Stata-MCP: A Research Report on AI-Assisted Empirical Research

TAN Song[1,2]， Feng Mu-yao[1,2] (Sep. 21 2025 New!)

1. School of Economics, Shanghai University, Shanghai 200072, China

2. Shanghai Bayes Views Co., Ltd., Shanghai 200444, China

Abstract: Data analysis is central to social science research, yet technical barriers still limit efficiency. While Stata lowers the learning cost of coding, further simplification is possible. This study develops Stata-MCP to streamline coding via AI and promote automation in empirical work. Using the ReAct agent framework and a custom evaluation set, we systematically assess LLMs across key stages of empirical research. Results show: (1) LLMs perform well in Stata code generation and debugging; (2) models differ notably in econometric understanding and coding, with the Claude series leading; (3) whether LLMs can conduct empirical analysis independently remains open. We also identify two key technical bottlenecks: (1) improving context management and overcoming window limitations through engineering; (2) fine-tuning the ReAct framework to better align with the needs of social science research.

Keywords: AI; Social science research; Stata; Empirical Strategy

Authors Notes: TAN Song, corresponding author (sepinetam@gmail.com), is affiliated with the School of Economics, Shanghai University, and CEO of Shanghai Bayes Views Co., Ltd. The authors acknowledge StataCorp for providing test licenses and support during the early stage of the project; we also thank the Shanghai University, the UM-SJTU Joint Insitute at Shanghai Jiao Tong Univeristy, PyCon China, and other platforms for offering opportunities for academic exchange that contributed to improving this project. The authors are especially grateful to Prof. Lian Yujun of Sun Yat-sen University for his valuable advice and guidance. As this work represents an internal research report rather than a formal academic paper, the authors welcome critical feedback and suggestions for improvement. The author commits to releasing all of the code of the project as fully open soure under the Apache-2.0 License on GitHub (https://github.com/sepinetam/stata-mcp). The related evaluation data and research reports will also be made publicly available within the project. For getting more information or tracing the newest research, please visit the project website (https://www.statamcp.com).

1 Introduction

The rapid advancement of artificial intelligence has opened up new opportunities across various industries and disciplines. In the natural sciences, several academic research paradigms have been fundamentally transformed. For instance, Google's AlphaFold compresses the inference from sequence to structure into a matter of minutes, drastically shortening research cycles under reproducible and accurate conditions [1], while its accompanying database—covering over 214 million predicted structures—has restructured workflows in structural biology [2]. In weather forecasting, AI systems such as GraphCast can generate 10-day global forecasts within one-minute, outperforming HRES on most evaluation metrics [3]; FourCastNet demonstrates orders-of-magnitude speed advantages over IFS, enabling large-scale ensemble and near real-time applications [4]; and GenCast surpasses ECMWF ENS in probabilistic forecasting with significant speed gains [5]. In the domain of PDE solving, Physics-Informed Neural Networks (PINNs) offer a unified framework for physics-constrained learning [6]. Subsequent developments in operator learning—such as the Fourier Neural Operator (FNO)—further elevate the paradigm from "solving equations" to "learning solution operators," achieving up to three orders of magnitude in speed improvements across various benchmarks [7]. However, to date, no comparable paradigm shift has occurred within the social sciences. In light of this gap, we introduce Stata-MCP [8], a framework designed to support and potentially transform research practices in the social sciences.

Stata-MCP is an open-source project designed for Stata users, aiming to help them leverage the power of AI to enhance research efficiency and seamlessly integrate AI into their academic workflows. Simply put, Stata-MCP bridges the gap between Stata and AI tools, enabling AI agents to interact directly with Stata in order to assist users in completing tasks more efficiently. Currently, Stata-MCP offers six tools and four prompts, supporting functionalities such as writing, appending and executing do-files, viewing log files, reading and displaying images, and accessing help documentation for Stata commands.

We developed a specialized evaluation database to examine whether Stata-MCP can effectively handle the coding aspects of empirical research. To minimize variability caused by different model providers, we conducted the tests using several models accessed via OpenRouter, a unified model provider platform. The final evaluation results were then scored and annotated using our internally

developed MASA System. Based on the examination results, we find that the Claude series of models already demonstrate sufficient proficiency in Stata coding and econometric model comprehension to be employed as practical tools in empirical research.

This study, through a systematic evaluation, finds that LLMs exhibit remarkable capabilities in Stata programming tasks: not only can they generate syntactically correct Stata code based on natural language instructions and autonomously detect and correct errors, but they are also capable of independently consulting help documentation, organizing and sequencing multi-step operations, and completing the full workflow—from data cleaning, processing, and importation to modeling and output—in complex tasks. Moreover, under the ReAct framework, the models demonstrate strong task scheduling and process management abilities, dynamically adjusting execution strategies to ensure systematic and coherent task completion.

However, the study also reveals critical limitations. First, despite the conceptual distinction between "thinking" and "non-thinking" models, empirical results indicate no statistically significant performance difference between the two, highlighting that current LLMs remain fundamentally predictive models and perform poorly in open-ended tasks such as research design. Second, context window limitations emerge as a key bottleneck: as task complexity increases and context length approaches model limits, interruptions and performance degradation become frequent. This issue is prevalent across models and points to the need for future research to enhance task efficiency through improved context engineering.

