IFPRI Discussion Paper 02424

June 2026

An Agentic AI Assistant for Country-Level Economic Modeling

Methods, Data, and Expert Evaluation

Askar Mukashov

Soonho Kim

Peixun Fang

Xinshen Diao

James Thurlow

Joshua Proctor

Alan Rennison

Foresight and Policy Modeling Unit
Markets, Trade, and Institutions Unit

INTERNATIONAL FOOD POLICY RESEARCH INSTITUTE

The International Food Policy Research Institute (IFPRI), a CGIAR Research Center established in 1975, provides research-based policy solutions to sustainably reduce poverty and end hunger and malnutrition. IFPRI's strategic research aims to foster a climate-resilient and sustainable food supply; promote healthy diets and nutrition for all; build inclusive and efficient markets, trade systems, and food industries; transform agricultural and rural economies; and strengthen institutions and governance. Gender is integrated in all the Institute's work. Partnerships, communications, capacity strengthening, and data and knowledge management are essential components to translate IFPRI's research from action to impact. The Institute's regional and country programs play a critical role in responding to demand for food policy research and in delivering holistic support for country-led development. IFPRI collaborates with partners around the world.

AUTHORS

Askar Mukashov (A.Mukashov@cgiar.org) is a Research Fellow in the Foresight and Policy Modeling (FPM) Unit of the International Food Policy Research Institute (IFPRI), Washington, DC.

Soonho Kim (Soonho.Kim@cgiar.org) is a Senior Data Manager in IFPRI's Markets, Trade, and Institutions (MTI) Unit, Washington, DC.

Peixun Fang (P.Fang@cgiar.org) is a Senior Research Analyst in IFPRI's FPM Unit, Washington, DC.

Xinshen Diao (X.Diao@cgiar.org) is a Senior Research Fellow in IFPRI's FPM Unit, Washington, DC.

James Thurlow (J.Thurlow@cgiar.org) is the Director of IFPRI's FPM Unit, Washington, DC.

Joshua Proctor (Josh.Proctor@gatesfoundation.org) is a Principal Research Scientist of AI, Deep-Learning, and Applied Mathematics at the Gates Foundation, Seattle, WA.

Alan Rennison (Alan.Rennison@gatesfoundation.org) is a Senior Program Officer in the Global Development Division of the Gates Foundation, Seattle, WA.

ABSTRACT

The demand for high-quality, rapid economic analysis to navigate complex issues faced by many low- and middle-income countries has led to the development of detailed structural simulation models, such as Computable General Equilibrium (CGE) models. Policy analysis with such models requires deep knowledge of their structure and applicability to the policy issues at hand. Policymakers in these settings often lack access to the expertise required for articulating, analyzing, and interpreting the relevant causal chains captured by the models. Attempting to circumvent these barriers by submitting complex economic questions directly to off-the-shelf large language models (LLMs) introduces severe analytical risks, including hallucinations and insufficient expert guidance. To resolve this limitation, we developed and empirically evaluated an agentic AI assistant called RIAPA-AI that integrates LLMs with a CGE model. We evaluated the performance of RIAPA-AI against expert human CGE modelers and a general-purpose plain LLM baseline across samples of complex economic scenarios, utilizing an independent panel of senior economists to grade the outputs. Our statistical analysis reveals no statistically significant difference in analytical accuracy between RIAPA-AI and human experts, while the AI accelerates reproducible policy analysis from weeks to minutes. Furthermore, by operating without manual processing limits, RIAPA-AI eliminates the 6.7% error rate observed among human modelers. Conversely, the general-purpose plain LLM exhibits profound failure rates, failing to achieve policy-ready scores in over 60% of depth evaluations. Without an underlying CGE model acting as a bounding force to reflect economic structural constraints, the general-purpose plain LLM defaults to linear economic assumptions and inserts unmodeled socio-political narratives. Crucially, by explicitly restricting the AI's narrative interpretation solely to deterministic numerical outputs, RIAPA-AI mitigates the risk of unverified assumptions and logic hallucinations. We conclude that by deploying an agentic AI assistant that layers a generative AI over a formal CGE model, RIAPA-AI successfully delivers sensible, rigorous, and rapid policy analysis.

Keywords: Computable General Equilibrium, Artificial Intelligence, Large Language Models, AI Agents.

ACKNOWLEDGMENTS

This publication was made possible through support from the Gates Foundation under the grant INV-073291 and the International Affairs Office at the Presidential Court of United Arab Emirates under the "Future Food Systems" grant.

This work was undertaken as part of the CGIAR Science Program on Policy Innovations. We would like to thank all funders who supported this research through their contributions to the CGIAR Trust Fund: https://www.cgiar.org/funders/.

We are deeply grateful to Sherman Robinson and Karen Thierfelder for their rigorous, independent evaluation and grading of the simulation outputs. Furthermore, we extend our sincere thanks to Rui Benfica, Luis Escalante, Sherwin Gabriel, Faaiqa Hartley, and Karl Pauw for their extensive work conducting the baseline manual CGE simulations required to construct the evaluation matrix.

1. INTRODUCTION

The demand for rapid, high-quality economic analysis has never been greater. Policymakers in low- and middle-income countries need comprehensive economic assessments to navigate complex economic issues and crises. Examples include extreme climate variations; global food, fuel, and fertilizer price shocks; and macroeconomic policy changes in trade, fiscal, public, and foreign investments. Traditionally, CGE models, such as the RIAPA (Rural Investment and Policy Analysis) model developed by the International Food Policy Research Institute (IFPRI, 2025), have provided this analytical backbone. These models allow researchers to evaluate the complex effects of various exogenous shocks and policy changes on distributional outcomes across socioeconomic groups and on trade-offs across the economy. For instance, RIAPA has been instrumental in guiding national responses to global commodity price spikes (Arndt et al., 2023) and assessing strategies for poverty reduction (Arndt et al., 2012).

