rows for types 1–6, the 190-panel adoption calls for type 7, and the Stage-1/Stage-2 candidate set for types 8–9). ! Error-prone rows are internal data-quality or panel-construction gaps, detectable and correctable by re-examining the row's own fields; • Partial marks an evidentiary split that is not itself an error; – Unresolved rows require either a primary-source adjudication pass or independent field re-derivation that has not yet been run.

Error type	Description	n	Representative case(s)
Internal data-quality gaps (human-audit file; detectable and correctable) ! Error-prone			
1. Label/citation gap	Coded "primary" but lacks a named legal instrument; discovery source is commercial/unrelated, leaving the tag unsupported.	11	Antigua and Barbuda, Bermuda, Cabo Verde, Dominica, Ecuador...
2. Discovery/verification conflation	Statute named, but discovery source is a commercial blog; the file conflates finding the lead with verifying the actual text.	15	Barbados, Estonia, Romania
3. Third-party source treated as primary	Discovery source is an official body (e.g., third-country embassy, int'l org), but not the target government, yet coded "primary."	7	Antigua and Barbuda, Barbados, Kyrgyz Republic, Montenegro...
4. Secondary derivation inside "primary" row	Legal-basis field admits the statute was found in a secondary source, contradicting the "primary" tag.	1	Portugal
5. Source-label mismatch	Discovery source actively contradicts the primary/secondary tag (likely a coding slip).	2	Seychelles, Bermuda
Evidentiary heterogeneity (not itself a defect) • Partial			
6. Non-adopter evidentiary heterogeneity	"No" calls mix affirmative checks of official portals (no DNV route listed) with simple absences of primary sources in the search results.	131	Afghanistan, Algeria, American Samoa...
Panel-construction issue ! Error-prone			
7. Date heterogeneity and missing values	Date fields mix formats or are missing for adopters, complicating panel construction.	5	Kazakhstan, Brazil, Croatia, Guatemala, Korea Rep.
Unresolved: pending adjudication or re-derivation – Unresolved			
8. Cross-coder disagreement	Human and LLM make different calls due to varying source discoveries or strictness, not internal data defects; no adjudication pass has yet been run.	19	Italy, Kenya, Colombia
9. Incomplete field-level re-derivation (Stage 2)	For many LLM adopters, Stage-2 values carry forward Stage-1 leads without independent re-derivation.	27	Albania, Dominica, Ecuador...

B Case Study: Thailand (Adopter) and Vietnam (Non-Adopter)

To illustrate how the verified benchmark can be read at the jurisdiction level, without introducing new causal claims beyond the main staggered-adoption panel analysis, we compare Thailand, a verified adopter, with Vietnam, which the benchmark's companion non-adopter file records as relying on short-term tourist or business-visa pathways rather than a dedicated remote-work visa category. The comparison of tourism share in GDP and unemployment rate is shown in Figure 7. Both countries share high pre-pandemic tourism dependence and a severe post-2020 contraction in tourism activity: Thailand's tourism-GDP share fell from 7.4% (2019) to a low of 0.84% (2021) before partially recovering to 2.8% (2022); Vietnam's fell from 9.2% (2019) to 2.0% (2021), a comparably sharp contraction from a comparably tourism-dependent baseline.

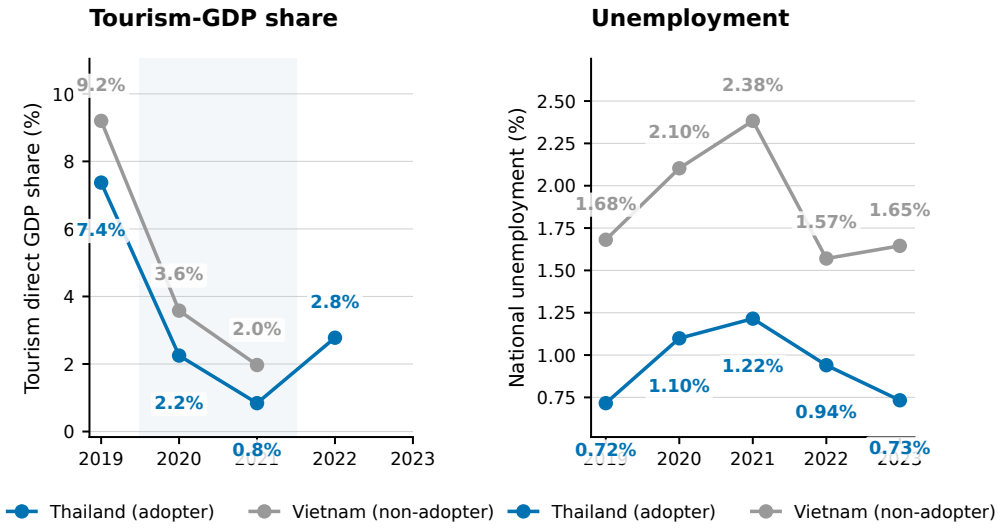


Figure 7: Thailand/Vietnam case study: tourism-GDP share and national unemployment, 2019–2023.

