Agentic AI Orchestration of Heterogeneous Economic Models for Rapid, Multi-Scenario Analysis of Energy Crises

Dana Golden [1] Brett Indelicato [2] Lav R. Varshney [3] Suzanne Thornsbury [4]

[1]CEWIT, Stony Brook University [2]Dept. of Economics, Stony Brook University [3]AI Innovation Institute, Stony Brook University [4]Institute for Food and Agricultural Studies, University of Florida

Motivation: Crises Move Faster than Analysis

- Energy-supply shocks propagate through **interlinked global systems** within days. Rigorous models take months to construct.
- The models needed **already exist**, but exist apart: different groups, languages, frameworks, and incompatible I/O.
- Manual integration can take **longer than the crisis**. Analysts narrow the question to fit one framework and run a handful of hand-crafted scenarios.

The binding constraint is not modeling capacity

It is the cost of orchestration: making heterogeneous models interoperate, substituting one for another as the question shifts, and chaining outputs across scales.

Contribution: Substitution as a First-Class Operation

An agentic framework for model orchestration

- **Typed-adapter registry:** every domain model wrapped in a Pydantic-validated adapter. Adding a model = registering one adapter.
- **Declarative configuration:** swap a scenario, retune a linkage, or change a tolerance band by editing YAML, not code.
- **LLM as orchestrator, not generator:** composes scenario narratives, extracts parameters, dispatches models level-by-level, synthesizes outputs.

Two Non-Negotiable Disciplines

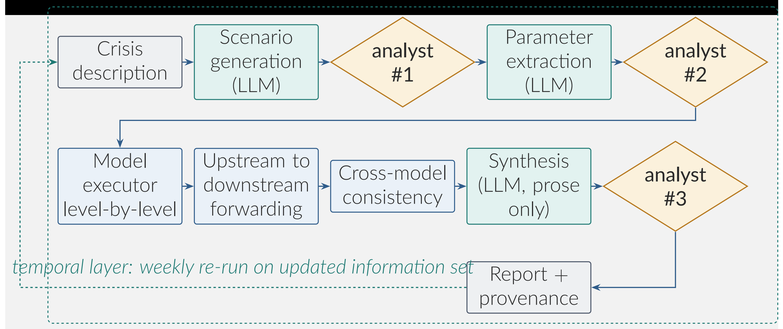
Human-in-the-loop

Three mandatory analyst checkpoints. No fully autonomous mode.

Non-fabrication invariant

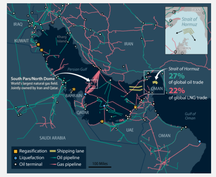
The LLM composes *prose only*. Every number is passed through unaltered from a model run, with full provenance.

The Orchestration Pipeline



Teal = LLM stages amber diamonds = mandatory analyst checkpoints

Demonstration: 2026 Closure of the Strait of Hormuz



Why this case?

- Pre-closure: ∼20M bbl/day of oil and ∼22% of global LNG transit.
- Operation Epic Fury (Feb. 28, 2026) forced operational closure. Brent peaked near $126/bbl.
- Five commodity systems hit: oil, LNG, fertilizer, helium, water.

Why it is a hard test

- Four nested analytical levels; 16 domain models orchestrated from a 34-adapter registry.
- Five scenarios, refreshed weekly for eight weeks; 8 to 16 models contribute per scenario, five anchors under all.

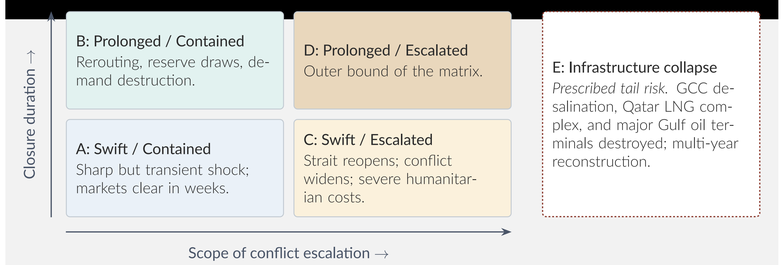
The crisis is the proving ground: the orchestration capability is the object of study, not a forecast of the crisis.