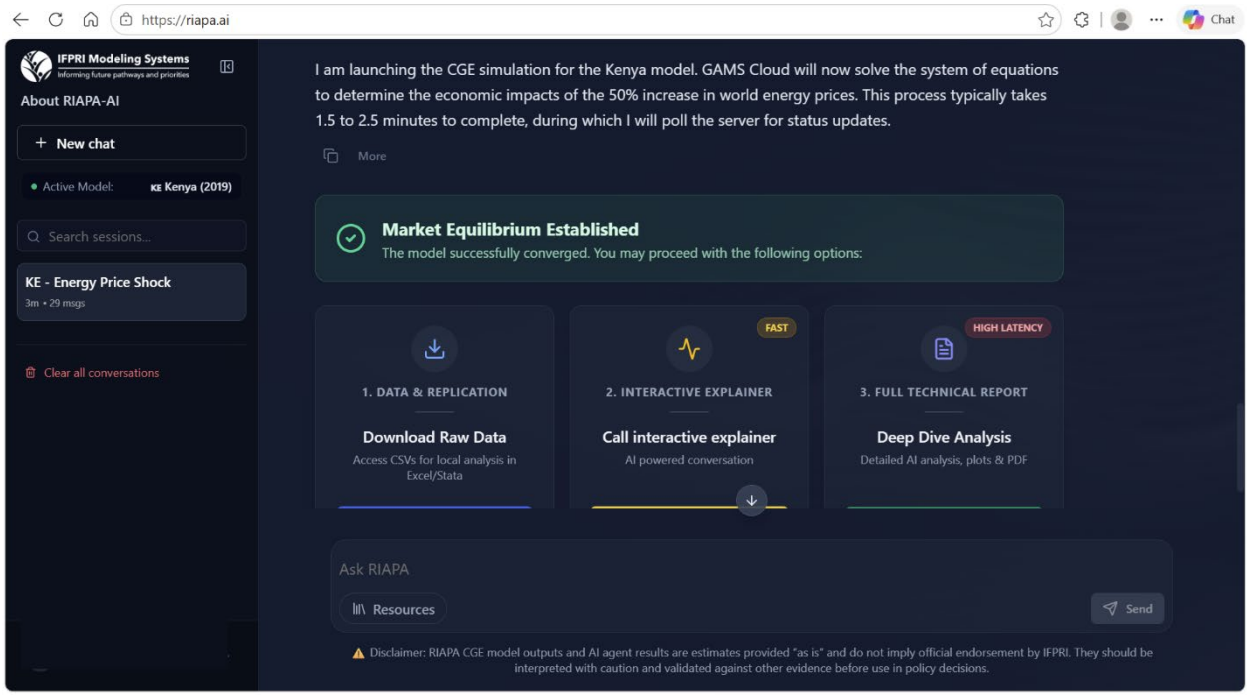
However, CGE modeling is constrained by significant operational bottlenecks. Designed for rigorous policy analysis, these models often contain thousands of mathematical equations representing an entire economy. Their implementation requires substantial CGE training and experience, making them inaccessible to many economists and non-modelers alike (see, e.g., EPA, 2017; Böhringer et al., 2003). With a large system of non-linear equations representing an economy in a general equilibrium setting, executing a CGE simulation requires that all economic agents interact with each other within a resource-constrained economy. A shock introduced into a CGE model causes *all* prices and quantities to adjust across sectors, commodities, and markets until a new equilibrium is reached. Beyond the reliance on advanced algebraic modeling software required to computationally solve this system of equations, accurately describing and interpreting the direct and indirect effects across thousands of economic variables demands deep and specialized expertise in both the specific model structure and general equilibrium theory. A single rigorous economic simulation can require hours of concentrated work by a well-trained CGE economist, with substantial variation in efficiency and judgment across analysts, and may take days, or even weeks, to

translate into a comprehensive analytical report when accounting for interpretation, validation, documentation, review, and normal constraints on human availability.

RIAPA-AI builds on a standard RIAPA CGE model powered by generative AI. RIAPA-AI operates as an agent designed to interact with users via LLMs to seamlessly specify simulation scenarios, run the CGE model, analyze the results, and draft analytical reports in a matter of minutes. The AI component of RIAPA-AI is a general-purpose LLM acting as a user interface, guided by IFPRI-developed instructions that translate user queries into specific, quantified model inputs. These inputs are routed to a cloud-based CGE model, which calculates the new economic equilibrium. The LLM subsequently fetches these results and interprets them by analyzing how thousands of economic variables deviate from their baseline values. Crucially, the underlying CGE model strictly preserves its formal structure, while the LLM is solely responsible for specifying the scenario inputs, fetching the numerical solutions, and translating those outputs into a readable report with consistent tables and figures. Because the LLM's narrative is strictly bound by the deterministically solved results of the CGE model, RIAPA-AI effectively eliminates the risk of logical hallucinations inherent in general-purpose LLMs. In this context, the first objective of this paper is to evaluate the performance of the RIAPA-AI agent against the CGE modelers from IFPRI who designed it: How accurately does the AI agent conduct CGE analysis in comparison to expert CGE modelers, and what is the operational advantage of RIAPA-AI?

The rapid development of general-purpose LLM technologies made the creation of RIAPA-AI possible; however, using a general-purpose LLM *per se* presents a fundamental methodological hazard. Policymakers and non-expert modelers, drawn by the low cost and convenience of general-purpose LLMs, may be tempted to circumvent formal economic modeling entirely by submitting complex economic questions directly to off-the-shelf LLMs (e.g., ChatGPT or Gemini). This raises a critical question that forms the second objective of this paper: Are the foundational training data and built-in reasoning capacities of general-purpose LLMs sufficient to provide satisfactory answers to complex economic questions on their own, without the underlying formal economic models? If so, RIAPA - and by extension, RIAPA-AI - would be redundant.

Figure 1. Screenshot of the RIAPA-AI user interface



To provide definitive empirical answers to our two objectives, we compare the performance of RIAPA-AI against both expert IFPRI CGE modelers and a plain, general-purpose LLM[1], using a Kenyan version of RIAPA as a case study. A custom expert evaluation methodology was designed to grade the quality of analytical responses to country-specific economic scenarios. The results reveal two primary findings. First, RIAPA-AI accurately implements CGE simulations, matching and operationally surpassing the CGE modeling experts who built it, thereby eliminating the 6.7% human error rate driven by high-volume manual processing limits. Second, relying on plain, general-purpose LLMs for economic assessments poses severe analytical risks. Even when the general-purpose LLM successfully identifies the correct direction of an economic shift, it suffers from a lack of quantitative precision. Without an underlying CGE model acting as a bounding force to capture the important structural constraints of an economy, the general-purpose LLM frequently hallucinates outcomes and provides strictly linear, one-sided, and often biased economic logic. Consequently, its output devolves into a loose collection of general directional changes, rendering it fundamentally unsuitable for precise policy formulation.

[1] We utilized Google's Gemini 3 Pro (gemini-3-pro-preview), accessed via API with a temperature parameter setting of 0 (an LLM parameter that defines response randomness).

The remainder of the paper is organized as follows: Section 2 explains the key structure of RIAPA-AI. Section 3 details the evaluation methodology. Section 4 presents the statistical evaluation results, and Section 5 concludes.

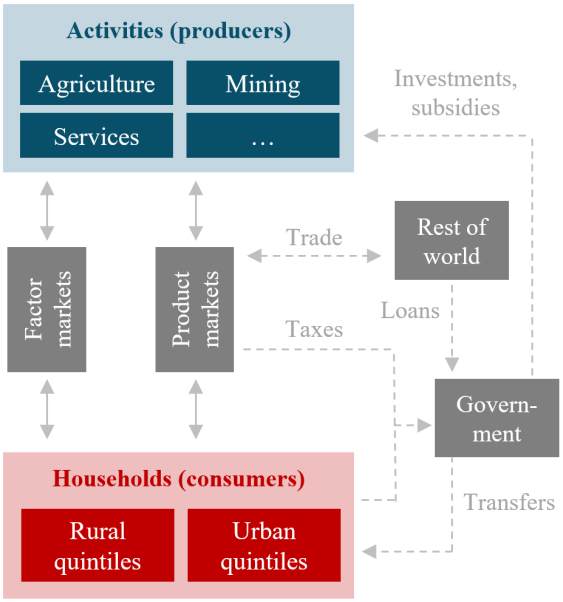
2. DESIGN OF RIAPA-AI

2.1. Economic Model at the Core

At the core of RIAPA-AI is IFPRI's standard single-country RIAPA CGE model (Table 1).