Thailand adopted its Destination Thailand Visa program in 2024. The verified record codes it as a territorial, remittance-based tax treatment, a 60-month duration (the longest in the benchmark) and a minimum fee of THB 10,000 ($\approx$ \$290). Thailand's financial-eligibility requirement is a savings balance rather than a monthly income flow; the verified benchmark leaves the income field blank for this reason, rather than forcing a savings requirement into a monthly-income field as though the two were the same eligibility mechanism. We flag this kind of field-level distinction as consequential for any future moderator analysis of policy generosity, although it cannot bias the pooled estimate reported in the main staggered-adoption analysis.

Vietnam, by contrast, has no dedicated visa or residence category for remote workers within the panel's coverage window. Nomads there rely on tourist visas, an investor visa, or a locally sponsored work permit, alongside a well-documented informal nomad presence in Da Nang, Hoi An, and Ho Chi Minh City. Over 2021–2023, Vietnam's national unemployment rate (1.6–2.4%) ran consistently above Thailand's (0.7–1.2%) even as both economies absorbed comparable tourism shocks. We do not read this gap as evidence of the visa program's effect: the two countries differ on many dimensions beyond visa policy, and this is a two-country comparison, not an identification strategy. The comparison illustrates the kind of jurisdiction-level pattern the verified benchmark is built to support, and is a reminder that Vietnam enters the benchmark only as a non-adopter comparison case, not as part of the verified policy-adopter set.

C Reproducible Protocol for the Two-Stage Audit

This appendix documents the operational protocol underlying the two-stage audit pipeline in enough detail to support independent replication.

C.1 Stage 1: Program-Discovery Phase

Objective. Classify all 190 panel jurisdictions into {Qualifying Policy, No Qualifying Policy, Announced but Inactive Policy, Unresolved Status}, using the eligibility-based DNV definition established in the main text: a legally distinct immigration status or permit whose eligibility criteria explicitly permit remote work for a non-domestic employer or non-domestic clients as a basis for admission or residence, independent of a domestic job offer.

Boundary-case adjudication rules. Applied uniformly across both passes:

1. *Freelancer visas*: included only where eligibility explicitly extends to remote employees of foreign firms, not solely to self-employed providers serving the local market.
2. *Remote-work residence permits*: included whenever the residence basis is remote work itself, regardless of branding.
3. *Entrepreneur/start-up visas*: excluded unless a remote-employment pathway exists within the same instrument.
4. *Tourist-visa pathways permitting incidental remote work*: excluded from the treated set entirely.
5. *Superseded programs*: coded to the original qualifying instrument for the adoption date, with replacement or material eligibility changes logged as policy-attribute changes.
6. *Multiple policy tiers*: coded to the tier establishing first qualifying eligibility, with tier-specific detail retained separately.
7. *Unresolved evidence*: coded Unresolved Status when the record is insufficient to adjudicate any of the above.

Source hierarchy. Applied independently, field-by-field: (1) statute or official gazette; (2) ministry or immigration-authority guidance; (3) consular or embassy pages; (4) secondary/discovery-only sources (law-firm bulletins, mobility-advisory alerts, press coverage), which are used only to locate a lead and are never treated as dispositive on their own.

Execution. Pass 1, a substantive, workbook-wide primary-source adjudication of all 190 jurisdictions, was run as interactive, browsing-enabled LLM-assisted research (OpenAI GPT-5.6 Luna via the ChatGPT interface), source cutoff and primary-source access date 2026-08-09; generation parameters were fixed at Temperature = 1, Top-P = 1, and Seed = null. Pass 2, a structured discovery/trace pass restricted to the 156 jurisdictions left Unresolved after Pass 1, was run against a fixed per-country query template requiring the model to (i) enumerate every plausible secondary hit, (ii) trace each toward a primary legal instrument, and (iii) either upgrade the classification with a cited primary source or leave the case Unresolved. Results are reported in Table 2.

C.2 Stage 2: Field-Level Audit

Objective. For each Stage-1 Qualifying Policy jurisdiction, independently re-derive: adoption/effective date, tax treatment, minimum income threshold, application fee, and visa duration, each from a primary government source.