The Analytical-Level Stack: 16 Models Orchestrated

Combat physical events	infrastructure damage, capability degradation
	physical disruption
Commodity markets	**Oil:** BKR · POLES · MarketSim **LNG:** GGM · LNGST · Energy Flux **Helium:** World Helium **Fertilizer:** World Fert. **Water:** CWatM · WEAP · SahysMod
	helium sectoral shortfall
Commodity-downstream	SimRLFab (fab throughput) **Argonne Helium ABM**
	commodity price & supply shock vector
Short-run macro	NEMS MAM NREL
	structural calibration
Long-run macro & energy systems	**CGE:** MPSGE.jl · pycge · MIRAGRODEP · OpenCGE **Energy systems:** OSeMOSYS · MESSAGEix · TEMOA

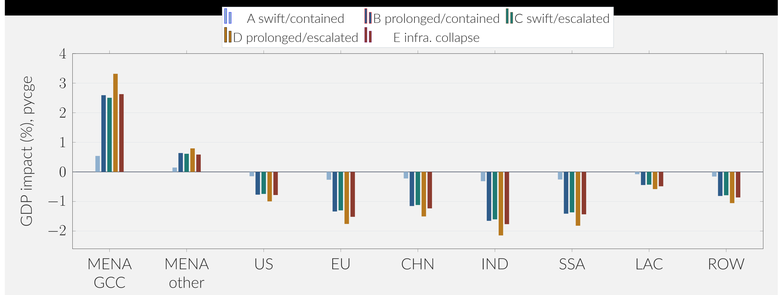
Parallel within a level, sequential across levels. Forwarded quantities are passed through with an audit trail, never re-derived by the LLM.

Scenario Architecture: 2×2 Matrix Plus Prescribed Tail



Scenario E activates the water adapters that A–D leave dormant.

Regional Incidence: Producers Gain, Importers Lose



- Gulf producers *gain* (terms of trade), roughly +2.5% to +2.7% GDP under escalation. Sub-Saharan Africa and India are hit hardest. Both CGE models agree qualitatively, disagree on magnitude.
- The US now sits on the *supply* side of the disruption. Fed oil-model peak GDP impacts cluster below 1% under A–C.

Findings Only a Portfolio Can See

The worst price shock is not the worst scenario

Fed oil impulse across A–E: +54.8, **+60.5**, +58.3, +51.8, +48.4%. The peak is under prolonged-but-contained B, not the tail. Visible only when all five scenarios are produced and compared in one view.

The water dimension is underpriced

Under escalation: ∼30% loss of GCC desalination capacity, unmet water demand of 9–13%, roughly 30 million people affected. Market-focused analysis largely missed this humanitarian exposure.

Disagreement is information, not embarrassment

Gas-signal readings span +25% to +126%. The consistency module flags the spread rather than averaging it, pointing the analyst to the question underneath: the models price different objects (wholesale LNG, forward index, delivered gas).

The Trust Architecture in Action

- **Forwarding holds mechanically.** Under Scenario C the oil shock appears as +58.25% in the Fed oil model, POLES-JRC, and both CGEs: agreement to the basis point.
- **Outliers are flagged, not deleted.** pycge reports global GDP −15.7% under the mildest scenario, orders of magnitude out of line. The number and the warning travel to the analyst together.

Takeaways and Limitations

- **Substitution as a first-class operation:** typed adapters + declarative config beat hand-wired pipelines.
- **LLM orchestrates, never originates quantities:** the non-fabrication invariant is structural, not exhortation.
- **Cheap substitution turns a one-off study into a monitored trajectory:** holding the registry fixed makes weekly changes attributable.
- *Limitations:* most adapters currently run as analytical-MVP reductions, so outputs are order-of-magnitude readings, not forecasts. Regional resolution rests on two CGE models; analyst checkpoints mitigate but do not eliminate extraction bias.