Table 1. RIAPA CGE model

Scheme of flows in the economy



Note: The detailed explanation of the model equations can be found in Lofgren et al. (2002) and Diao and Thurlow (2012); the description of poverty and undernourishment microsimulation modules can be found in Arndt et al. (2012) and Pauw et al. (2023).

Mathematical functional forms

Category	Functional form
Factors	Constant Elasticity of Substitution (CES)
Intermediates	Leontief
Imports	CES (Armington)
Exports	Constant Elasticity of Transformation (CET)
Consumption	Linear Expenditure System
Numeraire	Consumer Price Index (CPI) or Domestic Producer Price Index (DPI) is fixed
Current account closure[2]	Rule: Foreign savings or nominal exchange rate must be fixed.
Government closure[1]	Rule: Government savings or direct tax rates are fixed
Savings-investment closure[1]	Rule: Investment-driven, savings-driven, or balanced absorption adjustments
Factor closures[1]	Fully employed and mobile across activities; fully employed but immobile across activities; or flexible factor supply with fixed wage.

[2] Mathematically, closure rules ensure that the system of equations is uniquely solvable. They establish the economic boundary conditions of the model, ensuring the underlying equations resolve into a mathematically well-behaved general equilibrium. Economically, closures define which variables in the model remain fixed (exogenous) and which are allowed to adjust (endogenous) to satisfy systemic identities, clear markets, and achieve a new equilibrium.

The RIAPA CGE model is calibrated to a country-specific Social Accounting Matrix (SAM) that represents the country's real economy in detail. For the empirical application in this paper, the model employs the 2019 SAM for Kenya (Thurlow, 2021; IFPRI, 2021) with the following structural features:

- Sectors and products: 46 production sectors and 46 corresponding commodities covering primary agriculture (12), mining (2), agro-processing (8), manufacturing (9), utilities (2), construction (1), private services (8), and public services (4), with mathematical functions for their production, exports, and imports.
- Households: 10 distinct population groups covering household incomes and expenditures by expenditure quintiles and rural and urban locations, with per capita demand functions for each of the 46 commodities.
- Factors: 8 distinct factors of production covering labor (4), capital (3), and agricultural land (1), with factor demand functions for production sectors and factor supplies acting as part of the economic closure variables.
- Accounts: Government and rest of the world accounts allowing for the simulation of fiscal policy and external shocks.

Correspondingly, to define economic simulations users can adjust a vast continuum of structural levers (exogenous variables and parameters):

- International commodity market prices.
- Productivity and technology, including total factor productivity (TFP) and factor-specific productivity parameters in production functions, as well as elasticities of CES and CET trade functions.
- Factor supplies and household groups' population.
- Fiscal policy and tariffs, including direct taxes on households and indirect tax rates such as sales taxes, import tariffs, production taxes, and subsidies.

- Macroeconomic aggregates and international terms of trade, including government consumption and foreign savings or the exchange rate.
- Economic closures that allow users to define which model variables remain fixed versus endogenous, including savings-investment balances, government balances, rest-of-world balances, and factor market clearing conditions.

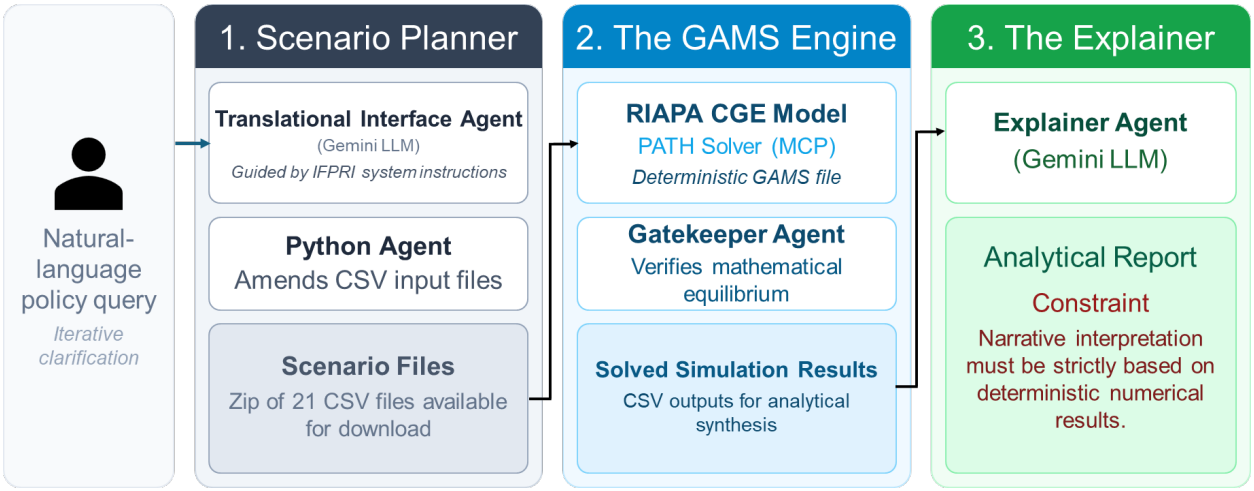
These model inputs are frequently applied in various economic simulations worldwide. The IFPRI standard CGE model is a workhorse tool in policy analysis across developing countries, utilized both by IFPRI (e.g., Arndt et al., 2012; Pauw et al., 2023; Arndt et al., 2023) and external researchers (e.g., Folsom, 2016; Shikur, 2020; Nechifor et al., 2022).

2.2. Architecture of RIAPA-AI

To evaluate the performance of RIAPA-AI, it is necessary to first establish its operational architecture. RIAPA-AI operates as an *agentic AI assistant system*—an integrated analytical pipeline consisting of a coordinated network of sub-agents that fuses a general-purpose LLM with a deterministic CGE model. Rather than operating as a standalone LLM predicting text based on training corpora, the RIAPA-AI agent functions as an autonomous wrapper around a hard-coded system of CGE economic equations.

A defining architectural principle of this workflow is the strict isolation of the LLM from the mathematical core. The CGE model is an immutable, deterministic structure of non-linear equations. The LLM does not alter, touch, or approximate this core logic. Instead, the general-purpose LLM acts exclusively as a translational interface on the input and output boundaries. The high-level workflow consists of three sequential components (Figure 2):

Figure 2. Operational architecture of the RIAPA-AI agentic modeling system



The workflow integrates LLMs with a CGE mathematical engine across three sequential phases. 1, The Scenario Planner translates natural-language policy queries into quantitative model inputs via a Gemini-powered interface. 2, The GAMS Engine acts as a cloud-based mathematical core, where a sub agent initiates the RIAPA model to solve for the new economic equilibrium. 3, The Explainer synthesizes the simulated results into a comprehensive analytical report, operating under strict system constraints to base its narrative solely on deterministic numerical outputs.