Logging requirements. For every re-derived cell: the controlling evidence span (quoted or faithfully paraphrased), source title, issuing authority, URL, access date, raw value with original unit/currency, any normalization rule applied, translation method for non-English sources, and a field-level confidence code (High / Medium-High / Medium / Low). Raw values are preserved; a normalized value, where computed, is never substituted for the legal threshold itself. A current-status exception field separately records whether a historically qualifying program remains operative at the observation date: e.g., Antigua and Barbuda is retained as a 2020 adopter even though its programme portal states the Nomad Digital Residency programme has since ended.

Human-audit counterpart. Conducted independently by a single domain-expert researcher across the full 190-jurisdiction panel, with Stage 2 scope limited to the adoption year (not duration, income, fees, or tax treatment). For every jurisdiction the audit recorded: whether a qualifying program was found; program name and type; legal/official basis; announcement and launch/effective dates; status during the observation period; the discovery evidence span and source; and whether that source was primary or secondary. Adopter ("yes") calls require an affirmative primary-source basis; non-adopter ("no") calls default from either an affirmative check of the jurisdiction's own official portal (no DNV route listed) or the absence of any primary source in the search results, and these two "no" subtypes are retained as distinct fields (error type 6).

Reconciliation status. Results are reported in Table 3 and Table 6. As of this release, no adjudication pass has been run against the underlying primary sources for the 19 cross-coder disagreement

cases (error type 8); replication attempts should treat these, and the 27 "field-level verification incomplete" Stage-2 cells (error type 9), as provisional.

D SEMI-SYNTHETIC SIMULATION: IMPLEMENTATION DETAIL

This section records the implementation of the semi-synthetic experiment (Table 4), consistent with the estimation code.

Estimators. *TWFE:* OLS of the outcome on a treatment indicator, country fixed effects, and year fixed effects, with country- and year-lagged log GDP, annual inflation, and internet-usage covariates; standard errors are cluster-robust by country. *CS-style:* an unadjusted group-time difference-in-differences estimator using never-treated countries as the comparison group. For each treated cohort g and post-period $t \geq g$, $\mathrm{ATT}(g,t)$ is the difference between the treated group's and the never-treated group's mean change in outcome from year $g-1$ to year t; this is the simplest never-treated-control version of the Callaway–Sant'Anna group-time estimand and is not presented as a general solution to endogenous treatment timing. Group-time estimates are aggregated to event time using treated-observation-count weights, with the aggregated standard error computed as the observation-weighted root mean square of the per-cohort standard errors, divided by the square root of the number of contributing cohorts.

Synthetic outcome-generating process. For each simulation replicate, a synthetic outcome is constructed per country-year as the sum of: a constant; a country fixed effect drawn $\mathcal{N}(0, 1.5^2)$; a country-specific linear time trend drawn $\mathcal{N}(0, 0.005^2)$ per year elapsed; a common year effect $0.20\sin((\text{year} - \text{year}_{\min})/2)$; $0.20\times$ the country-year's cross-sectionally standardized log-GDP residual; the imposed treatment effect $\beta = -1.0$ for country-years at or after the assigned cohort year; and an idiosyncratic AR(1) error with autoregressive parameter $\phi = 0.45$ and $\mathcal{N}(0, 0.7^2)$ innovations (redrawn fresh each replicate). This construction is shared by both the true-cohort and misdated-cohort runs within a replicate, so the two differ only in which year each treated country is coded as first exposed.

Misdating rule. For every qualifying jurisdiction with a recorded announcement date, the calendar year of that announcement date replaces the verified first-full-exposure year as the assigned treatment cohort; jurisdictions without a recorded announcement date keep their verified cohort. This substitution is deterministic given the audit data (not randomly sampled or applied with a probability), and is identical across all 200 replicates; only the outcome noise differs by replicate. We note one asymmetry in how dates are converted to cohort years: the verified cohort uses a month-aware rule (a January effective date counts as first-full-exposure that same calendar year; any other month rolls to the following year), while the misdated cohort uses the announcement date's raw calendar year with no equivalent month adjustment. Part of the measured attenuation may therefore reflect this date-to-year convention difference rather than the announcement-versus-effective gap alone; we did not have the underlying audit file available to quantify how much.

Inference and replicate handling. Confidence intervals use a normal approximation (estimate $\pm 1.96\times$ SE) at the per-estimate level; coverage is the share of the 200 replicates whose interval contains $\beta = -1.0$. A replicate-estimator combination is included in the summary statistics only if the point estimate is finite; non-finite results (e.g., from an empty comparison cell) are silently excluded from that cell's average rather than counted as a distinct failure category. Random seed: 12345.

Comparison group. Never-treated countries (no qualifying program under the LLM/human-audit determination used to build the analysis panel) serve as the control group in both the true-cohort and misdated-cohort regimes.