The rest of hte paper is organized as follows. Section 2 introduces the policy in China and the recent advancements in LLMs and related capabilities. Section 2 also introduces the existing literature on the application of AI in economics, highlighting current discussions and limitations in this area. Section 3 reports our research methodology and how this paper to evaluate the model’s capabilities using Stata-MCP. The data source and variables explanations in this paper also reports in section 3. Section 4 discusses the main evaluation results and section 5 analyzes the reasons behind this and discusses the limitations of generative AI in academic research as well as the issues we identified. In section 7 concludes.

2 Background

2.1 Policy in China

In August 2025, the China State Council issued the Opinions on Deepening the Implementation of the "Guiding Opinions on Further Advancing the AI Plus Initiative" (hereinafter referred to as the "Opinions"), which emphasized the enabling role of artificial intelligence (AI) in scientific research, social governance, and economic decision-making. The document calls for accelerating the pace of scientific discovery, promoting innovation in research and development models, and enhancing the overall efficiency of scientific inquiry. Notably, the Opinions explicitly advocate the exploration of AI-assisted research methodologies in the humanities and social sciences to improve the systematization, reproducibility, and intelligence of academic research.

In this era of integrating AI with disciplinary research—and under the broader policy framework set by national initiatives such as the New Generation Artificial Intelligence Development Plan, the AI Innovation Action Plan for Higher Education Institutions, and the NSFC Special Program on Intelligent Methods and Tools—there is clear encouragement for AI to empower scientific research, education, and social science inquiry. These policies collectively aim to advance the intelligent, reproducible, and efficient transformation of data analysis and research methodologies.

Against this backdrop, academic research—especially in the fields of philosophy and the social sciences—is undergoing a methodological transformation. Traditional statistical and empirical approaches are giving way to AI-driven tools that enable automated data processing, intelligent analytics, and interpretable results. Leveraging this policy momentum, the product aligns closely with the Opinions' call for "AI-driven scientific research platforms" and the goals of "enhancing research efficiency and promoting methodological innovation." It offers automated services for data cleaning, statistical analysis, regression modeling, visualization, and report generation—tailored to the needs of scholars and students in disciplines such as economics and sociology.

In line with national policy priorities and evolving market demands, Stata-MCP holds a first-mover advantage amid the growing wave of intelligent academic research. As the benefits of supportive policies are realized and the digitalization of education and research accelerates, the product is well-positioned to capture substantial growth opportunities in the scientific research tools

and services market.

2.2 Generative AI

Since Google proposed the Transformer architecture [9] and OpenAI released ChatGPT-3.5 at the end of 2022, generative AI (GenAI) represented by large language models (LLMs), has truly entered the public spotlight [10]. GenAI has brought convenience across various domains, and a number of studies have reported improved efficiency with the assistance and support of GenAI [11,12]. Despite the ongoing debates surrounding AI ethics, its potential in advancing academic research should not be overlooked [13–15].

Recent years, OpenAI and Google DeepMind' s tech-report discussed that the frontier and ability for large language models [16–19]. The capabilities of LLMs particularly in coding have been advancing rapidly, with multiple studies emphasizing their strong performance in tasks that are standardized and structurally well-defined [20,21].

In order to assess the current strength of AI's coding capabilities, I collected an open-source benchmark dataset from Hugging Face and conducted a visualization focusing on coding performance. Figure 1 illustrates how AI's coding ability has evolved over time. It is evident that AI has been consistently pushing the boundaries of what is possible in coding.

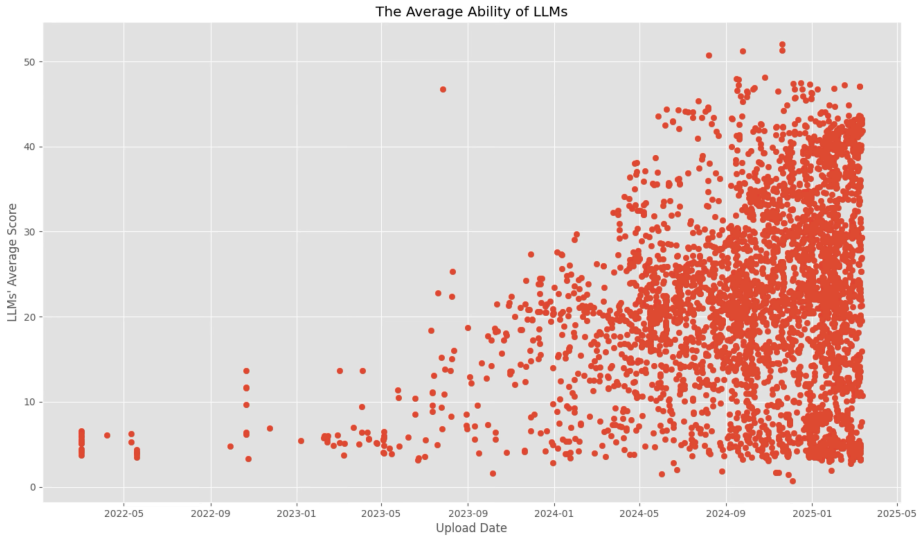

Figure 1. The coding ability of LLMs with time

2.3 MCP

In November 2024, Anthropic released the Model Context Protocol (MCP), which has since

