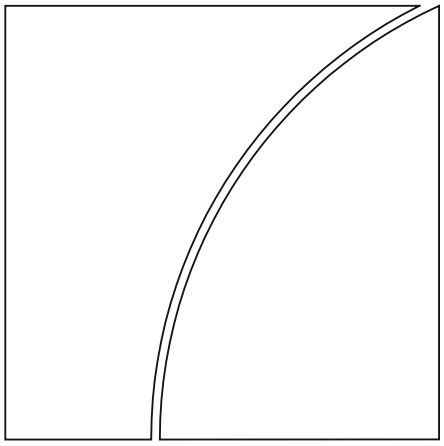


BIS Working Papers

No 1395

MaRCo@BIS: AI-agentic econometric analysis for central bankers

by Marco Lombardi, Taejin Park and Fernando Perez-Cruz

Monetary and Economic Department

November 2026

JEL classification: C32, C63, C87, E37

Keywords: agentic AI, large language models, structural VAR, econometric software

BIS Working Papers are written by members of the Monetary and Economic Department of the Bank for International Settlements, and from time to time by other economists, and are published by the Bank. The papers are on subjects of topical interest and are technical in character. The views expressed in this publication are those of the authors and do not necessarily reflect the views of the BIS or its member central banks.

This publication is available on the BIS website (www.bis.org).

© *Bank for International Settlements 2026. All rights reserved. Brief excerpts may be reproduced or translated provided the source is stated.*

ISSN 1020-0959 (print)
ISSN 1682-7678 (online)

MaRCo@BIS: AI-agentic econometric analysis for central bankers

Marco Lombardi · Taejin Park · Fernando Perez-Cruz

November 2026[1]

Abstract

Over the past half-century, econometric software has greatly reduced the burden of implementation, while leaving substantive choices about model specification, identification and sample largely to the researcher. Agentic AI can propose specifications and explain their economic rationale, while earlier automated selection methods could simply searched across candidate specifications using statistical criteria. This paper introduces MaRCo@BIS, the MAcroeconometric Research Companion: a framework of plain-text workflow files, a verified dictionary of canonical model specifications and a tested reference library, read by an off-the-shelf agentic coding environment. MaRCo@BIS proposes specifications and provides their rationale, retrieves and harmonises data, estimates, checks results against economic priors and logs every choice for replication, while the researcher confirms or overrides each (consequential) decision. We document the design, benchmark the agent's replications of three widely-used structural VARs against the originals, and illustrate the framework across five exercises ranging from a verified replication to an open design question. We also discuss the necessity of human supervision, the risks of specification search and automation bias, and the safeguards against these risks. The framework is released as a public good for the central banking community and macroeconomic researchers more broadly.

JEL classification: C32, C63, C87, E37.
Keywords: agentic AI; large language models; structural VAR; econometric software.

1 Introduction

For half a century, every advance in econometric software has automated the typing, rather than the thinking. Primitive punch cards and low-level languages such as Fortran gave way to specialised environments such as RATS, EViews, Matlab and Stata, and then to open-source languages such as R and Python. Each step relaxed a binding constraint. Computation grew faster, languages became friendlier, and estimation routines migrated from proprietary packages to flexible open libraries, each step cutting the time spent on technical implementation.

Choosing the right tool has remained the researcher's responsibility. Software packages estimate only what they are told to estimate; they do not help with the more substantive calls, e.g. which variables, which prior (how strongly estimates are pulled toward a conservative benchmark), which identification scheme (the assumptions that turn correlations into statements about cause and effect), which sample. The estimation layer itself is well served: BEAR (Dieppe et al (2016)), the BVAR_ toolbox (Canova and Ferroni (2021)), the VAR-Toolbox (Cesa-Bianchi (2026)), Dynare (Adjemian et al (2011)) and statsmodels (Seabold and Perktold (2010)) together cover most of the modern macroeconometric toolkit. But none of these tools reaches the layer where it is decided what to estimate. And because research questions can typically be approached in more than one way, say structural vector autoregressions (SVARs) vs local projections, answering one has typically meant switching between different software environments and

[1]Bank for International Settlements. Lombardi: Marco.Lombardi@bis.org; Park: Taejin.Park@bis.org; Perez-Cruz: Fernando.Perez-Cruz@bis.org. We thank Iñaki Aldasoro, Jon Frost, Byeungchun Kwon and Frank Smets for helpful comments. The views expressed in this publication are those of the authors and do not necessarily reflect the views of the BIS or its member central banks. The authors acknowledge the use of artificial intelligence in research assistance and editorial refinement. Responsibility for all errors remains with the authors. References to individual firms and projects are for illustrative purposes and should not be construed as any statement on the soundness or the legal or regulatory status of any firm or project.

languages, reconciling conventions, as well as manually adapting replication codes. Agentic artificial intelligence (AI) coding tools simplify this work considerably. They can adapt and translate code from one language to another, harmonise conventions and fold different estimation procedures into the same environment.

The adjective "agentic" matters here. Unlike a chatbot that answers a single prompt, an agentic tool is a large language model (LLM) set up to act, plan and execute multi-step work. It writes and runs code, reads documents, checks its intermediate output against its instructions, and pauses at defined checkpoints for direction. These instructions can be supplied as a "skill", a plain-text workflow manual that the model loads and follows. Evidence that generative AI raises the productivity of knowledge work is accumulating (Noy and Zhang (2023); Dell'Acqua et al (2023)), and Korinek (2023) catalogues its uses across the research pipeline. What that literature documents is, for the most part, assistance with individual tasks (drafting, coding, summarising), with the researcher still assembling the pieces.

But agentic AI can also reach beyond estimation, up to the specification layer. Earlier automated selection methods could at most use statistical criteria to choose among candidate specifications; an agentic tool can also explain the economic reasoning behind its proposals. Instead of merely simplifying the estimation of pre-determined models, it can integrate reasoning into the workflow that follows from a research question: proposing specifications based on the literature, checking results against theory, diagnosing issues (eg a price puzzle) and suggesting remedies, adapting models to specific economies taking into account differences in available data, all while documenting decisions for replication and review. For central banks, the case is clear. Scarce econometric expertise can focus on judgement and reasoning instead of mastering the different mechanics.

These capabilities need to be tested before an agent is entrusted with new analysis. We begin by examining whether it can reproduce published results, a task that has also proved difficult in research conducted by humans (Chang and Li (2015); Christensen and Miguel (2018); Vilhuber (2020)). The comparison requires care: the agreement we can expect depends on how the model is identified and whether the original data are available. Our benchmark in Figure 4 and Section 4.4 examines these differences and what they reveal about the agent's performance.

This paper introduces MaRCo@BIS, the MAcroeconometric Research Companion, a framework released as a public good for the central banking community, distributed as a self-contained package together with the paper (Section 7). The paper makes three contributions. First, for the applied economist, a working framework that removes the mechanical layer of econometric practice (writing and adapting estimation code, reconciling conventions across languages and packages, assembling and harmonising data) so that scarce econometric expertise concentrates on judgement: which question, which assumptions, whether the results make economic sense. As of the release date, the package contains a tested reference library of 24 estimation modules, 12 workflow skills, a dictionary of 30 canonical model specifications verified against the original papers, and a data layer with runtime connectors. The five exercises of Section 4 are this contribution's evidence, spanning the range from verified replication to open design question. Second, for developers of agentic research tools, we provide a design that combines modular plain-text skills, verified knowledge and tested code in a staged workflow. Its verification approach distinguishes between fixed components, which can be audited independently, and decisions requiring judgement, which are subject to human confirmation. Third, for the central banking community, we provide benchmark evidence on the capabilities and limitations of agentic AI, an account of where human judgement remains necessary, and a public release for others to test and develop. These contributions are intended to support debate and further assessment of the framework.

The remainder of the paper proceeds as follows. Section 2 relates MaRCo@BIS to two strands of related work, econometric toolkits and LLM-based agents for economic research, and sets out its distinctive focus on governed macroeconometric specification. Section 3 presents the framework: the principles it is built on, the workflow, and the model dictionary with its verification. Section 4 presents five exercises, examines replication using both stored model specifications and a supplied paper, and benchmarks the results against

the published originals. Section 5 follows an open research question from specification to results, bringing the framework's capabilities together in one session. Section 6 draws the lessons of building and using it, and states the limitations together with the safeguards against them. Section 7 describes the released package, and Section 8 concludes with the questions the community should now take up.

2 Related literature

The paper relates to two strands of work. The first concerns the development of software for applied econometric research. Toolkits and libraries such as BEAR (Dieppe et al (2016)), the BVAR_ toolbox (Canova and Ferroni (2021)), the VAR-Toolbox (Cesa-Bianchi (2026)), Dynare (Adjemian et al (2011)) and statsmodels (Seabold and Perktold (2010)) supply tested implementations of the modern macroeconometric toolkit, from estimation and identification menus to filters and forecast evaluation, and they anchor the reproducibility of applied work. Their common boundary is the one Section 1 described. The researcher supplies the specification, and the software estimates it. The exception is the automatic model selection programme of Hendry and Doornik (2014), which distils two decades of general-to-specific ("GETS") search (Hoover and Perez (1999); Hendry and Krolzig (1999)) into the Autometrics algorithm (Doornik (2009)). There, the researcher writes down a general unrestricted model nesting every candidate explanation; a multi-path search then deletes what fails statistical significance, subject to mis-specification and encompassing tests, with indicator saturation absorbing outliers and breaks, and the machinery extends even to more candidate variables than observations. Two features define that tradition. First, the algorithm selects specifications using statistical criteria. Economic reasoning enters through the researcher's formulation of the general model and the choice of variables to retain regardless of their statistical significance. Second, the statistical costs of specification search are explicitly analysed and found to be small in the settings studied.

The second strand applies LLM-based agents to economic research and structured modelling. Korinek (2025) provides a broad practical account of agents that plan and execute multi-step tasks including literature review, data retrieval and econometric coding. Dawid et al (2025) propose an end-to-end workflow spanning ideation, modelling, empirical analysis and interpretation, coordinated through specialised agents and human checkpoints. Afonso et al (2026) develop a more tightly governed application to dataset construction, separating research design from implementation and imposing explicit rules for evidence, provenance and uncertainty. Two contemporaneous systems come closer to the model-selection layer. The AgentEconomist of Chen et al (2026) translates economic intuitions into executable computational experiments using a domain-specific literature base, while the AI Economist Agent of Kato (2026) combines evidence retrieval, model selection, explicit model execution and traceable report generation. Together, these contributions establish that agents can organise substantial parts of the economic research process, while also highlighting the importance of structured workflows and human oversight.

MaRCo@BIS draws on both strands and fills the gap between them. From the first it keeps the discipline of a tested library. The estimation routines are implemented once for reuse across analyses and checked through automated tests. From the second it takes the agentic capabilities, but differs in object and verification strategy. It focuses specifically on applied macroeconometric specification and assigns the agent a deliberately bounded role over a verified dictionary of published models and a tested estimation library. Its contribution is therefore not agentic economic research in itself, but an architecture for governing it: fixed and independently auditable machinery wherever an answer can be checked, and explicit confirmation, logging and disclosure wherever econometric judgement cannot be verified in advance. Relative to the GETS tradition, the roles are reversed: Autometrics selects statistically within an economically specified universe, while MaRCo@BIS proposes the universe itself in economic terms (the variables, identification, sample and prior that GETS takes as given) and bars statistical criteria from those choices, confining data-driven selection to tuning parameters. Structural identification, outside the remit of reduced-form discovery, is the framework's main focus. And where Hendry and Doornik (2014)

discipline specification search by quantifying its statistical cost, MaRCo@BIS disciplines it procedurally, with gates, logging and disclosed revision counts. The two approaches are complementary: an agentic workflow could use GETS for tuning choices, such as lag selection, within a specification agreed with the researcher.

3 How MaRCo@BIS works

3.1 Design principles

First, the framework's defining behaviour is a guided, research-led workflow. At every consequential decision (the model class, the variable set, the identification scheme, the sample, or a data substitution), MaRCo@BIS proposes specific choices with their rationale and waits for the researcher to confirm or override; it does not decide silently. The agent contributes by proposing specifications grounded in the literature and remedies for anomalies. The researcher then assesses these proposals and decides whether to proceed. Both proposal and confirmation are necessary for the intended division of work.

Second, economic sensibility takes priority over statistical fit. A choice that is statistically convenient but economically implausible is flagged, and estimated results are screened against theory before they are accepted. The rule exists because statistical criteria alone will rationalise almost any specification in typical macroeconomic samples; economic logic is the one constraint that cannot be delegated to fit statistics. Anomalies are flagged, and none is accepted without the researcher seeing it.

Third, the workflow is acyclic. Any decision can be revised at any point, but revisions cascade forward. The framework re-runs the affected stages instead of restarting the loop, each revision is logged as a new version, and parallel specifications (eg alternative priors or lag orders, a BVAR alongside local projections) can be brought forward and compared side by side. Avoiding full restarts keeps long sessions coherent and preserves a readable record of what changed and why.

Fourth, the workflow is transparent. Every proposal names the knowledge source behind it (the sources have a fixed priority order, set out with the specification step below), and every decision enters the research log with its rationale, the alternatives considered and the human confirmation. Any run can therefore be audited and replicated after the fact.

Fifth, the framework divides labour deliberately between the stochastic and the deterministic. The stochastic AI handles tasks requiring judgement (economic reasoning, proposal, adaptation). Everything that can be fixed is fixed: canonical specifications are recorded in the dictionary, estimation code is maintained in the tested library, and the conduct of each workflow stage (what to propose, what to check, when to stop and ask) is set out in written instructions that the agent follows. Each time a task moved out of the model's hands into a fixed component, its errors became easier to find and fix. Recalling a paper's specification from memory gave way to reading a verified dictionary entry, and writing estimation code afresh to calling a tested library function. A corollary is that the framework is more general across AI models. There is no server and no bespoke model. The accumulated value is stored in plain-text files and code that the model merely reads, so an AI model upgrade or replacement calls for re-validation, probably without a complete rebuild.