1. **Scenario Planner**: The pipeline begins when a user submits a natural-language economic or policy question (for example, asking to evaluate the impact of a 20 percent increase in the world import price of energy commodities). The goal of this first sub-agent is to leverage advanced reasoning capabilities (for details, see Wei et al., 2022), guided by an extensive set of IFPRI-developed system instructions, to translate the user's narrative into quantitative shocks on specific structural inputs. Because real-world policy questions can be ambiguous, the sub-agent acts as a translational interface; users may iterate with RIAPA-AI across multiple conversational turns (exchanges of prompts and responses) to clarify the specifications of the question or simulation scenario. After the user's final confirmation, the sub-agent executes Python code to manipulate the data arrays, finalizing precise mathematical adjustments across the structural inputs defined in Section 2.1.
2. **The GAMS Engine**: The scenario file is zipped and sent to a cloud-based General Algebraic Modeling System (GAMS) engine. GAMS is a high-level mathematical programming and optimization software system widely used in operations research. Within this environment, the CGE model is formulated as a Mixed Complementarity Problem (MCP) and is solved using the PATH optimization solver (Dirkse & Ferris, 1995). The AI sub-agent responsible for this part of

RIAPA-AI has no influence over the CGE solution process; it acts solely as an initiator and an automated gatekeeper, verifying that the model solved without mathematical constraint violations.

3. **The Explainer**: After the CGE model deterministically computes the new general equilibrium, the final sub-agent fetches the solved results and interprets them by analyzing the deviations of all economic variables relative to their baseline values. Manually tracing how a shock propagates across thousands of variables (from primary effects to general-equilibrium price changes to downstream welfare impacts) is the most time-intensive stage of CGE analysis and the one most vulnerable to omissions. The Explainer traverses this space exhaustively, organizing the deviations into a coherent causal narrative. Because this sub-agent is explicitly restricted via system instructions to base its narrative interpretation solely on the deterministic numerical results, RIAPA-AI mitigates the risk of logical hallucinations inherent in general-purpose AI.

Overall, this workflow is rigidly structured around the CGE model. While the AI manages the simulation, the system remains fully transparent: advanced users can view the underlying data at any time and ask the agent to adjust structural elasticities or model closure rules. Furthermore, all modeling artifacts—including the underlying GAMS code, the SAM, parameters, scenario files, and simulation results—are fully transparent and available for download. This replicability allows users to manually verify the AI's logic using a local desktop installation of GAMS. Users can independently import the AI-formulated scenario input files into a desktop CGE model, solve it, and verify that the numerical results match the AI-generated report exactly.

Besides the time and analytical efficiency gains (considered below), it should be noted that this architectural design democratizes economic modeling by dismantling the barriers of software-related financial costs and specialized technical skills. Traditionally, outfitting a team of economists with standalone CGE modeling software can be cost-prohibitive. A standard single-user desktop GAMS license, equipped with a commercial solver like PATH, costs thousands of dollars per individual workstation (GAMS Development Corp., 2024). Furthermore, utilizing this software requires programming expertise, forcing resource-constrained institutions worldwide to frequently rely on expensive external experts to

conduct professional, CGE-based policy analysis. In this context, an important feature of the RIAPA-AI architecture is its utilization of a GAMS cloud solution. Rather than requiring local GAMS software, users access the underlying CGE model through the intuitive natural-language interface of RIAPA-AI, circumventing the coding barrier entirely. The fixed infrastructure cost for the GAMS cloud is an annual access fee, which provides unrestricted access for an unlimited number of users, while running a simulation incurs only a marginal compute cost[3]. Overall, this operational shift empowers resource-constrained government agencies and research institutions to conduct computationally intensive, rigorous economic analysis independently, without requiring software licenses or relying on external experts with GAMS knowledge.

Finally, it is also important to distinguish this architecture from a generic AI assistant. The agent operates as a fully self-contained, end-to-end analytical pipeline with absolute replicability. Users are not required to provide underlying SAMs, specify mathematical functional forms, or write modeling code (all components necessary to run simulations via an LLM interface, such as the SAM and all model parameters, are already structurally embedded within the system). In this regard, RIAPA-AI is not designed to be "transportable" in the sense that users can upload arbitrary, non-standard SAMs for the agent to run simulations. Instead, the agent architecture is designed to expand to multiple countries that have standard RIAPA CGE models. Following the 2019 Kenyan pilot, our goal is to scale the system to encompass 40 distinct, IFPRI-standardized, vetted RIAPA-AI country models.

[3] During our empirical testing, executing a single CGE simulation costs $0.025 for the GAMS Engine and up to $0.85 for LLM token processing and Google Cloud infrastructure (depending on the intensity of the user session and follow-up questions).

3. EVALUATION METHODOLOGY

Because RIAPA-AI is a novel system, rigorous empirical validation is a prerequisite for its broader deployment. As discussed earlier, the evaluation of RIAPA-AI is driven by two objectives: (1) assessing the analytical quality and efficiency of the agent's performance against human CGE modelers, and (2) comparing RIAPA-AI to a general-purpose LLM to assess overall analytical effectiveness.

Standard AI benchmark tests, which typically focus on factual recall or reasoning over written text (for details, see, e.g., Wei et al., 2022), are inapplicable to RIAPA-AI validation because the agent is explicitly designed for quantitative economic analysis, not simply for generating economic narratives. Moreover, the numerical results of RIAPA-AI are deterministically produced by the CGE model. Validation must therefore assess structural consistency and scenario translation in addition to economic logic. Consequently, we constructed a custom evaluation methodology, centered on expert evaluation, to address both comparative objectives.

Because human evaluators have finite capacity limits, exhaustively testing the possible CGE scenario space using formal statistical methods (e.g., Latin hypercube sampling) was unfeasible. To ensure broad coverage of model inputs under these time constraints, we adopted a pragmatic approach: prioritizing distinct economic domains and focusing on overall directional vectors (utilizing standardized positive and negative shocks of 10 and 20 percent) rather than formal parameter space sampling. We compiled a sample of 80 distinct economic scenarios covering all available RIAPA economic inputs, grouped into three primary blocks:

- Group 1: Supply-Side and Productivity: Encompassing total factor productivity, factor-specific productivity across production sectors, and supply adjustments across all factors.
- Group 2: External Trade and Macro: Encompassing world prices, tariffs, exchange rate and foreign savings fluctuations, and government consumption.
- Group 3: Fiscal Policy and Others: Encompassing direct and indirect taxes, production and factor subsidies, and other inputs.

IFPRI's CGE modelers were separated into two groups to execute this test: eight CGE modelers functioned as the composite human baseline, while a separate conclave of three senior experts functioned as the evaluators. The baseline group was specifically structured to represent an advanced and highly capable institutional user base; half of the participating modelers hold Ph.D. degrees in economics, and, on average, group members have over 10 years of professional CGE modeling experience. The evaluation conclave consisted of one active professor of economics, one professor emeritus, and one senior research fellow at IFPRI, with an average of over 30 years of professional CGE modeling experience[4].