become an industry standard. As an open-source specification, MCP aims to standardize the interaction between large language models (LLMs) and external tools. Its core objective is to establish a unified interface for communication between LLMs and external systems. By standardizing tool invocation, data source integration, and prompt template management, MCP addresses longstanding issues of inconsistency and fragmentation across different model ecosystems. At the industry level, MCP lays the foundation for interoperability between tools and models, promoting cross-vendor compatibility and resource reuse. As such, MCP not only enhances the scalability and maintainability of LLM applications but is also regarded as a major milestone in the standardization of agent-based and model-driven systems at the industrial scale.

Before MCP's emergence and widespread adoption, major AI vendors relied on proprietary standards for tool invocation, like function calling, tools and model extensions, resulting in significant variation in interface semantics and data formats. The primary challenge stemmed from the heterogeneity and siloed nature of model ecosystems: differences in tool registration, parameter declaration, and invocation mechanisms across models required redundant adaptation during migration. Moreover, tools developed for one ecosystem were often incompatible with others, limiting reuse. Against this backdrop, there was strong demand within the industry for a unified, cross-model standard—precisely the need that MCP was designed to address.

MCP, as an open protocol, streamlines the integration of LLMs with external tools and data sources, resolving the long-standing issue of fragmented tool invocation practices. By providing a standardized interface, MCP unifies tool registration, parameter declaration, and invocation procedures, thereby eliminating the need for repeated vendor-specific adaptations. It also breaks down ecosystem barriers, enabling tools developed for a single model to be reused across others, significantly reducing the cost of cross-platform migration. Furthermore, MCP offers a standardized framework for exposing external data sources and prompt templates, allowing models to access and execute external resources within a consistent contextual environment. These capabilities not only enhance the portability and maintainability of tool-based workflows but also provide a more robust foundation for production-grade LLM applications. Notably, MCP has made it feasible to operate Stata via multiple LLMs.

Stata-MCP represents a novel integration of AI and Stata. It is an open-source project released by Song Tan in March 2025. By adopting MCP’s unified invocation framework, Stata-MCP enables

large language models to directly generate and execute Stata commands for tasks such as regression analysis, data cleaning, and visualization. From both academic and industrial perspectives, Stata-MCP exemplifies the application of the MCP standard to social science research tools. It introduces a new paradigm of intelligent interaction for empirical researchers in fields such as economics and sociology.

2.4 Existing Approaches and Limitations

Machine learning, a subset of artificial intelligence, broadly refers to algorithms that identify patterns in data and use them to perform tasks such as prediction, classification, and clustering. The application of machine learning in economics generally presents on three areas: data generation, model prediction, and causal inference [22,23]. In econometrics, methods for causal inference have evolved from approaches such as Instrumental Variables, Difference-in-Differences, Regression Discontinuity Design, and Synthetic Control. With the advancement of machine learning techniques, researchers have further developed methods like Double Machine Learning [24,25].

In the aftermath of the release of ChatGPT-3.5, early literature recognized its potential utility across six key domains: ideation and feedback, writing, background research, data analysis, coding, and mathematical derivations [14]. Subsequent research has demonstrated that ChatGPT is not merely a tool for stylistic refinement in academic writing, but rather a transformative instrument capable of enabling qualitative leaps in scholarly research. Yet, the academic community still lacks a fully developed tool of this kind [26,27]. The most recent studies suggest that ChatGPT can serve as a tutor in economics, showing strong performance in explaining foundational concepts and solving multiple-choice questions, though it also exhibits notable shortcomings such as overconfidence, insufficient explanatory depth, and low-quality examples [15]. However, no feasible and stable solution has yet been proposed for applying large language models to empirical research in economics.

Although a team has boasted of developing a "Econometrics-Agent" that purportedly exceeds ChatGPT's performance by several-fold on their own dataset [28], the actual user experience is unsatisfactory: installation is cumbersome—non-technical users are unlikely to complete it successfully on the first attempt; methodological flexibility is limited, as all procedures are hard-coded and non-extensible; and the system is tethered to a specific platform, being a modification of MetaGPT that requires OpenWebUI as the local runtime. The authors further contend that LLMs

are ill-suited to writing Stata code or understanding econometric algorithms, but this stance neglects the capability bounds of pre-trained models. Accordingly, we empirically assess LLMs' capabilities in econometric and Stata programming under an open, minimally constrained system environment, using Stata-MCP as a lightweight and easier operable plug-in.