3.2 The workflow

The framework operates as a single end-to-end environment. The researcher poses a question in natural language, confirms key decisions in each step, and receives estimated results, diagnostic output and a research log of the choices made along the way. The framework is a set of plain-text instruction files (the skills introduced in Section 1) and reference code, read by an off-the-shelf agentic coding environment that provides the reasoning. The files form three layers. The *workflow* layer sets the process: the staged loop,

the decision protocols and the points at which the agent must stop and ask. The *model dictionary* supplies the specifications: a set of canonical papers, each with its own variables, identification, sample and priors, paired with executable replications, with results verified against the original. The *reference library* does the computing: a single, tested implementation of each estimation method. The current release holds 24 modules spanning VARs, BVARs and local projections; recursive, long-run, sign, narrative, proxy and heteroskedasticity-based identification; time-varying-parameter, mixed-frequency, factor and panel models; filters, connectedness and forecast evaluation. The agent proposes suitable econometric models based on the user's prompt. Once the user has selected a model, its application follows the process prescribed in the skill files.

A session begins with triage into one of three entry modes: replicating a paper already in the dictionary (mode A); ingesting a new paper, with the specification extracted from the document, verified and then executed (mode B); or designing a specification from scratch for an open research question (mode C). All three converge on the same staged loop shown in Figure 1: specification, data, estimation, a check of economic sensibility, reporting, and a final disposition of what the run produced. Triage is itself the first gate: the agent states the mode it has chosen and why, so a wrong guess can be corrected before any work is done.

Specification is the stage where the framework differs most from earlier tools, and its choices draw on an explicit hierarchy of knowledge sources. The researcher's instructions rank first and always prevail. The model dictionary follows, supplying verified anchors for questions the literature has already found convincing ways to answer.[2] The economy-specific knowledge base ranks next, then the AI model's own knowledge, used to adapt, combine and fill gaps. Finally, the tuning choices, such as lag order and prior hyperparameters, are settled by statistical fit. Lower-ranked sources can identify limitations in a specification drawn from a higher-ranked source and suggest adjustments, but any departure requires the researcher's confirmation. If a dictionary entry specifies seven variables with four lags but the local sample holds only 80 observations, the model's own reasoning flags the parameter count and proposes fewer variables or a tighter prior. Choices at every level must have an economic rationale.

Two rules govern specification choices. Variables and identification are never chosen by statistical fit. Data-driven criteria settle tuning parameters only, so the substantive content of a specification always has an economic or literature-based rationale. And specification search is disclosed: the final report states the number of sensibility-driven revisions, shows first-pass results alongside final ones whenever a revision occurred, and, for research integrity, notes that the reported bands are conditional on that search.

The workflow incorporates data collection and harmonisation, traditionally preparatory tasks that often account for a significant share of a project's effort. Pre-collected standard panels[3] cover common applications; runtime connectors retrieve additional series from public application programming interfaces (APIs); and frequency alignment, transformations, and vintage tracking are logged like any other decision. The economy-specific knowledge base makes the workflow portable. It records structural features (the exchange rate regime, the policy instrument, the principal transmission channels, the local data sources) and conditions the application of the other knowledge sources. Run with a different knowledge base, the same workflow should, in principle, produce a specification suited to that economy without code changes.

Data sources follow their own priority ladder, mirroring the specification hierarchy: pre-collected verified panels first; the knowledge base's local sources next; then the runtime connectors; and finally user-staged files for series available only from author archives, where the rule is to stop and ask, never to substitute or fabricate. Substitution, when it arises, is a gated proposal like any other. Sims's (1980) import-price series has no exact counterpart from the FRED database, so the framework proposes the closest substitute, explains the discrepancy and lets the researcher decide. Every fetched series is saved with its provenance

[2]The dictionary currently holds 30 canonical specifications; Section 3.3 details what an entry records, where the entries come from and how they are verified.

[3]These include the monthly and quarterly FRED-MD and FRED-QD macroeconomic datasets (McCracken and Ng (2016, 2021)).

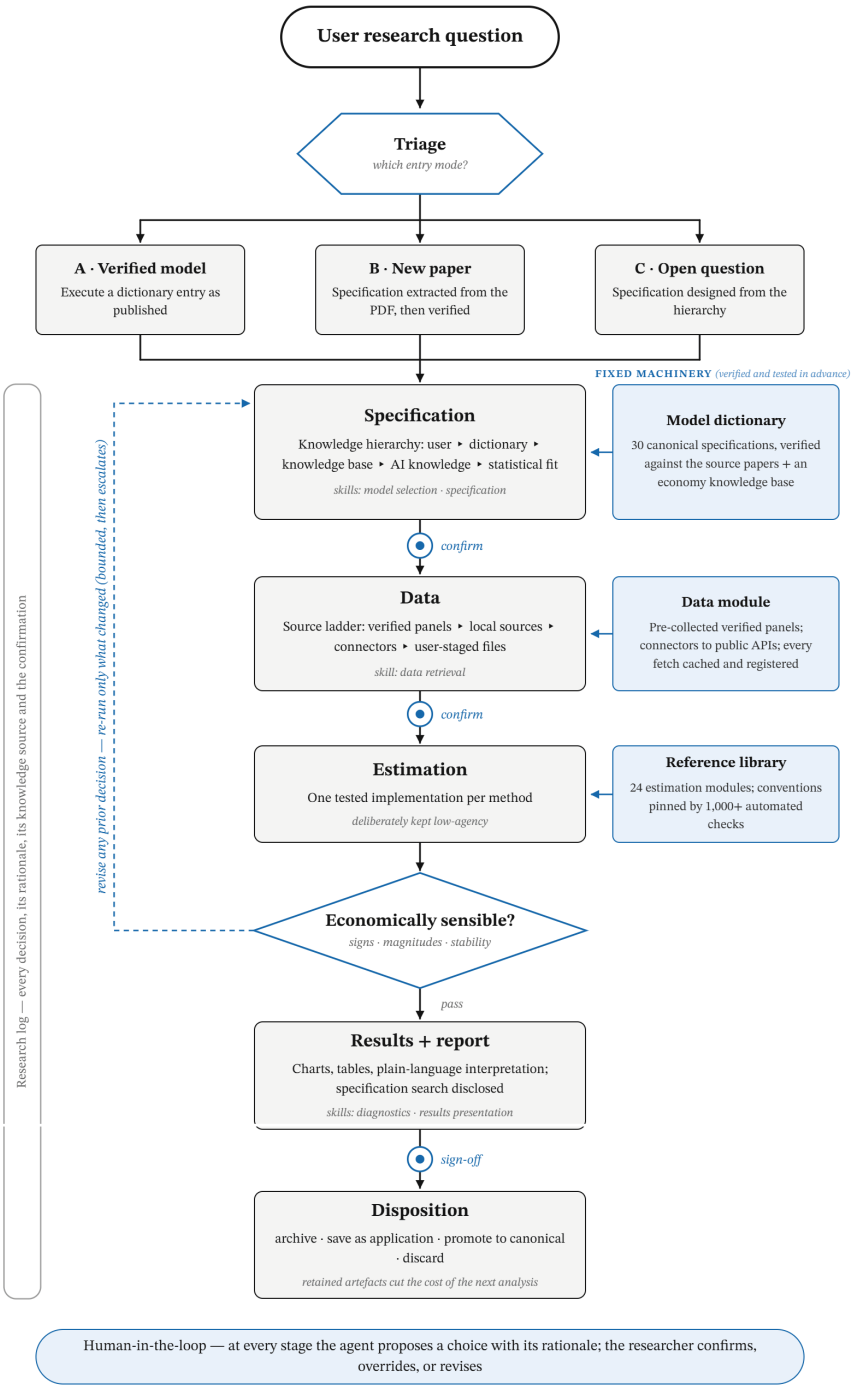


Figure 1. The framework's workflow at a glance

Notes: The centre spine is the agentic process: the stages in the order of Section 3.2, with a confirmation gate at every consequential decision. The right rail is the fixed machinery the process consults, verified and tested in advance (Section 3.3). Skills are noted at the stage where they act; three further skills support the loop throughout (benchmarking, advisor protocol, consistency audit). The dashed return path carries both revision channels: the sensibility check's automatic revisions, which are bounded and then escalate, and the researcher's own, available for any prior decision at any time. Based on Claude Code (Anthropic), running Claude Opus 4.8 and Fable 5, accessed May–July 2026.
Source: Authors' elaboration.

(source, vintage, fetch date, transformation) and registered for reuse. Later projects can load it without a new API call, making these data an institutional asset alongside retained analyses. Each first-time acquisition accordingly ends with its own disposition choice, the data-layer mirror of the loop's final stage.

Estimation, by contrast, is deliberately a less agentic stage. Every estimation method the framework offers is implemented once in the reference library. An analysis that needs sign restrictions calls the library's sign-restriction routine, and in the reference implementation the agent is required to call the library instead of writing estimation code of its own (when the user requests another language, the agent generates code against the library's documented conventions, and the rule remains to reuse a tested implementation, never to reimplement a standard algorithm). Two consequences follow. The same specification returns the same numbers across sessions, because the code that runs is the same code every time. And the test suite described in Section 3.3 applies to every analysis at once, because testing the single implementation tests the code they all share. Defaults come from a single configuration file; the agent's role is to assemble the call, not to improvise the econometrics.

After estimation, a sensibility loop checks the results against economic priors: whether the signs of the impulse responses are plausible, the magnitudes reasonable and the system stable. A failed check is diagnosed against known pathologies (eg a price puzzle, explosive dynamics) and matched to remedies from the literature. Not every flagged issue interrupts the researcher. Critical failures, such as an explosive root or severe autocorrelation, block and arrive with a proposed remedy; soft warnings, such as mild non-normality, are noted with a recommendation to proceed. Both enter the log either way. Revision is bounded: after a small pre-defined number of iterations, the framework stops and asks the researcher.

Reporting turns the estimates into charts, tables and a plain-language interpretation, with uncertainty and caveats shown in full and any specification search disclosed. Alongside the report, the research log is completed: every decision with its rationale, the alternatives considered, the knowledge source and the confirmation, plus a durable record of where the run stands, so that a session interrupted midway resumes instead of restarts.

Each analysis ends with a disposition step: the researcher chooses whether the work is archived, saved as a reusable application, promoted into the canonical library or discarded, and the choice is logged. Work retained this way leaves behind reusable assets (an indexed specification, a data connection, vetted estimation code, a log entry), so the marginal cost of related work is likely to fall with each exercise.

The stages share a common structure (a priority order over sources, a proposal with its rationale, a confirmation gate, a log entry and a disposition), which keeps the workflow legible and auditable. Twelve skills implement it (Table 1); each stage's full protocol ships with the framework's documentation.

Table 1. List of skills

Skill	Role
Orchestration	
Orchestrator	Drives the seven-stage, human-in-the-loop workflow from research question to final answer, proposing a specific choice with its rationale and asking for confirmation at every decision point (model class, variables, lags, priors, identification, sample, disposition).
Knowledge layer	
Model dictionary	Finds and develops model specifications: it matches the research question to a pre-verified entry (a paper's variables, identification, sample and prior) that the agent adapts without re-reading the source; drafts and verifies a new entry when the user brings a paper the registry does not yet cover.
Knowledge base	Holds institution- and economy-specific context (local data sources, conventions, regime dates, user-supplied papers) and governs when this local knowledge overrides the agent's general knowledge; the other skills consult it before falling back on the model's general knowledge.
Analytical workflow	
Model selection	Proposes appropriate models for an open research question, providing rationale and trade-offs, and gives a second opinion on a model the user has already chosen.
Specification	Settles variables, lag order, identification, sample and prior hyperparameters together, applying the framework's priority order with economic logic as the binding constraint.
Data retrieval	Retrieves data, preferring pre-collected verified panels and falling back to documented runtime sources or user-staged files, with registry checks, quality checks and transformations recorded.
Diagnostics	Runs residual and system checks (eg autocorrelation, heteroskedasticity, normality, eigenvalue stability, structural breaks), flags problems with model adequacy and proposes remedies.
Results presentation	Turns estimates into charts, tables and a plain-language interpretation, with uncertainty and caveats shown in full and any specification search disclosed.
Research log	Records every decision with its rationale, the alternatives considered, the human confirmation and the knowledge source that informed it, producing a timestamped, reproducible audit trail and the basis for a replication appendix.
Quality, escalation and maintenance	
Benchmarking	Evaluates any estimation method against baseline models, competitors and published results for forecast accuracy, economic sensibility and numerical correctness.
Advisor protocol	Escalates dense or unfamiliar mathematics (novel identification, non-standard prior derivations) to a more capable, separately configured reasoning model (which can be institution-hosted) when the executing agent's confidence is low.
Consistency audit	An on-demand maintenance pass, paired with an automated mechanical check, that audits the three layers (skills, library, dictionary) for drift and documentation-versus-behaviour mismatch.

Notes: Each skill may encompass several sub-skills. For instance, the data retrieval skill could create multiple sub-skills, each corresponding to a specific data source such as BIS, FRED, ECB, IMF, OECD, etc.
Source: Authors' elaboration.

```yaml
# Excerpt; full entry: skills/model-dictionary/entries/litterman_1986_bvar.yaml
id: litterman_1986_bvar
citation: {authors: Litterman, year: 1986, pages: 25-38,
           title: "Forecasting with Bayesian Vector Autoregressions—Five Years of Experience"}

model:
  class: bvar
  prior: minnesota
  variables:                                  # paper p. 29, in the paper's order
    endogenous: [real GNP, GNP deflator, unemployment rate, M1,
                 gross private domestic investment, commercial paper rate (4-6 month),
                 change in business inventories]
  lags: 6
  frequency: quarterly
  hyperparameters:
    lambda_1: 0.2     # overall tightness lambda; {.5, .3, .2, .1} tried, .2 adopted (p. 32)
    lambda_2: 0.2     # cross-variable weight theta = .2 (p. 31)
    lambda_3: 1       # harmonic lag decay: own-lag-j std = lambda/j
    delta_i: "1 (random-walk prior mean on own first lag) for ALL seven variables (paper p. 29)"

identification: {method: none, rationale: "paper focuses on forecasting, not structural analysis"}

sample: {start: 1948Q1, end: 1979Q3}                                  # paper p. 29

expected_results:
  forecasting:
    vs_unrestricted_var: BVAR substantially outperforms, especially at longer horizons
    vs_univariate: BVAR outperforms AR models for most variables
    vs_random_walk: RMSFE ratios < 1.0 for output and prices

verified:
  status: checked_against_paper
  date: 2026-06-03
  pdf_location: papers/litterman_1986_bvar.pdf
```

Figure 2. What a dictionary entry looks like

Notes: The entry shown is the Bayesian VAR with Minnesota prior of Litterman (1986).
Sources: Litterman (1986); authors' elaboration.