To answer their assigned ten economic questions, each human modeler utilized the desktop version of the RIAPA model in GAMS to produce a simulation report detailing the specific shocked inputs used, the simulation result tables, and a brief analytical summary. Ultimately, each agent (the human IFPRI modelers, RIAPA-AI, and the general-purpose LLM) answered the scenario questions. The expert panel then assessed the quality of these outputs across three dimensions: (i) *Correctness* (i.e., understanding the questions consistently with economic frameworks); (ii) *Depth and trade-offs* (i.e., tracing the economic shock to identify relative winners and losers); and (iii) *Objectivity* (i.e., evidence-based analysis and avoidance of ungrounded or normative speculation).

Unlike RIAPA-AI and the human modelers, the general-purpose LLM lacks a CGE model. To simulate a naive user interaction, the baseline Gemini 3 Pro model was prompted with standard inquiries in the format: "What is the economic and household impact of the following shock in Kenya: [insert scenario text]?". Therefore, to ensure a fair comparison, the evaluation criteria were leveled to focus on validity, economic sense, and mechanism tracing rather than exact numerical matching. Response outputs were graded using a categorical ordinal scale:

- 1.0 (Robust / Policy-Ready): The logic is sound, comprehensively traced, and meaningful for policy analysis.

[4] All experimental protocols were approved by the IFPRI Institutional Review Board (IRB protocol FPM-26-0409).

- 0.5 (Caution / Incomplete): Contains correct high-level directions but lacks necessary economic depth or omits secondary linkages.
- 0.0 (Fail / Danger): Hallucinates or provides incorrect directional vectors; policy advisory is fundamentally flawed or highly misleading.

To eliminate subjective bias, the evaluation panel of three experts operated under a strict conclave system requiring unanimous agreement to finalize every score. To quantify the variance in performance on this ordinal scale, we deployed two non-parametric statistical tests (Agresti, 2013). First, to assess absolute reliability, the data was binarized into "Policy Ready" (Score = 1) versus "Requires Intervention" (Score < 1). We then applied McNemar's Exact Binomial Test for matched pairs to calculate the mathematical probability of the observed variance in failure frequency. Second, to measure the overall severity of the analytical divergence, we utilized the Wilcoxon Signed-Rank Test, which evaluates the full ordinal variance (1, 0.5, 0). The complete catalog of simulation questions, the answers generated by all agents, and the expert grading rubrics are publicly archived and available for download via Mendeley Data (Mukashov, 2026).

4. RESULTS

The aggregate evaluation results demonstrate that reliance on a general-purpose LLM for complex economic analysis introduces severe, statistically significant analytical risks. Conversely, the CGE-backed RIAPA-AI agent achieves statistical parity with human CGE experts in analytical accuracy while substantially outperforming them in operational efficiency—drastically reducing execution time.

Table 2. Summary of Performances Across Evaluation Dimensions

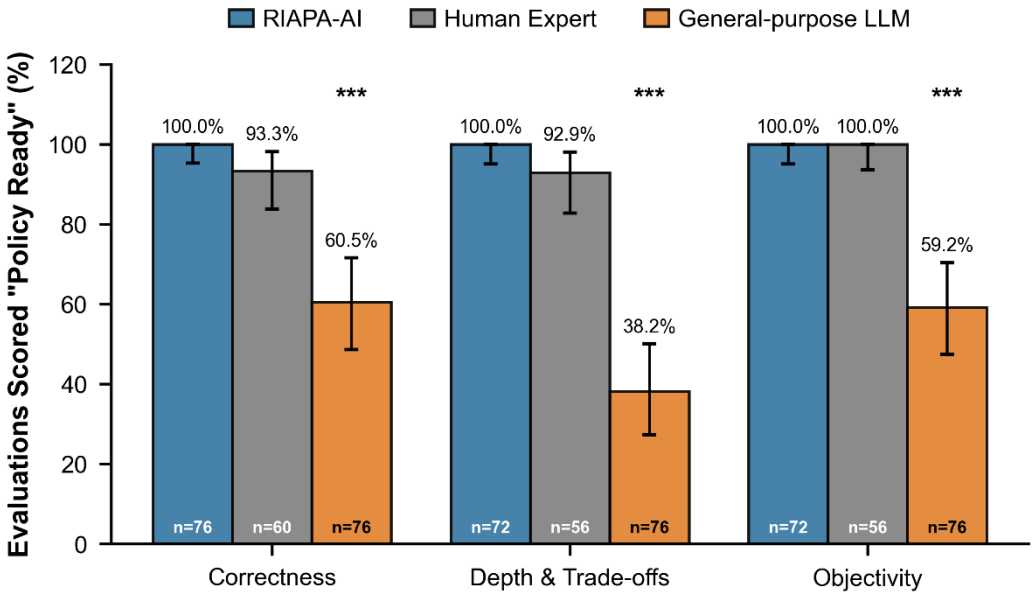
Agent	Dimension	"Policy Ready" (Score=1.0)	*n*	McNemar's Exact *p* (Isolates Binary Pass/Fail Discordance)	Wilcoxon Signed-Rank *p* (Evaluates Ordinal Magnitude of Failure)
RIAPA-AI	Correctness	100.00%	76	-	-
	Depth & Trade-offs	100.00%	72	-	-
	Objectivity	100.00%	72	-	-
General-purpose LLM	Correctness	60.50%	76	< 0.001***	< 0.001***
	Depth & Trade-offs	38.20%	76	< 0.001***	< 0.001***
	Objectivity	59.20%	76	< 0.001***	< 0.001***
Human Expert	Correctness	93.30%	60	0.125 (n.s.)	0.046*
	Depth & Trade-offs	92.90%	56	0.125 (n.s.)	0.046*
	Objectivity	100.00%	56	1.000 (n.s.)	1.000 (n.s.)

*Note: The variation in sample size (n) reflects specific scenario constraints (e.g., scenarios targeting non-existent export sectors where depth/objectivity metrics become inapplicable) and operational limits during human evaluation. Significance levels: ***p < 0.001; *p < 0.05; n.s. = not significant. P-values derived from exact binomial testing of discordant pairs (McNemar's) and paired ordinal variance (Wilcoxon).*

4.1. Overall Analytical Performance

Across the evaluation matrix, the CGE-backed RIAPA-AI agent maintained a perfect record, achieving a 1.0 ("Policy Ready") score in 100% of the tested scenarios (100% point estimate; 95% exact binomial confidence interval (CI): 95.3–100.0%) while completing comprehensive 10- to 12-page analytical reports in approximately two minutes per simulation (Figure 3). This performance stems from the architecture's reliance on a deterministic underlying core; by strictly limiting the agent's textual explainer to parsing processed numerical arrays from the CGE model, logic hallucinations are prevented. It should be emphasized that this 100% reliability reflects technical precision within the established boundaries of our scenario matrix, rather than open-ended infallibility across unvetted policy questions.