3 Research Method

3.1 Model Evaluation

3.1.1 Task Design

To construct the capability evaluation dataset, we randomly selected a number of questions from *Introductory Econometrics: A Modern Approach* [29], and supplemented them with additional question–answer pairs curated from our self-developed econometrics knowledge base. In total, 50 items were compiled to serve as the basis for experimental tasks. These exercises are designed to assess the performance of LLMs in three areas: understanding of economic concepts, comprehension of econometric models, and proficiency in writing Stata code. For each task, we predefined reference answers and established corresponding scoring criteria.

During the experiment, we instructed different large language models (LLMs) to complete the same set of tasks under standardized prompts. The models were required not only to produce step-by-step solutions, but also to include procedures such as data cleaning, code implementation, regression estimation, result interpretation, and visual presentation. We systematically documented each model's logical consistency in data processing, completeness of regression procedures, and accuracy of final outputs. Model performance was evaluated along multiple dimensions based on pre-defined scoring criteria. This process involved designing and maintaining a scoring framework, manually reviewing and comparing the generated content, and ensuring consistency and reproducibility in the evaluation.

In our implementation, we adopted LangChain and LangGraph as the foundational frameworks, utilized Chroma for the knowledge base architecture, and constructed a ReAct Agent to attain comparatively effective results [30]. All core code has been fully open-sourced on GitHub.

3.1.2 Model Choices

To more effectively evaluate the performance of models in empirical research and select among them, we include a range of mainstream models currently available on the market. These comprise

proprietary, closed-source models such as the ChatGPT series, Claude series, Gemini series, and Grok series, as well as open-source models including DeepSeek, Qwen, Kimi, and GLM.

These models exhibit markedly different characteristics, which enables us to assess which types of models are better suited to our specific tasks. For instance, ChatGPT, Claude, Gemini, and Grok are all closed-source models developed in the United States, whereas models such as Qwen and DeepSeek are open-source and developed in China. Some models, such as DeepSeek-R1, demonstrate relatively strong reasoning capabilities, while others, like DeepSeek-V3, perform less effectively in reasoning tasks. Furthermore, certain models like DeepSeek-V3.1 have been specifically optimized for agent usage.

To ensure consistency and minimize potential biases stemming from different model providers, we select OpenRouter as our designated model provider. Additionally, to ensure traceability and systematic documentation of all requests, we utilize Cloudflare's AI Gateway service to log, archive, and standardize the incoming queries.

3.1.3 MASA (Model Adversarial Scoring Architecture)

We constructed our evaluation framework by drawing on the assessment methodology of the GRA strategic [31] for evaluating model performance with different scenarios and the Scalar Reward Model [32], whose design principles inspired the construction of our custom stance model for generating evaluative feedback, including both positive and negative perspectives. We refer to this adversarial evaluating system as Model Adversarial Scoring Architecture (MASA).