3.3 The model dictionary

The dictionary is the layer users are likely to work with most. An entry records what an economist would need to rebuild a paper's baseline: the variables and their transformations, the identification scheme with its restrictions and their rationale, the sample, the prior and its hyperparameters, and the results the paper reports, the expected answers against which a replication is judged. Figure 2 shows the entry for Litterman (1986); Table 2 lists all 30, spanning the toolkit surveyed in Ramey (2016) and Kilian and Lütkepohl (2017), from Sims (1980) to sign restrictions sharpened by higher moments.

Each entry records our reading of the source paper's variables, identification, sample and priors, with page references to support checking. We checked the entries against the papers and documented identified departures from the published specifications, although errors or omissions may remain. The entries were prepared in several ways: most were drafted directly from the original papers; a number were based initially on the worked examples in Canova and Ferroni's (2021) toolbox and then checked field by field against the papers themselves; others were prepared through the mode B pipeline or from the authors' replication packages.

In use, an entry serves one of two purposes. Mode A runs its paired replication. Mode C treats it as a verified starting point that the agent adapts to the user's economy, sample and question, under the hierarchy of Section 3.2. Before either, the user can check the entry on their own machine. Each entry comes with a runnable replication, and one command re-runs the canonical specification and compares the output with the numbers the paper published or, where the paper reports no directly comparable number, with the authors' own results from the same replication. That one run checks the method's implementation, the data

Table 2. Canonical specifications in the model dictionary

	Paper	Journal	Model class	Identification or method
Bayesian VARs: forecasting, priors and estimation	Litterman (1986)	*J. Bus. Econ. Stat.*	Bayesian VAR (BVAR)	Minnesota prior; forecasting
	Bańbura, Giannone and Reichlin (2010)	*J. Appl. Econom.*	Large BVAR	shrinkage scaled with system size
	Bańbura, Giannone and Lenza (2015)	*Int. J. Forecast*	BVAR	conditional forecasts
	Giannone, Lenza and Primiceri (2015)	*Rev. Econ. Stat.*	BVAR	hierarchical prior selection
	Lenza and Primiceri (2022)	*J. Appl. Econom.*	BVAR	pandemic volatility scaling
Structural identification	Sims (1980)	*Econometrica*	Structural VAR (SVAR)	recursive (Cholesky)
	Blanchard and Quah (1989)	*AER*	SVAR	long-run restrictions
	Christiano, Eichenbaum and Evans (1996)	*Rev. Econ. Stat.*	SVAR	recursive; flow-of-funds responses
	Christiano, Eichenbaum and Evans (1999)	*Handbook of Macroeconomics*	SVAR	recursive
	Galí (1999)	*AER*	SVAR	long-run restrictions; technology shocks
	Rigobon (2003)	*Rev. Econ. Stat.*	SVAR	heteroskedasticity across regimes
	Romer and Romer (2004)	*AER*	shock series	narrative; intended rate changes purged of Fed forecasts
	Uhlig (2005)	*JME*	SVAR	sign restrictions
	Kilian (2009)	*AER*	SVAR	recursive (Cholesky)
	Mertens and Ravn (2013)	*AER*	proxy SVAR	narrative tax instruments
	Gertler and Karadi (2015)	*AEJ: Macro*	proxy SVAR	high-frequency external instrument
	Lanne, Meitz and Saikkonen (2017)	*J. Econom.*	SVAR	non-Gaussian maximum likelihood
	Antolín-Díaz and Rubio-Ramírez (2018)	*AER*	SVAR	narrative sign restrictions
	Andrade, Ferroni and Melosi (2023)	*Fed Chicago WP*	SVAR	Sign restrictions sharpened by higher moments
Local projections and nonlinear models	Jordà (2005)	*AER*	local projections (LP)	impulse responses by direct regression
	Auerbach and Gorodnichenko (2012)	*AEJ: Econ. Policy*	smooth-transition VAR	state-dependent fiscal multipliers
	Stock and Watson (2018)	*EJ*	LP-IV	external instruments
Factor, panel and mixed-frequency models	Bernanke, Boivin and Eliasz (2005)	*QJE*	FAVAR	recursive, with latent factors
	Canova and Ciccarelli (2013)	*AIE*	panel VAR	survey; pooled and hierarchical estimators
	Bańbura and Modugno (2014)	*J. Appl. Econom.*	dynamic factor model	maximum likelihood with missing data
	Schorfheide and Song (2015)	*J. Bus. Econ. Stat.*	mixed-frequency VAR	Bayesian state space
Time variation, spillovers and trend-cycle	Primiceri (2005)	*REStud*	TVP-VAR	stochastic volatility, time varying parameters
	Diebold and Yilmaz (2012)	*Int. J. Forecast*	generalised VAR	generalised variance decomposition; connectedness
	Hamilton (2018)	*Rev. Econ. Stat.*	Hamilton filter	trend-cycle decomposition
	Kamber, Morley and Wong (2018)	*Rev. Econ. Stat.*	Beveridge-Nelson filter	trend-cycle decomposition

pipeline and the local environment at once. The replication should be run before any adaptation, because the published numbers vouch for the canonical specification only; once the specification is adapted, its quality is governed by the workflow's gates instead. The dictionary can be extended locally through mode B ingestions, with new entries subject to the same audit standard.

Verification spans the whole framework. The fixed components are verified before release. The reference library's mathematical conventions are enforced by an extensive test suite;[4] each replication validates its output against figures and tables in the source paper or, where the paper reports no directly comparable number, against the authors' own results from the same replication; each dictionary entry has been checked against that paper; and a behavioural checklist specifies tests of the workflow itself, from mode triage to anomaly flagging. This includes discouraging requests to iterate specifications until results turn significant. A novel analysis has no known answer against which it can be verified. It is therefore governed by the confirmation gates, sensibility checks and disclosure rules already described.

The estimation library is also cross-validated against independent implementations: continuously against statsmodels inside the test suite, and during development against two MATLAB toolboxes, BVAR_ and the VAR-Toolbox, covering estimation, identification, impulse responses and both decompositions.

4 MaRCo@BIS put to work

4.1 Five exercises, one framework

We illustrate MaRCo@BIS with the kind of short request an applied economist might type. The five exercises in Table 3 all run through the same workflow of Figure 1; what changes from one exercise to the next is how much verified structure the framework can start from, ranging from a complete dictionary entry with data supplied for replication down to nothing but a question. Three runs in mode A are anchored on Uhlig (2005): the verified entry (A-1), data that the original specification cannot handle (A-2) and the application to another economy (A-3). A new paper ingestion (B) and an open design question (C) complete the set. In each case, the framework proposes a specification and raises the judgement calls that the instruction leaves open, while every key decision stays with the user.

The first exercise asks a question with a known answer: whether a monetary tightening reduces output, which Uhlig (2005) answered agnostically. A-1 simply instructs the framework to reproduce its results. Triage matched the request to the verified dictionary entry; the only decision the framework raised was data sourcing, and the analysis ran end-to-end from the entry's variables, identification, sample and priors.

A-2 asks the same economic question on data the original specification cannot handle. Extending Uhlig's sample to the present runs into the effective lower bound, non-borrowed reserves turning negative and pandemic outliers. The framework raised each obstacle itself, proposing the literature's remedies: the Wu and Xia (2016) shadow rate, total in place of non-borrowed reserves, and pandemic dummies or the Lenza and Primiceri (2022) volatility scaling. Every substitution passed through a confirmation gate before estimation.

A-3 asks whether the same identification travels to another economy. For the United Kingdom, the framework flagged breaks in the monetary policy regime, the absence of a domestic commodity-price index for the forward-looking supply role, and the lack of clean counterparts to the reserves variables; it sourced the series itself from the ONS and the Bank of England, and proposed a post-1997 sample (the era of Bank of England operational independence), the Krippner (2015) UK shadow rate, PPI input prices as the commodity proxy, and dropping the reserves variables. At the data stage the sample shortened further,

[4] Each method is checked against an independent, widely used implementation where one exists and must return identical results on re-runs; parameter recovery on simulated data confirms the estimator returns the true values, and coverage studies for recursive VAR and local-projection bands confirm, in simulation, that the stated 68%/90% bands contain the truth at close to the claimed rate.

to December 2008 onward, because the PPI input-price series only begins there; the constraint and its source are logged like any other decision.

Exercise B asks how commodity cost shocks pass through to inflation, the question of Peersman (2026), with nothing but the PDF as input. We chose the paper for its date. The vintage used appeared in January 2026 (its earliest public version in October 2025), at the edge of the training window of the models the sessions ran on, so the exercise tests reading more than recall. The framework recovered the SVAR-IV design and data requirements from the paper alone, flagged that one external instrument (the harvest-based food-supply shock) has no public release that we could locate, collected the public series including the Känzig (2021) oil-supply-news instrument, and narrowed the exercise to the energy cost channel, the only one attainable with public data.

Exercise C is an open design question: the effects of fiscal policy shocks on output and inflation when the granular military-spending data behind the standard instruments are unavailable. We chose the topic because many central bank economists do not work with fiscal VARs day to day, so the exercise shows what the framework contributes when its user works outside their own field. The framework ruled out the defence-news route (Ramey and Shapiro (1998); Ramey (2011)), flagged fiscal foresight (agents anticipating tax and spending changes) as the identification threat, and proposed two pre-specified routes with their trade-offs, Blanchard and Perotti (2002) recursive identification and Mountford and Uhlig (2009) sign restrictions, together with foresight remedies: forward-looking variables and the Auerbach and Gorodnichenko (2012) state-dependent design.

Across the five exercises, MaRCo@BIS carried the work end to end. It read its sources (the verified entry in the mode A runs, the staged paper in B, the literature in C), reasoned through the obstacles and proposed the specifications with their remedies, wrote the code, assembled the data pipeline and produced the report. Our time went to the calls that remain with the economist: judging the proposed framings, weighing the assumptions and evaluating the results. Adaptation across economies required no new code, and what each run left behind at disposition (a drafted dictionary entry, data connectors, estimation settings) stays available to the institution. What varied across exercises was where judgement entered: confirming data substitutions in the simplest runs, respecification in the intermediate ones, and scoping and route selection in the open-ended ones.

Table 3. Single framework, five illustrative applications

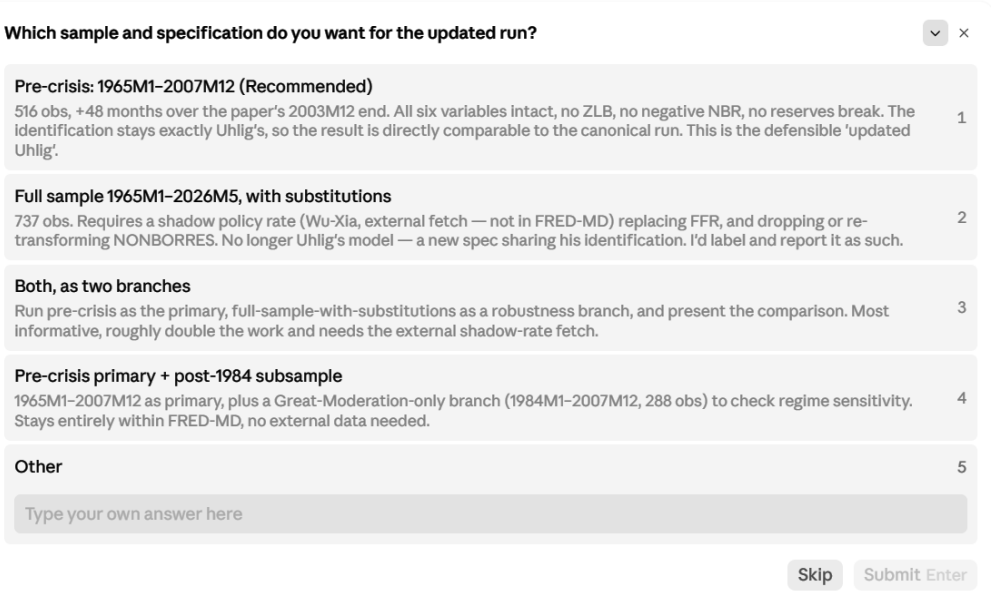
Exercise	Request (as typed)	Issues the framework raised	Solutions it offered
A-1 Verified entry	"Replicate the original results of Uhlig (2005)"	Data sourcing	Read the verified entry's variables, identification, sample and priors and ran the analysis end-to-end.
A-2 Data the spec can't handle	"Replicate Uhlig (2005) with updated US data"	Effective lower bound; non-borrowed reserves turning negative; pandemic-driven outliers.	Wu and Xia (2016) shadow rate; total reserves in place of non-borrowed reserves; dummies for the pandemic or Lenza and Primiceri (2022) volatility scaling.
A-3 Another economy	"Replicate Uhlig (2005) with UK data"	Breaks in the UK monetary policy regime; no domestic commodity-price index for the forward-looking supply role; no clean counterparts to the reserves variables.	Sourced the series itself (ONS, Bank of England); post-1997 sample (BoE operational independence), shortened to end-2008 onward by the commodity proxy's coverage; Krippner (2015) UK shadow rate; UK PPI input prices as the commodity-price proxy; drop the reserves variables.
B Paper ingestion	"Replicate this paper [. . .] on the commodity cost channel in the US"[1]	The SVAR-IV design and data had to be recovered from the PDF alone; one external instrument (Peersman's harvest-based food-supply shock) is not publicly available, as far as we could determine.	Extracted the VAR structure and collected the series, including external instruments such as the Känzig (2021) oil-supply-news shock; narrowed the exercise to the energy cost channel, the only one attainable with the publicly available data.
C Open design	". . . effects of fiscal policy shocks. . . no military-spending data. . . propose VAR-based alternatives"[2]	The missing data rules out the defence-news instruments (Ramey and Shapiro (1998), Ramey (2011)); and fiscal foresight (ie agents anticipating tax and spending changes) threatens identification.	Two routes with trade-offs: Blanchard and Perotti (2002) recursive identification and Mountford and Uhlig (2009) sign restrictions; plus foresight remedies: forward-looking variables and the Auerbach and Gorodnichenko (2012) state-dependent approach.