Figure 3. Comparative analytical performance across the core evaluation matrix.



The grouped bar chart visualizes the percentage of evaluations graded as "Policy-Ready" (Score = 1.0 on a three-tier categorical ordinal scale) across three testing dimensions: Correctness, Depth & Trade-offs, and Objectivity. The CGE-backed RIAPA-AI agent maintained a 100% policy-ready score across all domains, successfully operating without the manual processing limits that drove a 6.7% error rate among human modeling experts. In contrast, the general-purpose LLM exhibited severe failure rates when forced to navigate complex structural scenarios without an underlying mathematical engine. Sample sizes (n) for each evaluation dimension are denoted at the base of the bars. Asterisks indicate the statistical significance of the general-purpose LLM's failure rates compared to RIAPA-AI, calculated via McNemar's Exact Binomial Test for binary pass/fail discordance (*** $p < 0.001$).

In stark contrast, the general-purpose LLM exhibited high failure rates when forced to navigate complex economic questions without a CGE model. It achieved a policy-ready score in only 60.5% of Correctness evaluations (95% CI: 48.6–71.6%) and 38.2% of *Depth and Trade-offs* evaluations. The overlapping confidence intervals between RIAPA-AI and humans statistically confirm parity in their performance limits. As established in Section 3, we evaluated this divergence using both McNemar's Exact Binomial Test (which isolates binary pass/fail discordance) and the Wilcoxon Signed-Rank Test (which evaluates the ordinal magnitude of the failure on our 1, 0.5, and 0 grading scale). The results confirm that RIAPA-AI is significantly superior to the general-purpose LLM across all metrics in both failure frequency (McNemar's $p < 0.001$) and failure severity (Wilcoxon $p < 0.001$). A qualitative review revealed that rather than just omitting secondary effects (scoring 0.5), the general-purpose LLM frequently injected unmodeled socio-political narratives to fill analytical gaps, resulting in absolute failures (scoring 0). For example, in crop

failure simulations, the general-purpose LLM hallucinated that rural households would begin "pulling children out of school because parents can no longer afford fees."

When compared to human experts, RIAPA-AI achieved statistical parity in binary error frequency but demonstrated a statistically significant advantage in error severity. The human CGE modelers recorded deviations in 4 out of 60 evaluated instances, yielding a 6.7% error rate (93.3% accuracy point estimate; 95% CI: 83.8–98.2%). While McNemar's test yielded a p-value of 0.125 (indicating no statistically significant difference in the raw frequency of failures), the Wilcoxon Signed-Rank test yielded a p-value of 0.046 for both *Correctness* and *Depth* (see Table 2). This indicates that while humans do not fail significantly more often than the agent, the magnitude of the variance in ordinal scores is statistically significant when compared to RIAPA-AI's performance.

This divergence is rooted in operational constraints and fatigue. The human errors were characterized by minor omissions of secondary variables (scores of 0.5), driven almost entirely by the inherent constraints in manually processing thousands of economic variables during the complex, multi-step analysis of 10 simulation scenarios per IFPRI CGE modeler. Crucially, this specific constraint is largely an artifact of the high-volume testing environment required for statistical validation; in standard, high-stakes research environments, economists spend days or weeks iterating on a single simulation scenario to eliminate such omissions. Nonetheless, the comparison underscores the AI's operational advantage: operating at computational speed, RIAPA-AI exhaustively traced every important shift, generating comprehensive reports detailing economic, distributional, and poverty impacts in approximately two minutes. A full discussion of these human bottlenecks and their implications for policy analysis is provided in Section 5.

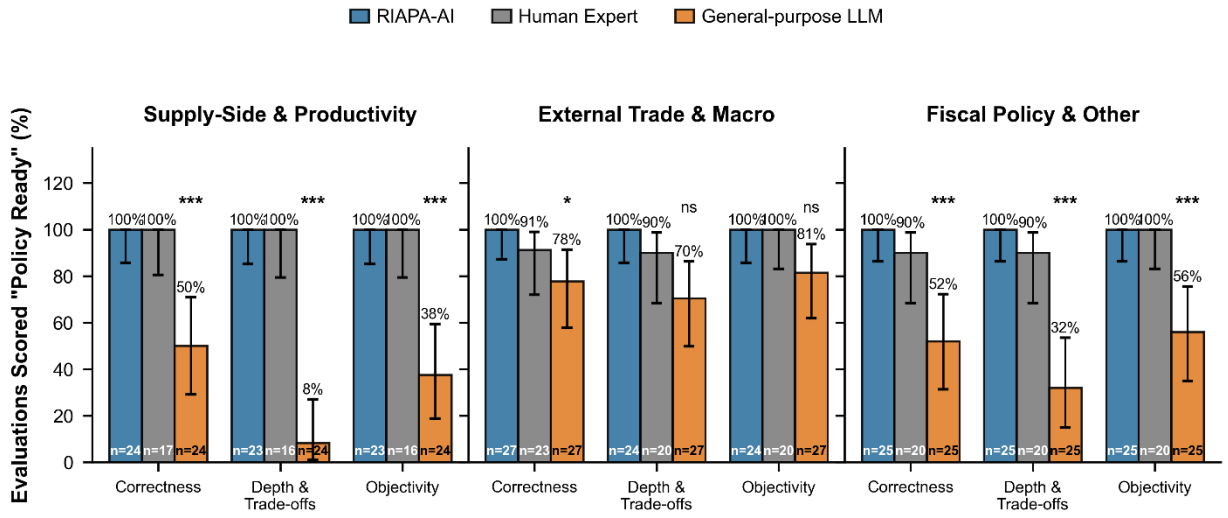
4.2. Performance Variability Across Economic Domains

Disaggregating the data into core economic domains further illustrates the specific structural environments where the general-purpose LLM fails, reinforcing the necessity of a CGE model (Figure 4).

- Supply-Side and Productivity: Within this domain, the general-purpose LLM failed significantly in measuring specific factor constraints (*Correctness*: McNemar's $p < 0.001$; Wilcoxon $p < 0.01$). The general-purpose LLM routinely omitted trade-offs. For example, during simulated positive agricultural productivity shocks, the general-purpose LLM incorrectly assumed a simple relationship between agricultural productivity and rural income growth. It consistently failed to consider the inelastic-demand responses that cause producer prices to fall when supply increases, demonstrating a lack of general equilibrium capability for assessing interactions between different economic forces. By contrast, human omissions in this domain were strictly logistical, such as failing to report secondary cross-sectoral shifts due to the sheer volume of data required for manual extraction. RIAPA-AI bridged this gap entirely, capturing both the structural logic and the exhaustive data extraction.
- External Trade and Macro: The general-purpose LLM performed its best in this category. Open-economy macroeconomics relies heavily on standard, predictable directional vectors, such as currency depreciation boosting exports or global commodity price spikes driving imported inflation. Because these mechanics are ubiquitous in standard economic textbooks, the general-purpose LLM successfully leveraged its foundational training to predict high-level trade shifts. However, while it accurately mapped these broad trends, RIAPA-AI maintained statistical superiority in depth (Wilcoxon $p < 0.05$) because the general-purpose LLM lacked the structural matrices required to identify which specific household groups would benefit or suffer from these trade shifts. Nonetheless, the general-purpose LLM matched RIAPA-AI in *Objectivity* (McNemar's $p = 0.063$, n.s.).
- Fiscal Policy and Other: The general-purpose LLM proved structurally incapable of accounting for systemic balancing mechanisms, such as tax recycling and government investment dynamics. It frequently operated under the flawed assumption that fiscal policy shocks exclusively affect households' disposable incomes, entirely omitting government revenue and expenditure dynamics in the general equilibrium. Consequently, RIAPA-AI vastly outperformed the general-purpose

LLM in both *Correctness* (Wilcoxon $p < 0.001$) and *Depth* (Wilcoxon $p < 0.001$) by perfectly tracing the general equilibrium impacts of tax incidence and capital distributions.