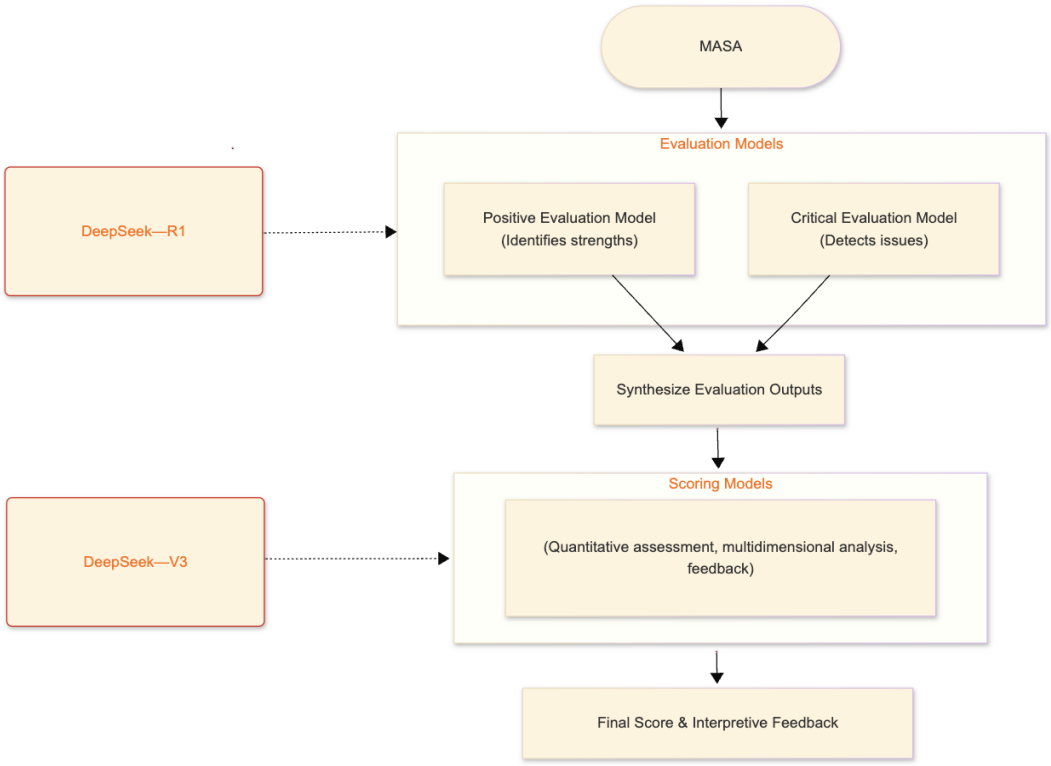

Figure 2. Workflow of MASA

MASA is composed of two functionally distinct classes of models: evaluation models and scoring models. Figure 2 depicts the MASA framework. The evaluation models provide qualitative assessments of the task execution process without assigning numerical scores. The positive evaluation model is designed to identify strengths and highlights in task performance, while the critical evaluation model focuses on detecting potential issues and shortcomings. The outputs from this pair of evaluative models are then synthesized and passed to the scoring model. Drawing on predefined scoring criteria, reference answers, the execution process, final outcomes, and the evaluative feedback, the scoring model conducts a comprehensive, multidimensional quantitative assessment and generates both a final score and interpretive feedback.

Regarding the choice of base models, the evaluation models are built on DeepSeek-R1, owing to its outstanding chain-of-thought capabilities, which enable the generation of structured and coherent reasoning processes during evaluation. Unlike models that provide only final conclusions, DeepSeek-R1 is particularly well suited for explanatory evaluation tasks, ensuring that both positive and critical assessments remain traceable and consistent, thereby enhancing transparency and professionalism. For the scoring stage, MASA employs DeepSeek-V3 as the base scoring model,

which demonstrates robustness in multi-turn dialogue, complex instruction following, and structured output generation. Its sufficiently large context window further allows for the integration of diverse input sources, enabling multidimensional quantitative analysis and result generation. Therefore, MASA framework effectively integrates interpretability with quantitative rigor, offering a powerful tool for the comprehensive evaluation of complex tasks.

3.2 Data and Variables

The data are derived from our own testing results, which were subsequently digitized and structured for analysis. Specifically, we recorded the final overall scores, task-level scores across different scenarios, and model-specific attributes such as whether the model exhibits reasoning capabilities. Table 1 presents the descriptive statistics of the dataset, and it also describes the meaning of variables.

Table 1. Data Descriptive Statistics

VarName	Obs	Mean	SD	Min	Max
score	348	65.934	22.905	0.000	100.000
is_thinking	348	0.417	0.494	0.000	1.000
is_opensource	348	0.500	0.501	0.000	1.000
context	348	2.79e+05	2.31e+05	1.28e+05	1.00e+06

3.3 Empirical Strategy

We employ a variety of models and specify our benchmark regression as follows:

$$Score_{i,j} = \beta_0 + \beta_1 \cdot thinking_i + \beta_2 \cdot opensource_i + \beta_3 \cdot log(context) + e$$

where $Score_{i,j}$ denotes the performance score of model i on task j, as evalated by MASA. The variable $thinking_i$ is a dummy variable meaning whether this model is a thinking model, $opensource_i$ is also a dummy indicating whether the model is open-source, and $context$ is the model's context window limitation.

4 Results

We present violin plots of the scores across different model providers in Figure 3. The figure illustrates the distribution of model performance, and it is evident that Anthropic—whose models are well known for their strong programming capabilities—also demonstrates superior proficiency in generating Stata code. Moreover, its outputs appear comparatively more stable than those of other models. Beyond the graphical evidence, tabulated comparisons of means and variances further corroborate that the Anthropic Claude model demonstrates notable strengths in both Stata

programming and econometric model interpretation.