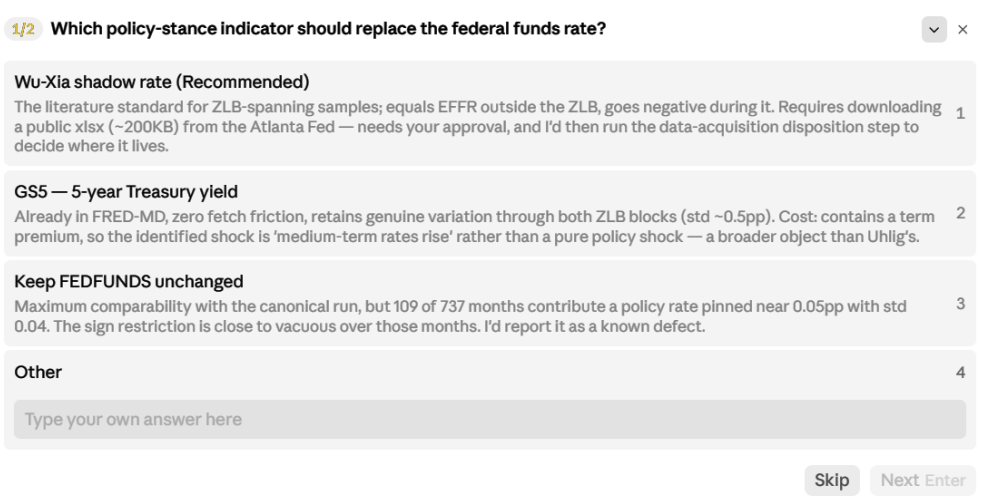
Notes: [1] The paper provided is Peersman, G (2026).
[2] The full prompt used is "I want to estimate the effects of fiscal policy shocks on output and inflation, but I have no access to the granular military-spending data usually used. Can you propose VAR-based alternatives using exclusion or sign restrictions?"
Source: Authors' elaboration.

4.2 A session, step by step (Mode A)

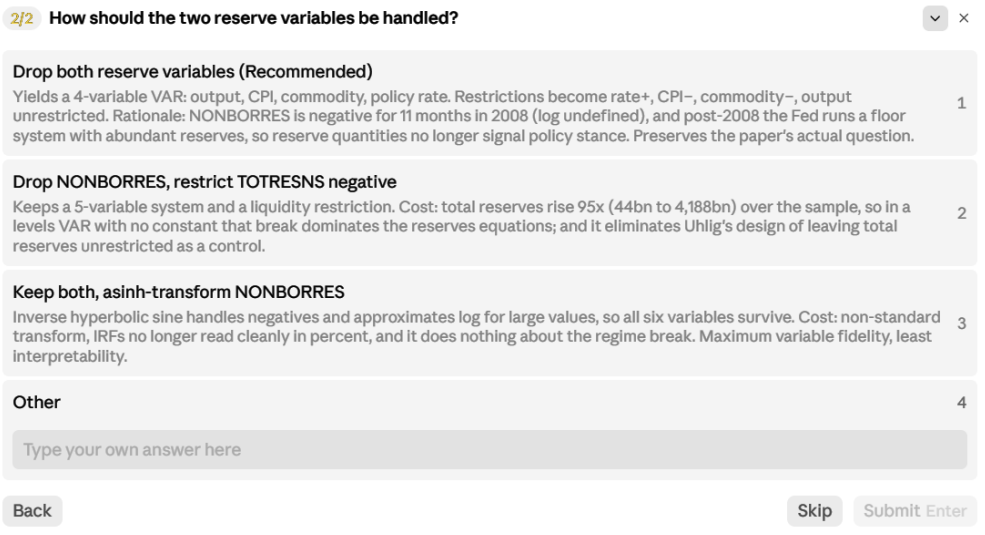
Table 3 summarises whole sessions in single rows; this subsection follows one in detail. The session is a fresh run of exercise A-2 on the released package, in a clean installation with no memory of the earlier runs: the request, as before, was "Replicate Uhlig (2005) with updated US data". Uhlig's identification restricts only the variables theory is confident about (after a contractionary shock, prices and reserves fall and the policy rate rises) and leaves the output response free; the well-known finding is that the data then do not give a clear answer. The adaptation produced in this session differs in detail from the run behind Table 3, and this difference is informative. A fresh session can produce a different, equally defensible specification from the same request. Across sessions, consistency is maintained through the same confirmation gates, logging and disclosure procedures. We describe four steps of the session; the decision cards reproduced between the paragraphs are unedited screenshots, shown as we saw them.

The first step was the opening proposal. Before estimating anything, the agent assessed the obstacles and declined to run the naive extension, noting that "one of them is arithmetically impossible": non-borrowed reserves turn negative for eleven months in 2008 (bottoming at −$333.5 billion, so the log transform is undefined), the federal funds rate spends 109 months at the zero lower bound (with a standard deviation of 0.04 percentage points there, the "rate rises" restriction is vacuous), and total reserves grow 95-fold under quantitative easing. Its proposal kept the paper's identification, re-based the policy rate on a shadow rate spliced into the two zero-lower-bound windows only, and put the contested choices (the sample, the policy indicator, the reserves block) to us as explicit options with trade-offs.





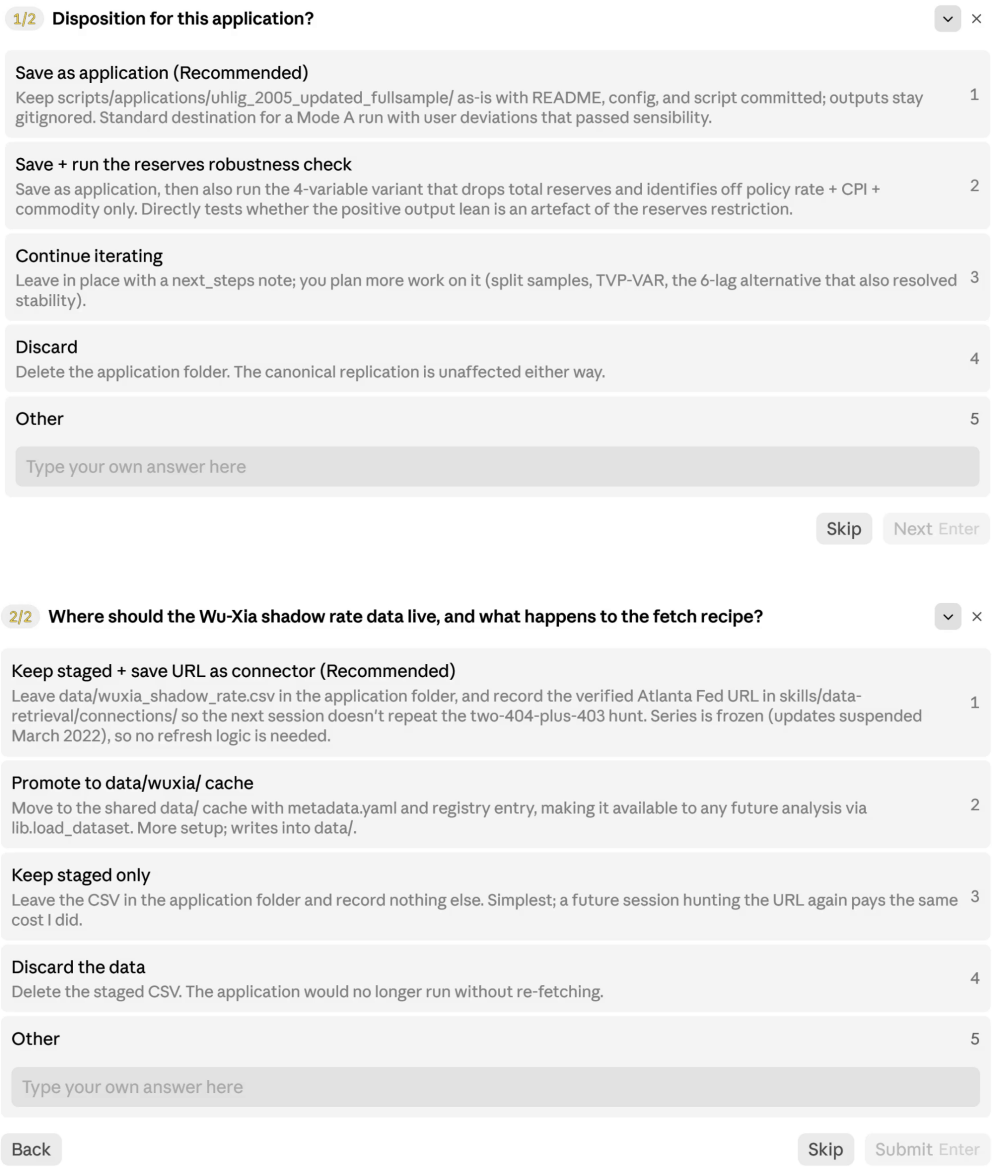
The second was an override. On the reserves block the agent recommended dropping both reserve variables; we instead chose to keep total reserves, restricted to fall, to stay closer to the paper's design. The choice and its reason entered the log, and it had a consequence the agent reported unprompted. The reserves response came out roughly seven times the canonical magnitude, a sign that the identified shock was partly a reserves shock, as the agent had warned ("the deviation I flagged as most likely to bite"). The check that would settle it, dropping reserves entirely, was recorded in the run state as the next action.



The third was the sensibility gate. The first estimate failed the stability check. The largest root came out at 1.0002, formally explosive, and the agent stopped before presenting any impulse response. The diagnosis across variants cleared the updated data and traced the failure to the paper's own no-constant design, which turns out to be explosive even on the original 1965–2003 window once the variable set changes. One bounded revision followed (adding a constant); the revised system was stable, and the rejected first run was kept and reported alongside, its long-horizon output drift identified as a side effect of the explosive root. The report's disclosure states it plainly: one sensibility-driven revision, both runs reported, the bands conditional on that revision.

The fourth was the close, in three parts. The application was saved with its code, configuration and both output sets. The fetched shadow-rate series got its own data disposition: the file stayed staged with the application, and the verified source address was recorded as a connector, together with a warning about the two plausible addresses that return a page instead of the file, so the next session will not repeat the hunt.

And the tool review flagged one helper function as a candidate for the shared library without promoting it; that write, too, is gated.



In the canonical run and both updated runs, the 90% band on the output response includes zero in each of months 1 to 24. Uhlig's agnostic finding therefore also holds in these adaptations, despite the additional twenty-three years of data and the changes to the policy rate and reserves variables. Figure 3 reproduces an excerpt of the final report: the specification summary, with the declared deviations from the paper and the headline finding, and the impulse responses with 68% and 90% bands.



Figure 3. Excerpt from the walkthrough session's output report

Notes: The specification summary, including the declared deviations from the paper and the headline finding, and the impulse responses with 68% and 90% bands. The full report adds diagnostics, variance decompositions and the plain-language narrative with the specification-search disclosure; every run produces the report and a research log.
Sources: Uhlig (2005); authors' elaboration.

4.3 From PDF to validated replication: ingesting a new paper (Mode B)

Exercise B starts from the situation mode B was built for. We gave the framework a paper (Peersman (2026), thirty pages, and nothing else) and asked for a replication on US data. The paper studies the commodity cost channel of inflation in the United States and the euro area, modelled jointly. It isolates exogenous energy and food cost shocks with external instruments and finds that these shocks explain only a modest part of the post-pandemic surge in inflation. In a counterfactual, overall commodity cost movements, including their endogenous responses to macroeconomic conditions, account for most of the surge in both economies. The choice of paper doubles as a control, though a soft one: the vintage used appeared in January 2026, and even the earliest public version, an October 2025 working paper, sits at the edge of the knowledge cutoff of the models the sessions ran on. Recall from training data can therefore not

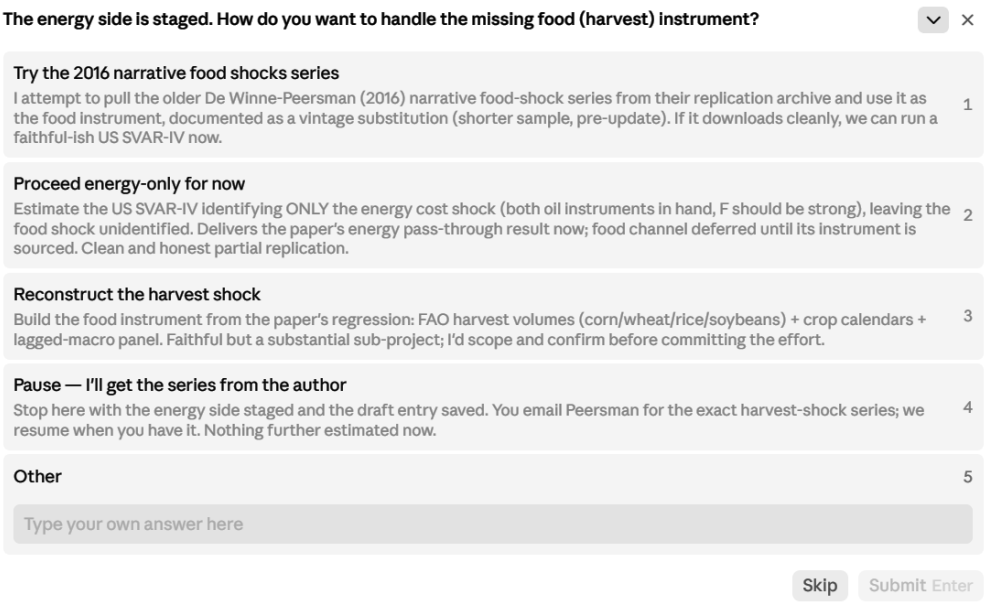
be relied on, but it cannot be ruled out either. The protocol requires the session to take everything it uses from the paper itself. The session opened with the ingestion protocol accordingly: verify the file, check the dictionary for an existing entry (there is none), extract the empirical core from the PDF, and draft a dictionary entry with a page reference on every substantive field, all before any estimation. The page references make the extraction checkable: every field of the resulting entry points at the passage it came from.