Figure 4. Analytical performance disaggregated by core economic domain.



The multi-panel bar chart illustrates the variance in evaluation scores ("Policy-Ready") across three distinct structural clusters of simulation questions: Supply-Side & Productivity, External Trade & Macro, and Fiscal Policy & Other. The data isolates the specific economic environments where the general-purpose LLM fails; while it successfully anticipates standard directional vectors in open-economy trade scenarios, it exhibits severe failure rates in the Supply-Side and Fiscal domains where it lacks the mathematical capacity to map rigid structural constraints or complex macroeconomic closures (e.g., tax recycling and investment crowding-out). Conversely, RIAPA-AI demonstrates absolute statistical parity with human experts across all three domains, effectively operating without the human omissions driven by data exhaustion in complex, multi-step scenarios. Sample sizes (n) are denoted at the base of the bars. Significance levels for differences in failure rates between the general-purpose LLM and RIAPA-AI are calculated via McNemar's Exact Binomial Test (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, ns = not significant).

5. DISCUSSION AND CONCLUSION

Analytical Risks of General-purpose LLM Use in Economic Analysis

The statistical breakdown reveals a fundamental flaw in utilizing general-purpose LLMs for economic and policy analysis. While the general-purpose LLM grasps standard economic theory to anticipate directional impacts, it is inherently incapable of navigating country-specific economic structures. When scenarios involve multi-dimensional shocks with divergent trajectories (e.g., producers versus consumers), the general-purpose LLM defaults to generic, linear assumptions. Consequently, its answers are frequently incorrect.

This lack of a formal model is particularly problematic for country-specific policy formulation. For example, during our evaluations of the Kenyan economy, we asked the models to simulate a 20 percent decrease in the productivity of the cereals sector (acting as a proxy for a severe national drought). During our test, the general-purpose LLM performed a back-of-the-envelope calculation of the impact on national GDP, estimating a contraction between 1 and 2 percentage points, which is not far off from the RIAPA-AI estimate of a 0.83 percent contraction. However, the general-purpose LLM did not provide further numeric assessments, and all its subsequent assessments on sectoral and household welfare impacts were purely qualitative. Most importantly, the general-purpose LLM predicted a crisis that would first and foremost devastate rural household welfare. Conversely, RIAPA-AI revealed a numerically backed, nuanced general equilibrium reality: the supply contraction caused food prices to increase. Because rural Kenyan households are net sellers of crops, this price surge buffered some of their income losses, and it was the urban poor—who are net food consumers—who bore the brunt of the shock. Because the general-purpose LLM lacks a model to account for such nuances, it generated a flawed analysis. Furthermore, to fill these analytical gaps, the general-purpose LLM injected unmodeled narratives, such as the possibility of farmers pulling children out of school to save money for food. While the general-purpose LLM may be useful for initial brainstorming, final policy design requires absolute quantitative precision. Utilizing these imprecise assessments for policy formulation introduces significant analytical risk.

Overcoming Human Bottlenecks

The comparison between RIAPA-AI and human CGE modelers yields important operational insights. Statistically, both human experts and RIAPA-AI possess an identical grasp of the CGE framework. However, a standard CGE model features a vast array of structural inputs, a continuum of possible input values, and thousands of economic variables. In our empirical tests, human modelers spent up to 75 minutes per simulation. Their 6.7% error rate was understandably driven by cognitive fatigue, manifesting as the omission of secondary variables in complex scenarios (e.g., failing to include petroleum commodities as part of a global energy price shock scenario). At the same time, it is important to note that this human fatigue factor is specific to the high-volume CGE simulations we asked IFPRI CGE modelers to conduct

for this study. In the reality of high-stakes policy analysis, economists spend days or weeks refining scenarios via dynamic trial-and-error in desktop GAMS software, deeply investigating complex causal chains; in those settings, cognitive fatigue factors most likely do not apply.

However, it is also undeniable that RIAPA-AI processes data at a speed incomparable to human IFPRI CGE modelers. In approximately two minutes, RIAPA-AI reads a natural language question, formulates a simulation scenario confirmed by the user, runs and solves the CGE model across all economic variables, and generates a comprehensive report. This report summarizes the model results by quantifying economywide impacts on total GDP, sectoral GDP, and other economic variables, alongside distributional impacts across household groups and micro-level impacts on poverty, hunger, and other development outcomes. The summary report also clearly identifies relative winners, losers, and trade-offs resulting from the simulated shock, providing critical analytical information frequently requested by governments in low- and middle-income countries. The AI agent consistently navigates these complex relationships on both the input and output sides of the CGE model, reducing the possibility of error in CGE analysis.

Conclusion

The empirical evidence strongly indicates that a general-purpose LLM lacking an underlying model calibrated to real country data cannot safely replace formal economic modeling. In the case of RIAPA-AI, a CGE model provides the structured numerical constraints required to generate robust analysis. By layering a generative AI interface over a formal CGE model, RIAPA-AI bridges the gap between LLM accessibility and structural modeling rigor.

Ultimately, for high-stakes policy questions, human auditing of RIAPA-AI output remains highly recommended. While RIAPA-AI achieved a 100% policy-ready score across our specific evaluation matrix, these results should be interpreted cautiously; they reflect performance within tested boundaries. Cases of unpredictable LLM behavior are documented in the AI literature. These include performance degradation over time due to API updates, known as model drift (Singh et al., 2025), and the “lost in the middle” phenomenon, where models fail to retrieve critical information buried within long context windows (Liu et al., 2024). In this context, it is important to reemphasize that, by design, RIAPA-AI is a transparent tool

built for replicability. All underlying artifacts—the scenario zip file, exact GAMS code, and raw simulation results—can be downloaded and audited at every step of the simulation analysis. While we did not formally estimate the time gains between the two *modi operandi* (i.e., a CGE modeler working from scratch versus a CGE modeler auditing RIAPA-AI), we can reasonably speculate that reviewing AI-generated artifacts demands significantly less time than building them from scratch.