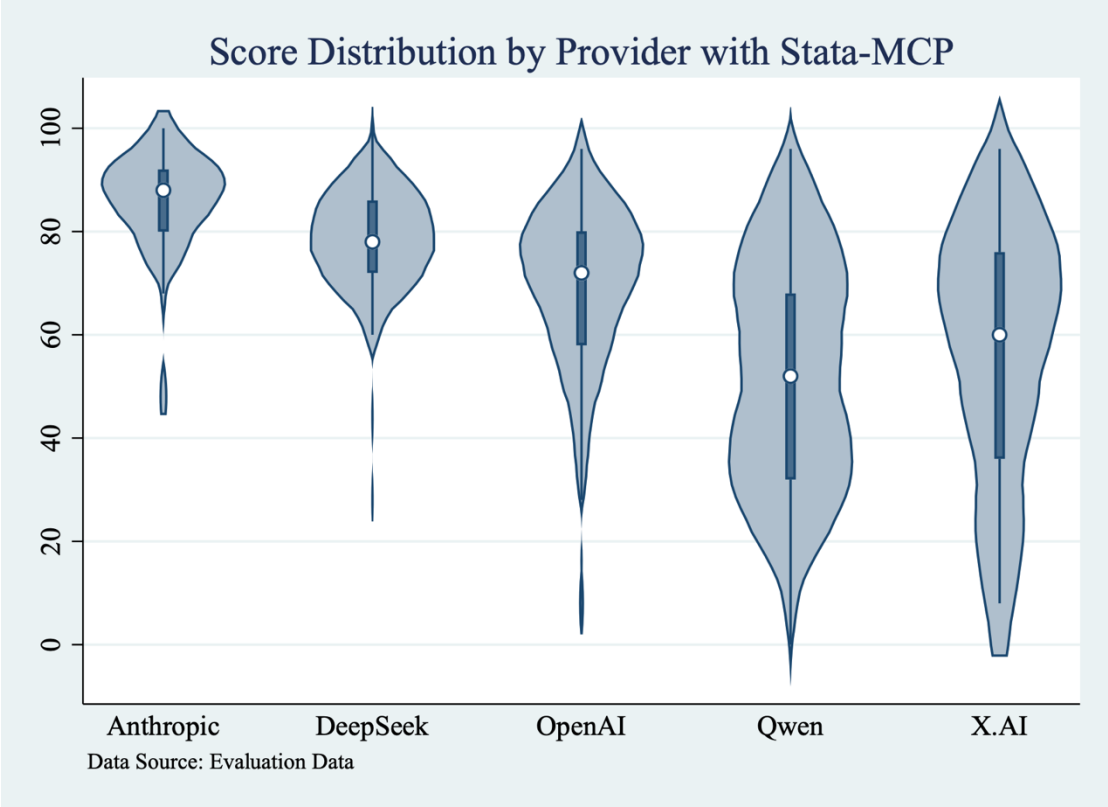

Figure 3. Score Distribution by Provider with Stata-MCP

Table 2. Different model score

Provider	Obs	Mean	SD	Min	Max
Anthropic	58	85.74138	10.05173	48	100
DeepSeek	58	77.48276	11.49865	28	100
OpenAI	87	66.18391	20.1434	8	96
X.AI	58	55.34483	25.8949	8	96
Qwen	87	51.83908	21.9274	0	96

In order to examine whether certain factors affect model performance, we also estimate a regression model and assess the significance of the coefficients. The results, presented in Table 3, suggest that we do not have sufficient grounds to establish a robust relationship between model performance and either openness (open-source status) or whether the model is designed for reasoning, given the large number of uncontrollable factors involved. However, from the details of task execution, we find compelling evidence that model performance is closely related to the size of the context window. During task execution, it becomes apparent that under the current ReAct framework all preceding context is passed in as background information. This substantially lengthens the prompt, often causing many models to hit their context window limit and thereby

impairing performance. A more detailed discussion is provided in the *Limitations* sub-section.

Table 3. The result of regssion

	(1)	(2)	(3)	(4)
VARIABLES	Model 1	Model 2	Model 3	Model 4
is_thinking	-2.138	1.414		
	(2.861)	(2.493)		
is_opensource	-4.291		-8.994***	
	(2.836)		(2.411)	
context	2.30e-05***			
	(6.62e-06)			
l_context				9.998***
				(2.127)
Constant	62.57***	65.34***	70.43***	-57.43**
	(2.715)	(1.609)	(1.705)	(26.27)
Observations	348	348	348	348
R-squared	0.075	0.001	0.039	0.060

5 Reasoning and Limitations

5.1 Ability

5.1.1 Coding on Stata

Our test data primarily consists of exercises selected from *Introduction to Econometrics*, which concentrate on model analysis and comprehension within econometrics. Put more simply, the focus lies in translating econometric models into executable Stata code, thereby allowing us to identify and substantiate AI's competence in Stata programming. As shown in the *Results* section, the model is capable not only of generating syntactically valid Stata code from simple natural language instructions, but also of autonomously detecting and correcting errors. When confronted with more complex tasks, it can consult help documentation on its own. In multi-step tasks, the model is able to independently organize and sequence operations, thereby accomplishing the entire workflow of data cleaning, processing, importing, modeling, and output.

5.1.2 ReAct Framework

As noted earlier, the ReAct framework emphasizes the alternating integration of reasoning and acting. In practical task execution, the model demonstrates strong scheduling and process-management capabilities, adhering closely to the logical sequence prescribed by the ReAct

framework. Upon receiving a task, the model typically begins by decomposing it and organizing its logical structure, thereby clarifying the required operations and methods at each stage. It then translates these reasoning outcomes into concrete execution steps, operationalized through the invocation of appropriate Stata commands. During execution, the model dynamically adjusts subsequent actions based on the outputs of preceding steps, thereby maintaining both continuity and adaptability. This mechanism not only ensures systematicity and coherence in task completion, but also provides reliable technical support for multi-step procedures in complex empirical economic research.