The extraction identified the method (a proxy SVAR, which the library supports in the required multi-instrument mode) and, more importantly, the constraint. In the session's words, the estimator is not the problem; the data is. Identification rests on three external instrument series, none of them in the pre-collected panels and several living only in author archives. The session proposed auditing what is actually obtainable before fixing any scope, and stopped at the end of the audit, reporting the result reproduced below: the entire energy side recovered, including the Känzig (2021) oil-supply-news instrument in the revised vintage of Mori and Peersman (2024) that the paper uses and the Baumeister and Hamilton (2019) flow oil-supply shocks, and the food instrument declared missing, with three costed routes to it (an older vintage, a reconstruction from crop data, or an email to the author).

I stopped before any estimation, as agreed. Here's what's obtainable:

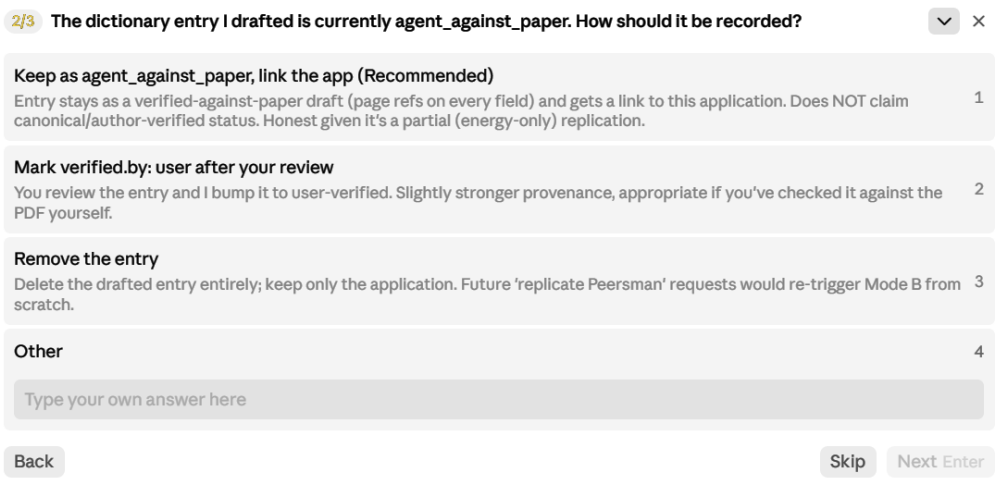
Component	Status	Source
Energy instr. 1 — Känzig oil supply news, Mori–Peersman revised	✅ staged	`shocksMP.x1sx`, monthly 1975:01–2024:08 (this is the exact revised series the paper uses)
Energy instr. 2 — Baumeister–Hamilton flow oil supply shocks	✅ staged	Baumeister datasets, 1975M2–2026M1
World IP (endogenous var) — Baumeister–Hamilton index	✅ staged	1958M1–2026M4
Commodity prices — World Bank Pink Sheet	✅ obtainable	public, worldbank.org
Real S&P 500 — Shiller	✅ obtainable	public, Yale
US macro — energy/food/core CPI, real GDP, 1yr rate	✅ obtainable	FRED
Food instr. — De Winne-Peersman / Peersman harvest surprises	⚠️ not cleanly posted	see below

Two scope decisions followed, each a choice among labelled options. The first reduced the paper's joint US and euro-area model to a nine-variable US block, recorded as a documented deviation. The request was for the US, and the euro-area block would roughly double the data assembly. The second came after the audit: with the food instrument missing, we chose the energy-only partial replication, leaving the food shock unidentified until its instrument is sourced. The alternatives included pausing the run and emailing the author for the series.



The assembly met the by-now familiar obstacles (two more bot-protected hosts, and a subtle quarter-labeling mismatch between the pre-collected panel and the fetched series, which the session diagnosed and fixed), and the estimation validated on the first pass: a first-stage F statistic of 24.2 against the paper's published 23.1, a stable system, and the paper's qualitative story reproduced, with energy costs passing through strongly to energy prices and only modestly to core inflation. No sensibility revision was needed; the report's specification-search count reads zero.

The session ended with three steps to make the work reusable. The application was saved as a validated partial replication, reusable as a proxy-SVAR template. The drafted dictionary entry was recorded with its provenance stated exactly: verified against the paper with page references on every field, linked to the application, and explicitly not claiming canonical status, so that a future request to replicate this paper routes through mode A. And the download routes for the recurring public sources became connectors. The write-up is explicit about what was not reproduced. The paper's headline, that commodity costs almost fully account for post-pandemic core inflation, needs both shocks and a counterfactual, so it stays out of reach until the food instrument is sourced. What is validated is the energy-channel building block.



Mode B therefore does not guarantee a full replication. It provides a documented route from a paper's PDF to a replication whose scope depends on the available data. The audit records the work needed to obtain those data, the scope decisions and the parts that could not be replicated.

4.4 Benchmarking against the originals

Modes A and B rest on the claim that the framework reproduces published results; this subsection tests that claim. We benchmark the agent's replications against the published originals (Figure 4). We start with the Uhlig (2005) replication of exercise A-1 (Panel A) and the Peersman (2026) ingestion of exercise B (Panel B). Both comparisons, however, carry ambiguity: a set-identified median in one case, a partly private dataset in the other. The third exercise is added to resolve it. Kilian (2009) is a deterministic design with deposited data, so any gap between reproduction and original could only be the framework's error. It is the control experiment against which the discrepancies in the other two can be read.

Uhlig's (2005) design is deliberately agnostic. It imposes only uncontroversial sign restrictions (prices and reserves fall, the policy rate rises) and stays silent on output, the variable of interest, so the data admit many models, not one, and the reported median depends on the posterior draws and on a prior over those models that the data do not update. Peersman's (2026) original is a 14-variable multi-country model; with its euro-area block and one bespoke instrument unavailable for this exercise, the reproduction is the energy half re-estimated on the public US subset.

Kilian (2009), added as the control, asks what moves oil prices: supply disruptions, booming global demand, or oil-specific precautionary demand. A three-variable model separates the three shocks by their timing, with supply moving first within the month; the specification pins down a unique answer, and the author deposited the data.

The three exercises therefore reproduce in three different ways, and we expect the same distinctions to arise in other applications. Where the specification pins down a unique answer on deposited data (Panel C), reproduction is exact: the input series are identical to the author's archive, and all nine impulse responses match the output of the author's own code to roughly 10^{-12}, the tiny residual showing that the numbers were recomputed independently, not copied. Where the design is set-identified (Panel A), the reported median is a draw-dependent summary of an identified set, so two correct implementations agree in shape and in finding but not to the decimal. Where the original data are partly private (Panel B), reproduction degrades predictably. On the public subset the headline pass-through magnitude appears broadly similar while its persistence does not; the missing series are a plausible cause, though this cannot be confirmed without them. Reproducibility is thus a property of the identification scheme and of data availability as much as of the tool, and exact agreement should be expected only where the design admits it.

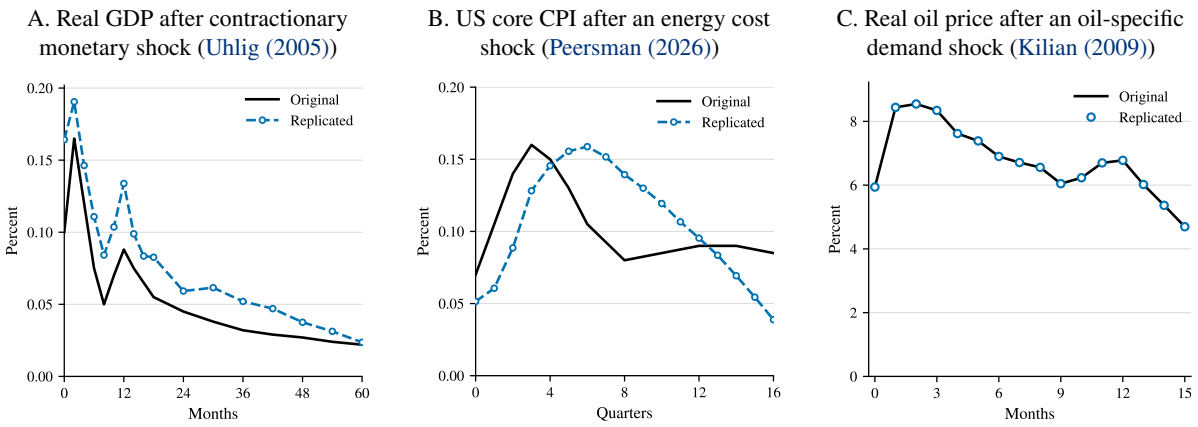


Figure 4. Published SVARs: MaRCo@BIS estimates and reference comparisons

Notes: Impulse response functions, in per cent. The lines labelled "Original" in Panels A and B are our rough visual approximations of the cited figures, included for qualitative reference only; they should not be used to assess numerical agreement. Panel A uses the dataset supplied for the Uhlig replication. Its median depends on the parameter and rotation draws, and the sampling procedure differs from the paper's. Panel B is a nine-variable US-focused adaptation of the original 14-variable model. It uses the public Känzig (2021) instrument alone, whereas the paper combines the Mori–Peersman revision of that series with Baumeister–Hamilton oil-supply shocks. The estimated responses are scaled using a rough reading of the published peak energy-price response. Panel C compares the framework's estimates with a stored copy of output from Kilian's deposited replication code.

Sources: Uhlig (2005); Peersman (2026); Kilian (2009); authors' calculations and visual approximations.

4.5 Paired sessions with and without MaRCo@BIS

The exercises so far show what the framework does; they do not show what it adds over an off-the-shelf frontier AI model, which is itself capable of econometric work. To make that margin visible, exercises A-1 and A-2 were each run twice on the same day, on the same underlying model (Claude Code), from the same one-sentence prompts shown in Table 3: once using MaRCo@BIS, on a clean installation, and once relying on Claude Code alone, in a plain coding-agent session started in an empty folder with no instructions beyond the request itself. All four sessions correctly reproduced Uhlig's headline finding, with the credible band on the output response straddling zero throughout. The comparison is therefore about process: where the decisions sat, what was verified, and what survives as a record. Table 4 provides a summary.

The Claude-only sessions performed well. Given the prompt "Replicate the original results of Uhlig (2005)" (our A-1 exercise), the model did not ask for additional input. It searched the web, found a reconstruction of Uhlig's dataset in a replication archive, downloaded it, ported the accompanying estimation programs to several hundred lines of fresh Python, and reproduced six of the paper's figures.

With the prompt "Replicate Uhlig (2005) with updated US data" (our A-2 exercise), it opened with one well-chosen question card (language, sample handling across the effective lower bound, scope) and then worked autonomously to a finished write-up: it spliced a shadow rate over the lower-bound years, handled the negative non-borrowed reserves of 2008 with a transform, added pandemic dummies, ran the original and the updated samples side by side, and volunteered a well-argued caveat that the post-2008 reserve aggregates have a different economic meaning. A frontier model can therefore ask relevant questions and reason about the underlying economics without being explicitly instructed to do so.

The first difference is in how data enters the analysis. What Claude found on the web for exercise A-1 was, on one dimension, the better choice: the reconstruction sits closer to the paper's original series than the documented public proxies the framework's entry uses. Yet the process behind it is riskier. The model searched, picked a file, downloaded it and estimated on it in one step, with no gate in between. So the only evidence that the data actually matched Uhlig's original dataset was the source's own claim, and a search re-run later, or by someone else, may surface a different file. Nothing in the process distinguished that

source from an arbitrary upload, and nobody checked what licence the file carried. MaRCo@BIS instead draws on known sources (pre-collected panels with re-download recipes, named connectors, a vintage registry) and puts the source decision to the researcher, so that data from outside the known set enters the analysis only by an explicit decision.

The reconstruction file Claude downloaded also carried a specification it should not have. The session estimated the VAR with a constant term, following the reconstruction's convention, while the original paper explicitly states that his specification included neither a constant nor a time trend. This is the very deviation the framework's entry audit caught in its own canonical script during development (Section 6.1), and an independent rebuild made the same mistake from a different direction. Including the constant leaves the headline finding intact. This makes the specification error easy to overlook when checking the figures against the paper. MaRCo@BIS automatically checked its results against the dictionary entry's numerical targets and used a specification verified against the paper with page references. Claude reported that its figures matched the originals, based on its own comparison. The implementations also differed: MaRCo@BIS reused estimation routines covered by automated tests, while Claude wrote new estimation code for the session. Agreement between the figures provides some reassurance, but does not establish that the underlying code correctly implements the specification.

The remaining differences concern what the sessions checked and what they left behind. Neither Claude-only session ran residual diagnostics; MaRCo@BIS's A-2 run ran the full battery, found the residuals clean of autocorrelation, rejected normality and homoscedasticity, and attributed the rejections to the crisis and pandemic outliers, flagging them. And while MaRCo@BIS left behind a research log, a run state, registry entries and an explicit disposition step, Claude Code left only clean code and a readme, with the decision trail recoverable only by reading the chat transcript.[5]

The last comparison is cost. Model use is billed in tokens, the units of text a model reads and writes, and the two setups spend them differently.[6] On A-1 the two sessions cost about the same in total, for different reasons. The framework session wrote far less new text, because it reused its tested replication script; the Claude-only session wrote all of its estimation code from scratch. The framework session instead spent more on reading, because its instructions, entries and records travel with the session and are re-read at every step. On A-2 the framework session cost roughly two-thirds more. Part of the gap had nothing to do with the framework. The institutional network blocked every automated route to the data, and working around the block took many attempts. The rest is work the framework genuinely adds, namely the diagnostics, the report, the log and the closing steps. In these sessions, the framework shifted spending from code generation to reading instructions and completing the additional workflow steps, without necessarily reducing total costs. These two comparisons do not establish how costs would differ for other tasks. The framework's reading overhead grows the longer a session runs and every time a session is reopened; its saving returns every time a task can reuse a verified specification and script instead of writing new code, and newly written text is billed at several times the rate of re-read text. Which effect dominates depends on how the tool is used; two short sessions cannot conclude.