Furthermore, it is important to note that the formulation of a valid simulation scenario is itself a field where economists with deep knowledge of a country's context and the CGE model's domain of applicability remain essential. In this study, our goal was to cast a wide net, prioritizing broad coverage of distinct economic domains and focusing on overall directional vectors (utilizing standardized positive and negative shocks of 10 and 20 percent) rather than exhaustive parameter space sampling. In practice, CGE modelers and country experts spend considerable time formulating the magnitude of scenarios by examining historical data and policy discussions. Consequently, while RIAPA-AI proves exceptionally capable of executing clearly formulated scenarios, the validity of the scenarios themselves can still be subject to scrutiny. In this context, as a next step in RIAPA-AI's development, we are developing a Scenario Facilitator engine. This module will assist users through structured, multi-turn conversations to provide precise, scientifically grounded, and verifiable scenario quantification that leverages available country-level data. However, at this point, the validation of modeling scenarios (and therefore, modeling results) must still be overseen by professional economists and CGE modelers.

Even with this current need for human oversight in scenario design, RIAPA-AI already represents a critical step toward equity in access to data, models, and rigorous policy decision-making. Historically, outfitting a team of economists with advanced programming skills and modeling software has been highly cost-prohibitive, and RIAPA-AI dismantles these dual barriers of coding and cost. Furthermore, the agent intrinsically encodes the best practices of senior IFPRI modelers for both building scenarios and interpreting complex outputs. Looking forward, this framework has the potential to fundamentally transform economic analysis in resource-constrained environments. For analysts and non-modelers in low- and middle-income countries, it offers speed and ease of use, allowing them to jump into structural modeling directly via natural

language. Ultimately, this approach paves the way toward integrated, real-time decision-making, where AI and structural economic modeling work in tandem to deliver rapid, high-fidelity insights during emerging national and global crises. In this context, RIAPA-AI is an important step toward democratizing high-quality, reproducible economic analysis for policymakers and researchers in low- and middle-income countries.

Software Access Note

To ensure infrastructure sustainability, RIAPA-AI currently operates under a managed access model. Because running the cloud-based mathematical engine and processing LLM tokens incurs costs, access is prioritized for research and policy evaluation. Researchers and institutions interested in getting access may request login credentials by contacting the project team at IFPRI-FPM-RIAPA-AI@groups.cgiar.org.

REFERENCES

Agresti, A. (2013). *Categorical data analysis* (3rd ed.). John Wiley & Sons.

Arndt, C., Diao, X., Dorosh, P., Pauw, K., & Thurlow, J. (2023). The Ukraine war and rising commodity prices: Implications for developing countries. *Global Food Security*, 36, 100680. https://doi.org/10.1016/j.gfs.2023.100680

Arndt, C., Hussain, M. A., Jones, E. S., Nhate, V., Tarp, F., & Thurlow, J. (2012). Explaining the evolution of poverty: the case of Mozambique. *American Journal of Agricultural Economics*, 94(4), 854-872.

Böhringer, C., Rutherford, T., & Wiegard, W. (2003). Computable general equilibrium analysis: Opening a black box. *ZEW Discussion Papers*, Zentrum für Europäische Wirtschaftsforschung, Mannheim. https://www.econstor.eu/bitstream/10419/130210/1/85663686X.pdf

Diao, X., & Thurlow, J. (2012). A recursive dynamic computable general equilibrium model. In X. Diao, J. Thurlow, S. Benin, & S. Fan (Eds.), Strategies and priorities for African agriculture: economywide perspectives from country studies (pp. 17-50). International Food Policy Research Institute.

Dirkse, S. P., & Ferris, M. C. (1995). The PATH solver: a non-monotone stabilization scheme for mixed complementarity problems. Optimization Methods and Software, 5(2), 123-156.

EPA. (2017). *SAB Advice on the Use of EconomyWide Models in Evaluating the Social Costs, Benefits, and Economic Impacts of Air Regulations*. Tech. Rep., US Environmental Protection Agency, Washington. https://nepis.epa.gov/Exe/ZyPDF.cgi/P100X9WI.PDF?Dockey=P100X9WI.PDF

Folsom, B. (2016). The effect of easing traffic congestion in the Philippines using the Standard CGE model. Presented at the *19th Annual Conference on Global Economic Analysis*, Washington DC. Purdue University, Center for Global Trade Analysis. https://www.gtap.agecon.purdue.edu/resources/res_display.asp?RecordID=5128

GAMS Development Corp. (2024). GAMS Commercial Price List. Retrieved from https://www.gams.com/sales/pricing/

International Food Policy Research Institute (IFPRI). (2021). *2019 Social Accounting Matrix for Kenya*. Harvard Dataverse. https://doi.org/10.7910/DVN/ALUXSI

International Food Policy Research Institute (IFPRI). (2025). RIAPA Data and Modeling System project. https://www.ifpri.org/project/riapa model/

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., & Liang, P. (2024). Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics*, *12*, 157–173. https://doi.org/10.1162/tacl_a_00638

Löfgren, H., Harris, R. L., & Robinson, S. (2002). A Standard Computable General Equilibrium (CGE) Model in GAMS. *Microcomputers in Policy Research 5*, International Food Policy Research Institute.

Mukashov, Askar (2026), "An agentic AI assistant for country-level economic modeling: evaluation data", Mendeley Data, V1, doi: 10.17632/f8mbb6cykv.1

Nechifor, V., Basheer, M., Calzadilla, A., Obuobie, E., & Harou, J. J. (2022). Financing national scale energy projects in developing countries – An economy-wide evaluation of Ghana's Bui Dam. *Energy Economics*, 111, 106065. https://doi.org/10.1016/j.eneco.2022.106065

Pauw, K., Ecker, O., Thurlow, J., & Comstock, A. R. (2023). Measuring changes in diet deprivation: New indicators and methods. *Food Policy*, 117, 102471. https://doi.org/10.1016/j.foodpol.2023.102471

Shikur, Z. H. (2020). Agricultural policies, agricultural production and rural households' welfare in Ethiopia. *Economic Structures*, 9, 50. https://doi.org/10.1186/s40008-020-00228-y

Singh, H., Xia, F., Gossmann, A., Chuang, A., Hong, J. C., & Feng, J. (2025). "Who experiences large model decay and why?" A Hierarchical Framework for Diagnosing Heterogeneous Performance Drift. *arXiv*. https://doi.org/10.48550/arxiv.2506.00756

Thurlow, J. (2021). *2019 Social Accounting Matrix for Kenya: A Nexus Project SAM*. https://doi.org/10.2499/p15738coll2.134819

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems, 35, 24824-24837.

ALL IFPRI DISCUSSION PAPERS

All discussion papers are available **here**

They can be downloaded free of charge

INTERNATIONAL FOOD POLICY RESEARCH INSTITUTE
www.ifpri.org

IFPRI HEADQUARTERS
1201 Eye Street, NW
Washington, DC 20005 USA
Tel.: +1-202-862-5600
Fax: +1-202-862-5606
Email: ifpri@cgiar.org