5.2 Limitation

5.2.1 Model Thinking Ability

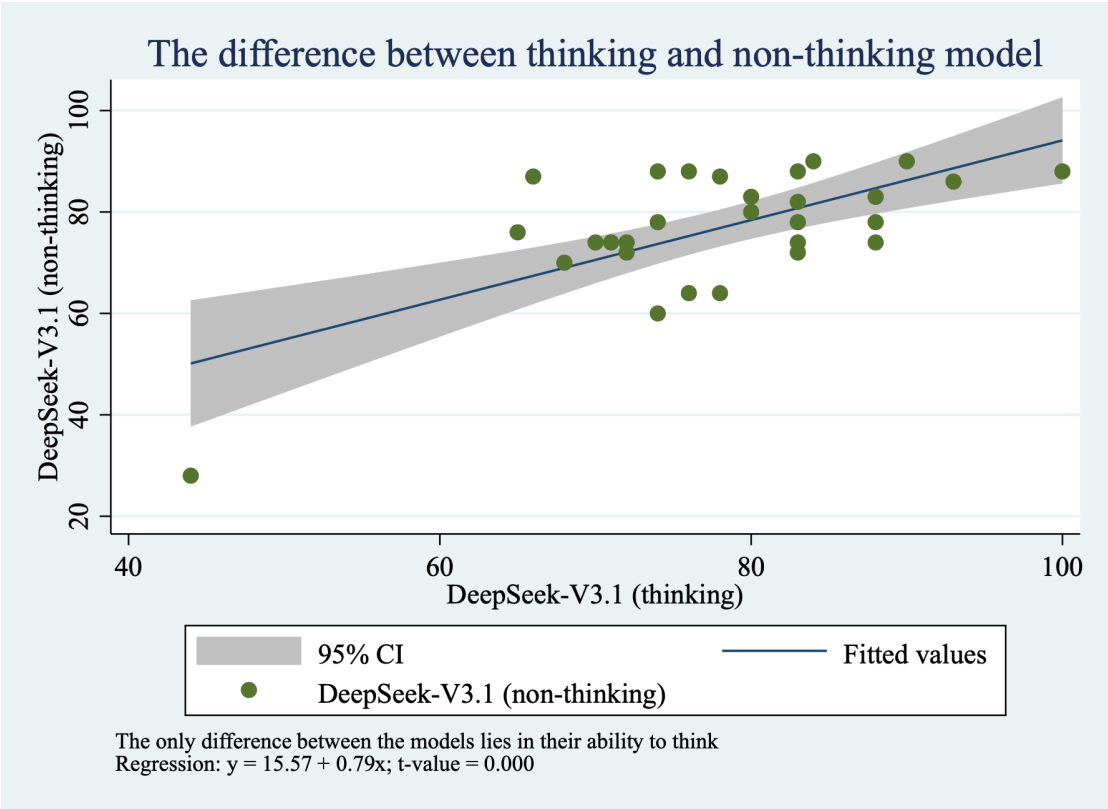

Figure 4. The difference between thinking and non-thinking model

Since LLMs are essentially still predictive models, and there is insufficient evidence to demonstrate that they are truly capable of "thinking", it becomes necessary to delineate their capabilities. Predictive models tend to perform well in tasks requiring structured output, but their performance is often underwhelming when it comes to open-ended tasks such as research design.

To further explore this distinction between thinking and non-thinking models, we switched our model API provider from OpenRouter to DeepSeek, which only offers two options: *deepseek-chat* and *deepseek-reasoner*, corresponding respectively to DeepSeek-V3.1 (non-thinking) and DeepSeek-V3.1 (thinking). As expected, the performance difference between the two was not statistically significant, the related result as present following figure.