Although limited in scope, the comparison illustrates the paper's main point: both setups reproduced the headline finding, while the framework made it easier to trace how the results were obtained.

[5]The two Claude-only sessions produced separate implementations, using different data sources and different treatments of constants and other deterministic terms. One fixed the random seed; the other did not. Although each implementation was internally coherent, these choices were not coordinated across sessions. MaRCo@BIS's shared routines are intended to reduce unintended differences of this kind.

[6]All four sessions ran the same usage accounting. Token prices differ by kind: re-reading text already held in the session is cheapest, adding new text to the session's memory costs more, and text the model newly writes costs the most, several times the base rate. The totals weight each kind by its price. On A-1 both sessions used about 1.6 million weighted tokens; on A-2 the framework session used about 3.7 million against about 2.2 million without.

Table 4. The same requests with and without the framework

Dimension	With the framework	Without the framework
Decisions put to the researcher	Gates at the binding moments: the data source and its substitutions, the redesign across the 2008 regime break, the disposition of the run and of new code.	One model-initiated card at the start of A-2 (language, sample handling, scope); none in A-1. The deep design calls (transforms, splices, pandemic dummies) made autonomously.
Data sourcing	Known sources: pre-collected panels with re-download recipes, named connectors, a vintage registry; the source choice confirmed by the researcher.	Web search at run time; in A-1 a dataset reconstruction found, downloaded and used in one autonomous step.
Specification	Verified dictionary entry, each field checked against the paper.	Rebuilt per run from the model's knowledge and the downloaded material; the two runs differ in deterministics.
Estimation code	Shared library backed by the test suite; the canonical script reused unchanged in A-1.	Several hundred lines of fresh numerical code per run, untested.
Validation	Machine-checked targets from the entry's validation contract.	Self-asserted match to the paper's figures.
Diagnostics	Sensibility gate plus residual diagnostics; soft failures flagged and attributed.	None run in either session.
Record left behind	Research log, run state, data registry entries, disposition and tool review.	Clean code and a readme; the decision trail lives only in the chat transcript.

Notes: Paired sessions of exercises A-1 and A-2, run on the same day on the same underlying model: with the framework on a clean installation, and without it in a plain coding-agent session started in an empty folder.
Source: Authors' elaboration.

5 From an open question to results (Mode C)

The previous sections showed the framework replicating and extending existing results. This section moves one step further and shows an exercise that is closer to everyday central bank practice: a policy question with no explicitly stated method or reference model and subject to data constraints. For this purpose we chose a question on the impact of fiscal policy. The state-of-the-art approach, based on granular military-spending instruments, cannot be readily applied to most other economies, so an economist facing the question of how fiscal shocks move output and inflation has to find alternative approaches.

The session replayed here ran on the released package from the request recorded in Table 3 (the full prompt is in the table's note) through estimation, a report, and three rounds of follow-up questions of the kind an economist might ask a fellow econometrician after seeing first results. Our main interest is the exchange. We examine how the agent reasoned about the economics, raised concerns, challenged our choices and proposed alternatives, and where our own judgement was needed. The decision cards and the excerpt reproduced below are unedited screenshots from the session.

5.1 Understanding the question

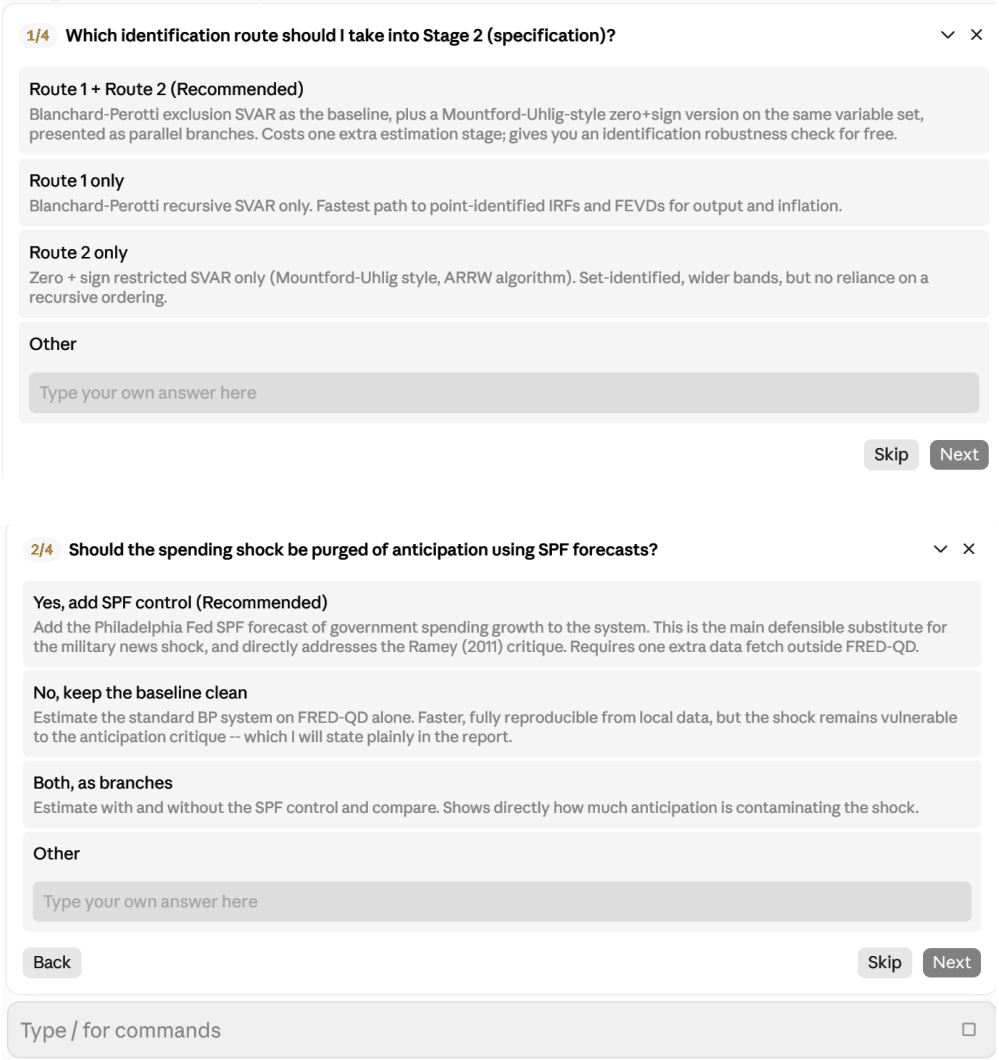
The agent's first move was to restate the problem in economic terms before proposing anything. Abandoning the military-news approach gives up one specific protection in identifying the fiscal shock, protection against anticipation: spending shocks identified from timing restrictions alone are partly forecastable. News of defence build-ups, by contrast, is treated as exogenous and dated at announcement (Ramey (2011)). Hence, alternative strategies need to consider this problem. The agent suggested using professional forecasts to account for anticipated changes in spending, without relying on military data. It then laid out three routes, each with its identifying assumptions spelled out.

The first, recommended as the baseline, follows the exclusion approach of Blanchard and Perotti (2002). It assumes that discretionary spending does not respond within the quarter to output, prices or taxes, on the grounds that policy decisions and implementation take time. This allows the spending shock to

be identified recursively, with spending ordered before those variables. The agent also noted that the paper's tax shock is not recursive: it requires a calibrated output elasticity of net taxes imposed before the recursive step. The library does not implement this adjustment, which is unnecessary if only spending shocks are of interest.

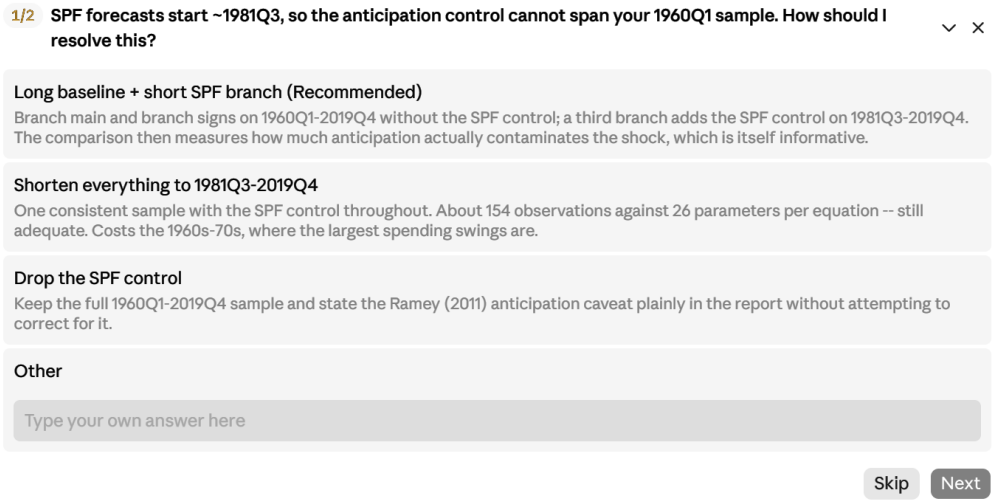
The second proposed route follows Mountford and Uhlig (2009) and is based on sign restrictions. Here the agent pointed to a pitfall. A naive "spending rises" restriction picks up the business cycle, because booms raise spending. Hence it suggested identifying a business-cycle shock and a monetary shock first and then extracting a spending shock orthogonal to both, with output and inflation left unrestricted so that the answer is informative instead of assumed. The agent also noted that the proposed implementation combines sign restrictions with zero impact responses of spending to the business-cycle and monetary shocks. It therefore retains some of the timing restrictions used in the recursive approach.

The third route used external instruments and was left for possible follow-up. A narrative tax instrument would shift the analysis to tax shocks; a spending forecast-error instrument could provide a check on the spending results under different identifying assumptions. The agent closed with a warning to be considered before any estimation. Spending shocks routinely produce a small or negative price response in this class of models (a known and well debated result), and the agent would flag such a result if it appeared, without changing the specification to remove it. Four decisions were then put to us (route, anticipation control, tax scope, and economy and sample), each with a recommendation and its negative sides; the two reproduced below are the ones with economic content.



5.2 Specification, and a conflict the framework surfaced

The initial six-variable quarterly VAR included spending, net taxes, output and the price level in logs, together with the professional-forecast growth rate and interest rate, and a constant and trend. Each variable was tagged with the knowledge source behind it (the dictionary entry for the fiscal instrument, the specification skill's core table for output and prices, parametric knowledge for the rest). The agent also expressed the caveat that the framework imposes no cointegrating restriction. It then highlighted a conflict between two choices we had just confirmed. The professional-forecast series begins in 1981, so the anticipation control cannot span the full sample starting in 1960. Three resolutions were offered with their trade-offs; we chose one consistent sample from 1981. Two smaller judgement calls the agent made without asking, but logged and left reversible (no per-capita normalisation, and the definition of net taxes), show the other side of the same practice: these choices were not gated, but they were recorded.



5.3 Results, the agent's dissent and our override

Under recursive identification, the impact multiplier was less sensitive to specification choices than the longer-horizon results. It was about 0.85 in the initial specification and remained roughly between 0.6 and 1.1 in checks using two, four or eight lags, three shrinkage settings, and omission of the anticipation control. The one-lag specification gave a lower estimate of 0.43. Beyond one year, the results were much less robust. Table 5 reproduces the specification range the agent put in front of us: the eight-quarter cumulative multiplier swings from -0.56 to $+1.04$, and the price response changes sign depending on the (apparently innocuous) choice of the lag order and the prior. The diagnosis was that the size of the sample does not enable the persistence of the shock to be pinned down properly: according to OLS estimates, the spending response decays to zero within three years and then oscillates, the ill-determined long-horizon behaviour of 26 parameters per equation on some 150 observations with a largest root of 0.98. The agent noted that Minnesota shrinkage would remove the lag sensitivity, but warned about its implications for persistence. The prior drives the largest root to 0.9999 and makes the spending shock nearly permanent, so the tidy long-run multiplier near one would reflect the prior more than the data. It recommended headlining the impact result and reporting the rest as unidentified.

Table 5. The specification range the agent reported before the baseline was chosen

Specification	Impact multiplier	Cumulative multiplier, 8 quarters	Output at 8 quarters (%)	Deflator at 12 quarters (%)
OLS, 1 lag	0.43	−0.05	−0.070	−0.014
OLS, 2 lags	0.70	−0.14	−0.098	+0.056
OLS, 4 lags (initial baseline)	0.84	−0.56	−0.116	+0.064
OLS, 8 lags	1.11	+0.58	−0.065	+0.104
Minnesota BVAR, 4 lags	0.87	+1.04	+0.158	−0.039

Notes: Impact multipliers are dollar multipliers at horizon zero; cumulative multipliers sum output over spending responses to the horizon shown. All specifications include the professional-forecast anticipation control; sample 1981Q4–2019Q4. OLS entries are point estimates and the Minnesota BVAR entries point responses at the posterior-mean parameters, so the OLS rows differ somewhat from the bootstrap medians plotted in Figure 5.
Source: The session's estimation output.

Minnesota shrinkage does fix the lag sensitivity — p=4 and p=8 become nearly identical. But I am not going to present that as the answer, because it buys stability by driving the largest eigenvalue to 0.9999, which makes the spending shock *permanent* (g stays near 0.6% out to h=60). The tidy, textbook-looking long-run multiplier of ~1.0 is then the unit-root prior speaking, not the data. Adopting it silently because it produces the expected sign is precisely the move this framework exists to prevent.

The sign-restricted branch is the most informative result of the session. Drop the recursive ordering and the output response is unsigned at *every* horizon — the 68% band spans −0.51 to +0.30 at h=8. Even the impact result does not survive. That is Uhlig's agnostic finding transplanted to fiscal policy, and it is the direct measure of how much work your identifying assumption is doing.

One caveat on the anticipation control: dropping the SPF variable moves the impact multiplier from 0.85 to 0.87. I would not read that as "anticipation doesn't matter" — more likely the SPF forecast is too coarse to capture what defence news captures.

Specification-search disclosure: zero revisions to the baseline. Everything above was run as a diagnostic, not as a search for a spec that produces the right sign, and all of it is reported.

Diagnostics for the record: residuals clean everywhere (Ljung-Box p > 0.19) except the deflator equation, p = 0.018; ARCH present in the SPF and tax equations; the spending shock explains 4.5% of output forecast-error variance at h=1, 1–3.5% thereafter.

We deliberately chose otherwise, adopting the Minnesota BVAR to see how the framework would handle a choice it had argued against. The agent recorded its position and built the run as instructed, with the disagreement written into the output itself: the OLS alternative is estimated on every run and displayed as the first robustness chart, the specification table carries the flag that the unit-root prior binds, and the report's narrative states that the persistence is imposed. The BVAR is estimated with a constant only, as the library's BVAR takes no trend term. Figure 5 shows what that contrast looks like. Under the prior the spending shock never decays and output settles permanently higher; under OLS the shock fades within three years, but output turns negative after only two quarters, while spending is still about half its initial level.

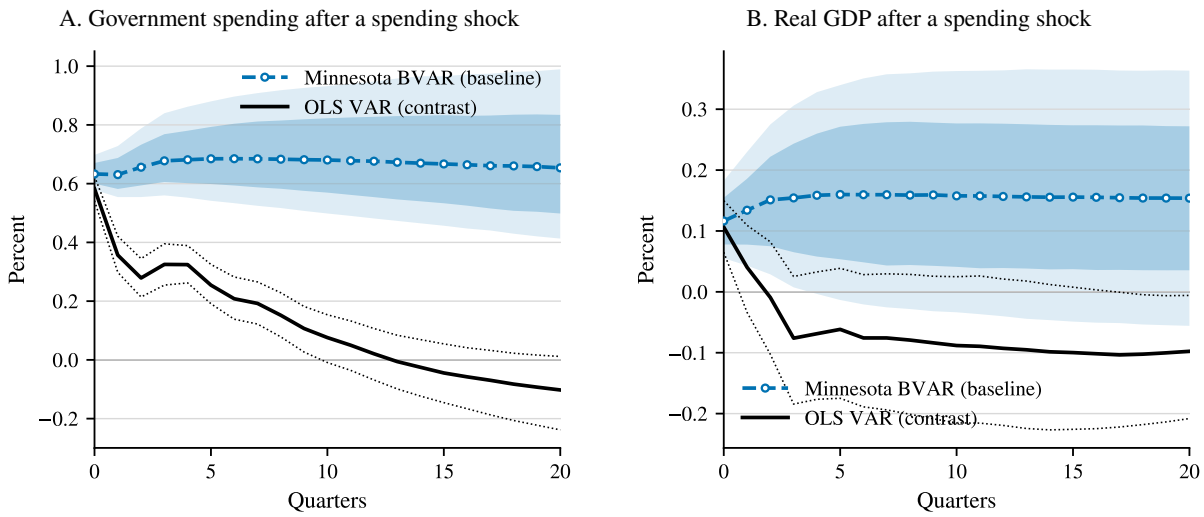
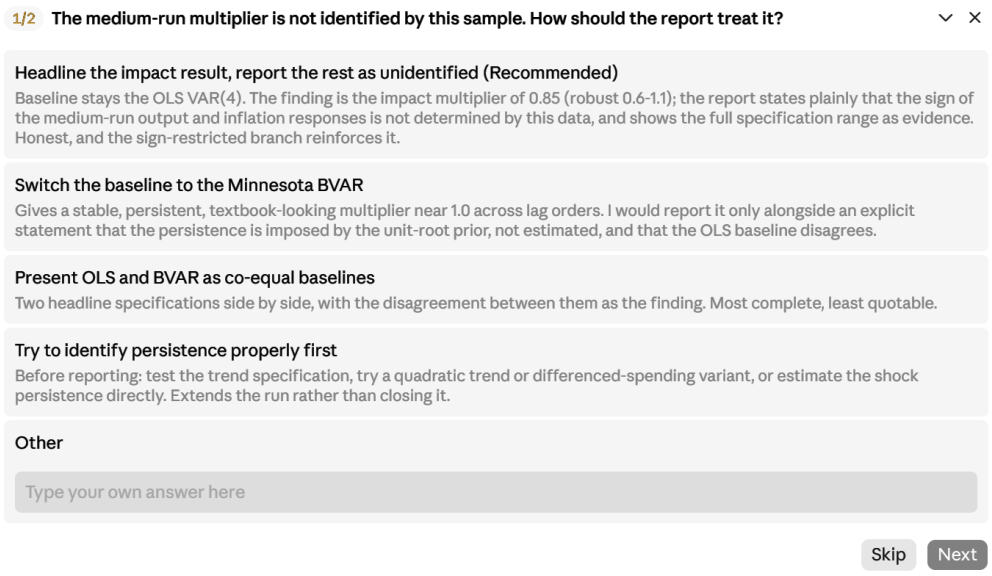


Figure 5. Government spending shock under a Minnesota BVAR and under OLS

Notes: Impulse responses to a one-standard-deviation government spending shock, in per cent; US quarterly data, 1981Q4–2019Q4, six-variable VAR(4) with the professional-forecast anticipation control, recursive identification with spending ordered ahead of output, prices and the policy rate. Dashed blue: Minnesota BVAR baseline, posterior median with 68% and 90% bands. Solid black: OLS contrast, median of the bootstrap distribution with 68% bootstrap band (dotted).
Source: Authors' calculations, from the session's own estimation output.



Two further results from the same stage were flagged. Under the alternative identification combining sign and zero restrictions, the output response was no longer clearly positive: its reported uncertainty intervals included zero at every horizon. The agent regarded this as the session's most informative result. As in Uhlig's analysis, the direction of the output effect remained uncertain when its sign was left unrestricted. The other was a caveat on the anticipation control: dropping it barely moved the impact multiplier. The agent declined to interpret this as "anticipation does not matter", suggesting instead that the forecast series is too coarse to capture what defence-spending news would capture. The initial diagnostic stage recorded zero revisions to the baseline. The subsequent switch from OLS to the BVAR was logged as one revision, made at our request.

5.4 Follow-up questions after the report

This session differed from the earlier ones in what happened after the report: we asked three follow-up questions, and each drew a different kind of contribution.

First, we asked for a historical decomposition of growth and inflation into spending and tax shocks under both baselines. The agent built it but warned against reading half of it as intended: it reminded us that the design does not identify a tax shock, and the evidence was strong enough to put the warning on the figure itself. The tax-shock impulse response was in fact wrong-signed (a tax increase raising output), and 2009Q1 registered a tax "shock" of nearly five standard deviations. This is odd in a quarter when net taxes collapsed because output and profits collapsed; in fact the decomposition had attributed part of the 2009 contraction to tax shocks with the causality running backwards. The agent offered the fix (the calibrated tax elasticity of Blanchard and Perotti (2002), scoped out at the specification stage) and, when we chose instead to treat tax shocks as unidentified, switched the decomposition to partial identification, so that the decision lives in the code and the bars still sum to the data.

We then asked how to rationalise the finding that the negative spending shocks of the post-GFC consolidation appear inflationary. The answer began "your premise is half right, and the half that is not is the important half". The half that is right: the shock series is real. The estimated spending shocks by year trace the stimulus fading in 2010, the budget caps of 2011 and the sequestration of 2013; the model picked up austerity where it happened. The half that is not: the inflation sign is mechanical. The baseline's price response to a positive spending shock is negative at every horizon, so a negative shock contributes positively by construction, and the OLS contrast reverses the sign in exactly those years. The agent then named the one structural story that could rationalise a negative price response: the neoclassical channel in which spending imposes a negative wealth effect, households supply more labour and potential output rises, consistent with the finding of Ramey and Shapiro (1998) that real wages fall after spending shocks. And it gave three reasons not to believe that story here. Standard analysis at the lower bound predicts the opposite (spending is more inflationary when the policy rate cannot offset it), and the shadow-rate robustness check had shown the lower bound was not driving the estimates. The deflator embeds the government consumption deflator, largely a public wage index, which consolidation should push down. And anticipation applies once more, since the caps and the sequestration were legislated years ahead, exactly the case in which dating the shock at implementation misplaces the price response. Its recommendation was to report a negligible contribution with a sign that is not robust across specifications, and to resist the structural reading. In this exchange we supplied the question and the economic hunch, the agent supplied the structural candidate, the counter-arguments and the evidence, and the judgement about what to conclude stayed with us.

The last question was whether the model's silence on monetary policy, conventional or unconventional, could confound the spending shock. The answer separated identification from interpretation. In a recursive scheme the spending shock depends only on the covariance of the residuals and the variables ordered ahead of it; adding a monetary instrument therefore could not change it by construction, so the only genuine test is to change the variable set. The agent ran that test (Table 6): adding a long-term yield, where unconventional policy actually operated, and swapping the funds rate for a shadow rate, alone and together, essentially leaves the identified spending shock unaffected. The agent then named where the concern does bite. A linear model estimates one average policy reaction function over a sample in which the lower bound binds for about a fifth of the quarters, while the literature predicts a larger multiplier when the policy rate cannot offset fiscal expansion, so the estimate applies cleanly to neither regime. It closed with four ways forward ranked by value per unit of effort, from the cheap one, splitting federal from state and local spending, to a state-dependent estimation around the lower bound and a jointly identified monetary shock, each with its cost stated.

Table 6. Testing the confounding concern: the spending shock under alternative monetary policy variables

Specification	Impact multiplier	Output at 8 quarters (%)	Deflator at 12 quarters (%)	Correlation with baseline shock	Correlation, 2010 onward
Baseline (federal funds rate)	0.87	0.158	−0.039	1.000	1.000
Plus ten-year Treasury yield	0.85	0.149	−0.037	0.998	0.999
Shadow rate instead of the funds rate	0.84	0.152	−0.045	0.998	0.995
Shadow rate plus ten-year yield	0.83	0.138	−0.045	0.996	0.995

Notes: Minnesota BVAR baseline throughout; the last two columns report the correlation of the identified spending-shock series with the baseline's, over the full sample and over 2010 onward.
Source: The session's estimation output.

5.5 Closing the run

The disposition step ran as in the earlier sessions, with two details worth recording. The data disposition included an offer the agent made but did not execute without consent: the downloaded FRED-QD file was exactly what was needed to restore the repo-level quarterly panel missing from this installation, a write outside the application folder, so it was proposed as an option and flagged when carried out. And the tool review surfaced five helper functions written during the session; we expressed no preference, so the framework's default applied and nothing was promoted to the shared library, with the candidates queued for later review. The library was untouched throughout.

An error encountered during the run shows why outputs need careful reading. The corrected shadow-rate robustness check found almost no change in the results. Output and price paths moved by less than a hundredth of a percentage point at every horizon. The agent reported that its first parser had taken the wrong column of the downloaded workbook, the effective funds rate instead of the shadow rate, which would have compared the baseline against itself and produced a false null indistinguishable from the genuine one it eventually found. It fixed the parser to match on the column name and to refuse to run otherwise, and it said so. The documented error and correction give the reader a basis for assessing the eventual result.

5.6 What the exchange shows

Taken as a whole, the session shows how the work was divided. The agent supplied the economics of the design (what the missing instrument actually protects against, and what substitutes for it), warnings registered in advance and honoured (the price response, the persistence), dissent when asked to adopt a specification it considered unsafe, refusals to let an output be over-read (the tax shocks), and answers to open questions, with the structural argument, the counter-arguments and a diagnostic run to settle what could be settled. Every consequential choice, including the one made against advice, was ours, and every one entered the log with the alternatives it displaced.

This session also illustrates why the results require scrutiny. The agent warned that the Minnesota baseline's persistent response depended heavily on the prior. The tax decomposition appeared plausible even though tax-policy shocks had not been identified, and the post-GFC inflation contribution changed sign across specifications. The comparison charts and warnings made these limitations visible, but the user still had to decide what conclusions the results supported.

6 Lessons, limitations and safeguards

6.1 Lessons

Building and testing the framework yielded four lessons that may generalise to other agentic research applications.

First, verification requires several kinds of review. Automated tests check the code, source-paper comparisons check the recorded specifications, and researchers review decisions requiring econometric judgement. These checks serve different purposes. For example, the source-paper audit found that a dictionary entry included a constant term absent from the published specification, even though the test suite passed.

Second, reliability improved as we moved tasks that could be specified in advance from the stochastic model to deterministic components. Early drafts relied on the model to recall specifications and write estimation code during each session. We encountered fewer errors after introducing checked dictionary entries and shared library functions. Errors in those components could then be corrected centrally, with the corrections available to subsequent runs. This experience informed the design principle in Section 3.1.

Third, the agent's limited context led us to distribute instructions and reference material across many short files. Keeping those files consistent became a recurring maintenance task: counts, cross-references and descriptions of behaviour often fell out of step during development. We therefore added checks to the test suite that flag discrepancies and cause the relevant tests to fail.

Fourth, a model change need not require rebuilding the framework. When our model was updated during development, the instructions, model specifications and code accumulated up to that point remained usable. The new model's behaviour still needed to be checked, so re-validation remains necessary after an update or replacement.

6.2 Limitations and safeguards

We discuss the main limitations and the safeguards intended to reduce the associated risks.

Economists still need to frame policy-relevant questions, interpret results in their institutional context and assess identification strategies, particularly in unfamiliar settings or when structural breaks are present. Outputs require expert review before they inform policy work. In our exercises, that review led us to intervene and materially change the specifications that were estimated.

Agentic tools make it easier to revise and rerun models. This can support careful analysis, but it also makes searching for a preferred result easier. MaRCo@BIS includes safeguards against such practices, although users can bypass them. The workflow requires the report to count sensibility-driven revisions, show first-pass results alongside final ones and label the uncertainty bands as conditional on the search (Section 3.2). The disposition log also records the attempts made within the session. These records make the search easier to review and reduce the risk that attempts go unreported. They cannot prevent a user from starting a fresh session and reporting only the chosen specification.

Even users acting in good faith may place too much confidence in a well-written answer from the agent, a risk known as automation bias. At each confirmation gate, users must assess the agent's proposals and decide whether to proceed. This requires econometric expertise, so the framework is intended for trained economists. Reproducibility checks, periodic re-audit of the dictionary (the framework ships a consistency-audit skill for exactly this) and explicit sign-off before release mitigate automation bias. Some mathematical problems may exceed the model's capabilities. The advisor protocol provides for consulting a stronger reasoning model in such cases.

7 Deployment

MaRCo@BIS accompanies this paper as a self-contained archive for use in an agentic coding environment. The included documentation explains how to set it up. The package contains a dictionary of published models, an estimation library covering VAR, BVAR, local-projection and factor models, and US data tools connected to public statistical APIs. It is a proof of concept: fixed random seeds, automated tests and comparisons with published results support reproducibility, but the agent's behaviour depends on the model and the user's instructions. Errors may remain in the code, recorded specifications or documentation, so users should check results before they inform research or policy work. The package is distributed as is under the Apache-2.0 licence, without warranty.

Further development could add published models, estimation methods and economy-specific knowledge bases. Our experience suggests particular value in documenting local data sources, licensing restrictions and changes in series definitions, which caused more difficulties than the estimation methods themselves. This knowledge could be reused across applications and shared among institutions, including those with limited technical capacity. New model entries and methods should undergo the same verification procedures as those in the released package.

The distributed package is a frozen snapshot of the framework described in Section 3. All counts reported in this paper refer to that release. Each dictionary entry includes a runnable replication, and the documentation provides commands for running the automated checks and comparing replication results with their targets.

8 Conclusion

The replications, adaptations and open design exercise show how MaRCo@BIS can support macroeconometric specification while researchers retain control of consequential decisions. Tested components provide the estimation machinery; confirmation gates, logging and disclosure make specification choices explicit and available for review. The framework helps researchers implement and examine an analysis, while its economic interpretation still requires their judgement.

Our exercises leave broader questions about how econometric practice should change as implementation becomes cheaper. How should specification search be disclosed so that readers can assess uncertainty bands when models can be revised within minutes? Who should curate verified dictionaries of published models, and to what standard? Which parts of econometric work should be assigned to fixed components, given the reliability gains we observed? MaRCo@BIS provides a working example through which researchers and central banks can examine these questions. We hope readers will try the framework, adapt it to local needs and share what they learn.

References

Adjemian, S, H Bastani, M Juillard, F Karamé, J Maih, F Mihoubi, G Perendia, J Pfeifer, M Ratto and S Villemot (2011): "Dynare: reference manual, version 4", *Dynare Working Papers*, no 1, CEPREMAP.

Afonso, S, S Galiani, R H Gálvez and R A Sosa (2026): "Deep research on a loop: using AI agents to construct economic datasets", *NBER Working Papers*, no 35188.

Andrade, P, F Ferroni and L Melosi (2023): "Identification using higher-order moments restrictions", *Federal Reserve Bank of Chicago Working Papers*, no 2023-28; revised (2025) as "Higher-order moment inequality restrictions for SVARs", *Federal Reserve Bank of Boston Working Papers*, no 25-3.

Antolín-Díaz, J and J Rubio-Ramírez (2018): "Narrative sign restrictions for SVARs", *American Economic Review*, vol 108, no 10, pp 2802–29.

Auerbach, A and Y Gorodnichenko (2012): "Measuring the output responses to fiscal policy", *American Economic Journal: Economic Policy*, vol 4, no 2, pp 1–27.

Bańbura, M, D Giannone and M Lenza (2015): "Conditional forecasts and scenario analysis with vector autoregressions for large cross-sections", *International Journal of Forecasting*, vol 31, no 3, pp 739–56.

Bańbura, M, D Giannone and L Reichlin (2010): "Large Bayesian vector auto regressions", *Journal of Applied Econometrics*, vol 25, no 1, pp 71–92.

Bańbura, M and M Modugno (2014): "Maximum likelihood estimation of factor models on datasets with arbitrary pattern of missing data", *Journal of Applied Econometrics*, vol 29, no 1, pp 133–60.

Baumeister, C and J Hamilton (2019): "Structural interpretation of vector autoregressions with incomplete identification: revisiting the role of oil supply and demand shocks", *American Economic Review*, vol 109, no 5, pp 1873–910.

Bernanke, B, J Boivin and P Eliasz (2005): "Measuring the effects of monetary policy: a factor-augmented vector autoregressive (FAVAR) approach", *The Quarterly Journal of Economics*, vol 120, no 1, pp 387–422.

Blanchard, O and R Perotti (2002): "An empirical characterization of the dynamic effects of changes in government spending and taxes on output", *The Quarterly Journal of Economics*, vol 117, no 4, pp 1329–68.

Blanchard, O and D Quah (1989): "The dynamic effects of aggregate demand and supply disturbances", *American Economic Review*, vol 79, no 4, pp 655–73.

Canova, F and M Ciccarelli (2013): "Panel vector autoregressive models: a survey", in T Fomby, L Kilian and A Murphy (eds), *VAR Models in Macroeconomics – New Developments and Applications: Essays in Honor of Christopher A Sims* (Advances in Econometrics, vol 32), Emerald, pp 205–46.

Canova, F and F Ferroni (2021): "A hitchhiker's guide to empirical macro models", *Federal Reserve Bank of Chicago Working Papers*, no 2021-15.

Cesa-Bianchi, A (2026): *VAR Toolbox 4.0*, https://github.com/ambropo/VAR-Toolbox (accessed July 2026).

Chang, A and P Li (2015): "Is economics research replicable? Sixty published papers from thirteen journals say 'usually not"', *Finance and Economics Discussion Series*, no 2015-083, Board of Governors of the Federal Reserve System.

Chen, J, J Piao, X Xu, S Li, T Xia, X He and Y Li (2026): "AgentEconomist: an end-to-end agentic system translating economic intuitions into executable computational experiments", *arXiv preprint*, arXiv:2604.27725.

Christensen, G and E Miguel (2018): "Transparency, reproducibility, and the credibility of economics research", *Journal of Economic Literature*, vol 56, no 3, pp 920–80.

Christiano, L, M Eichenbaum and C Evans (1996): "The effects of monetary policy shocks: evidence from the flow of funds", *Review of Economics and Statistics*, vol 78, no 1, pp 16–34.

Christiano, L, M Eichenbaum and C Evans (1999): "Monetary policy shocks: what have we learned and to what end?", in J Taylor and M Woodford (eds), *Handbook of Macroeconomics*, vol 1A, Elsevier, pp 65–148.

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

Dell'Acqua, F, E McFowland III, E Mollick, H Lifshitz-Assaf, K Kellogg, S Rajendran, L Krayer, F Candelon and K

Lakhani (2023): "Navigating the jagged technological frontier: field experimental evidence of the effects of AI on knowledge worker productivity and quality", *Harvard Business School Working Paper*, no 24-013.

Diebold, F and K Yilmaz (2012): "Better to give than to receive: predictive directional measurement of volatility spillovers", *International Journal of Forecasting*, vol 28, no 1, pp 57–66.

Dieppe, A, R Legrand and B van Roye (2016): "The BEAR toolbox", *ECB Working Paper Series*, no 1934.

Doornik, J (2009): "Autometrics", in J Castle and N Shephard (eds), *The Methodology and Practice of Econometrics: A Festschrift in Honour of David F Hendry*, Oxford University Press, pp 88–121.

Galí, J (1999): "Technology, employment, and the business cycle: do technology shocks explain aggregate fluctuations?", *American Economic Review*, vol 89, no 1, pp 249–71.

Gertler, M and P Karadi (2015): "Monetary policy surprises, credit costs, and economic activity", *American Economic Journal: Macroeconomics*, vol 7, no 1, pp 44–76.

Giannone, D, M Lenza and G Primiceri (2015): "Prior selection for vector autoregressions", *Review of Economics and Statistics*, vol 97, no 2, pp 436–51.

Hamilton, J (2018): "Why you should never use the Hodrick-Prescott filter", *Review of Economics and Statistics*, vol 100, no 5, pp 831–43.

Hendry, D and J Doornik (2014): *Empirical model discovery and theory evaluation: automatic selection methods in econometrics*, MIT Press.

Hendry, D and H-M Krolzig (1999): "Improving on 'Data mining reconsidered' by K D Hoover and S J Perez", *The Econometrics Journal*, vol 2, no 2, pp 202–19.

Hoover, K and S Perez (1999): "Data mining reconsidered: encompassing and the general-to-specific approach to specification search", *The Econometrics Journal*, vol 2, no 2, pp 167–91.

Jordà, Ò (2005): "Estimation and inference of impulse responses by local projections", *American Economic Review*, vol 95, no 1, pp 161–82.

Kamber, G, J Morley and B Wong (2018): "Intuitive and reliable estimates of the output gap from a Beveridge-Nelson filter", *Review of Economics and Statistics*, vol 100, no 3, pp 550–66.

Känzig, D (2021): "The macroeconomic effects of oil supply news: evidence from OPEC announcements", *American Economic Review*, vol 111, no 4, pp 1092–125.

Kato, M (2026): "AI Economist Agent: an agentic framework for model-grounded economic analysis with RAG, knowledge graphs, and large language models", *arXiv preprint*, arXiv:2606.20041.

Kilian, L (2009): "Not all oil price shocks are alike: disentangling demand and supply shocks in the crude oil market", *American Economic Review*, vol 99, no 3, pp 1053–69.

Kilian, L and H Lütkepohl (2017): *Structural vector autoregressive analysis*, Cambridge University Press.

Korinek, A (2023): "Generative AI for economic research: use cases and implications for economists", *Journal of Economic Literature*, vol 61, no 4, pp 1281–317.

Korinek, A (2025): "AI agents for economic research", *NBER Working Papers*, no 34202.

Krippner, L (2015): *Zero lower bound term structure modeling: a practitioner's guide*, Palgrave Macmillan.

Lanne, M, M Meitz and P Saikkonen (2017): "Identification and estimation of non-Gaussian structural vector autoregressions", *Journal of Econometrics*, vol 196, no 2, pp 288–304.

Lenza, M and G Primiceri (2022): "How to estimate a vector autoregression after March 2020", *Journal of Applied Econometrics*, vol 37, no 4, pp 688–99.

Litterman, R (1986): "Forecasting with Bayesian vector autoregressions – five years of experience", *Journal of Business & Economic Statistics*, vol 4, no 1, pp 25–38.

McCracken, M W and S Ng (2016): "FRED-MD: a monthly database for macroeconomic research", *Journal of Business and Economic Statistics*, vol 34, no 4, pp 574–89.

McCracken, M W and S Ng (2021): "FRED-QD: a quarterly database for macroeconomic research", *Federal Reserve Bank of St Louis Review*, vol 103, no 1.

Mertens, K and M Ravn (2013): "The dynamic effects of personal and corporate income tax changes in the United

States", *American Economic Review*, vol 103, no 4, pp 1212–47.

Mori, L and G Peersman (2024): "Estimating the macroeconomic effects of oil supply news", *CESifo Working Papers*, no 11532.

Mountford, A and H Uhlig (2009): "What are the effects of fiscal policy shocks?", *Journal of Applied Econometrics*, vol 24, no 6, pp 960–92.

Noy, S and W Zhang (2023): "Experimental evidence on the productivity effects of generative artificial intelligence", *Science*, vol 381, no 6654, pp 187–92.

Peersman, G (2026): "Understanding post-pandemic inflation fluctuations: the commodity cost channel", *CESifo Working Papers*, no 12415.

Primiceri, G (2005): "Time varying structural vector autoregressions and monetary policy", *The Review of Economic Studies*, vol 72, no 3, pp 821–52.

Ramey, V (2011): "Identifying government spending shocks: it's all in the timing", *The Quarterly Journal of Economics*, vol 126, no 1, pp 1–50.

Ramey, V (2016): "Macroeconomic shocks and their propagation", in J Taylor and H Uhlig (eds), *Handbook of Macroeconomics*, vol 2, Elsevier, pp 71–162.

Ramey, V and M Shapiro (1998): "Costly capital reallocation and the effects of government spending", *Carnegie-Rochester Conference Series on Public Policy*, vol 48, pp 145–94.

Rigobon, R (2003): "Identification through heteroskedasticity", *Review of Economics and Statistics*, vol 85, no 4, pp 777–92.

Romer, C and D Romer (2004): "A new measure of monetary shocks: derivation and implications", *American Economic Review*, vol 94, no 4, pp 1055–84.

Schorfheide, F and D Song (2015): "Real-time forecasting with a mixed-frequency VAR", *Journal of Business & Economic Statistics*, vol 33, no 3, pp 366–80.

Seabold, S and J Perktold (2010): "statsmodels: econometric and statistical modeling with Python", *Proceedings of the 9th Python in Science Conference*.

Sims, C (1980): "Macroeconomics and reality", *Econometrica*, vol 48, no 1, pp 1–48.

Stock, J and M Watson (2018): "Identification and estimation of dynamic causal effects in macroeconomics using external instruments", *The Economic Journal*, vol 128, no 610, pp 917–48.

Uhlig, H (2005): "What are the effects of monetary policy on output? Results from an agnostic identification procedure", *Journal of Monetary Economics*, vol 52, no 2, pp 381–419.

Vilhuber, L (2020): "Reproducibility and replicability in economics", *Harvard Data Science Review*, vol 2, no 4.

Wu, J and F Xia (2016): "Measuring the macroeconomic impact of monetary policy at the zero lower bound", *Journal of Money, Credit and Banking*, vol 48, no 2–3, pp 253–91.