5.2.2 Context Window

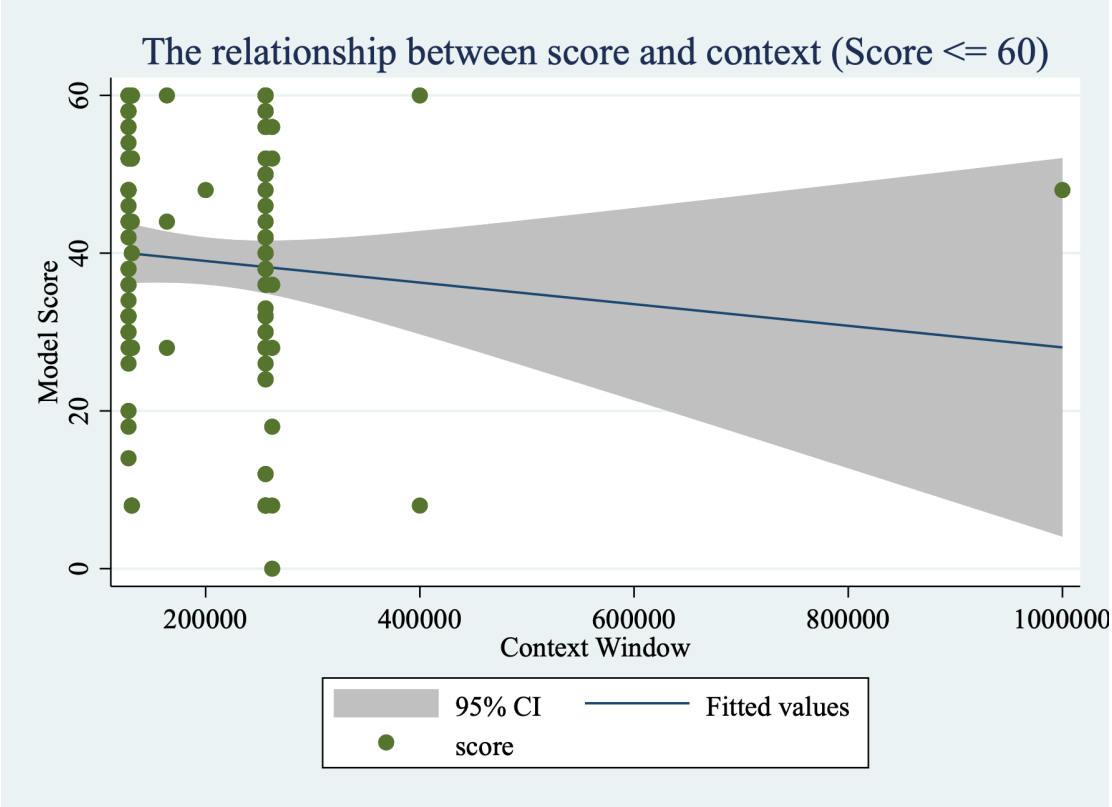

Figure 5. The mainly context window while score lower than 60

We observed that when using *deepseek-chat*, context truncation occurred frequently that tasks initiated by the agent were often prematurely terminated due to the context length reaching its maximum limit. To further investigate this issue, we collected and compiled information on context length limits for models available via OpenRouter, as illustrated in the figure below. Our findings suggest that performance degradation due to insufficient context length is a widespread phenomenon.

This observation also issues a new research direction for Stata-MCP: complex tasks are prone to failure when constrained by limited context length. There are two potential solutions to this

problem: one is to select models with larger context windows, and the other is to process and optimize the context itself. While the former largely depends on advancements made by model providers, our next step will focus on the latter — specifically, compressing the context from the perspective of context engineering in order to improve efficiency and task completion under context length constraints.

6 Conclusions

Through comprehensive testing across a range of econometric tasks, we find that large language models exhibit remarkable capabilities in both Stata programming and the comprehension of econometric models. The models can not only generate syntactically correct Stata code from natural language instructions, but also autonomously detect and correct errors, independently consult help documentation, and systematically organize and execute multi-step operations in complex tasks. Notably, the Anthropic Claude series demonstrated exceptional stability and accuracy in our evaluation, with an average score of 85.74. These results provide strong evidence that AI is now capable of handling the full empirical research workflow—from data cleaning, processing, and importation to modeling and output—effectively mastering the core technical skills required in empirical research.

Although current models have demonstrated solid foundational capabilities, realizing the broad application of AI in social science research still requires the development of more intelligent and domain-specialized agent systems. Our use of the ReAct framework supports this view: under this framework, the models exhibited strong task scheduling and process management abilities, dynamically adjusting execution strategies to ensure systematic and coherent task completion. However, the lack of significant performance differences between “thinking” and “non-thinking” models suggests that deeper innovation is needed at the level of agent design. Future research should focus on developing more specialized agents for economics and the social sciences—agents that can better understand discipline-specific research paradigms, methodological requirements, and analytical frameworks, thereby offering researchers more accurate and effective support.

Although AI performs impressively in empirical research, limitations in context window size remain a key bottleneck to its further advancement. Our findings show that as task complexity increases and context length approaches the model’s limit, interruptions and performance

degradation occur frequently—a phenomenon observed consistently across different models. Regression analysis reveals a significant positive correlation between model performance and context window size, indicating that context length is a critical factor influencing outcomes.

Thus, the Stata-MCP project still holds considerable room for optimization. On one hand, improvements will depend on model providers offering larger context windows; on the other hand—and more importantly—progress must come from the perspective of context engineering. This involves enhancing techniques in context compression, task decomposition, and information filtering to improve efficiency and task success rates under constrained context conditions. This is not only a technical challenge, but also a necessary step toward the deeper integration of AI in academic research.

Reference:

[1] JUMPER J, EVANS R, PRITZEL A, et al. Highly Accurate Protein Structure Prediction with AlphaFold[J]. Nature, 2021, 596(7873):583-589.

[2] VARADI M, BERTONI D, MAGANA P, et al. AlphaFold Protein Structure Database in 2024: Providing Structure Coverage for over 214 Million Protein Sequences[J]. Nucleic Acids Research, 2024, 52(D1):D368-D375.

[3] KARLBAUER M, CRESSWELL-CLAY N, DURRAN D R, et al. Advancing Parsimonious Deep Learning Weather Prediction Using the HEALPix Mesh[R]. 2024.

[4] PATHAK J, SUBRAMANIAN S, HARRINGTON P, et al. FourCastNet: A Global Data-Driven High-Resolution Weather Model Using Adaptive Fourier Neural Operators[R]. 2022.

[5] PRICE I, SANCHEZ-GONZALEZ A, ALET F, et al. Probabilistic Weather Forecasting with Machine Learning[J]. Nature, 2025, 637(8044):84-90.

[6] RAISSI M, PERDIKARIS P, KARNIADAKIS G E. Physics-Informed Neural Networks: A Deep Learning Framework for Solving Forward and Inverse Problems Involving Nonlinear Partial Differential Equations[J]. Journal of Computational Physics, 2019, 378:686-707.

[7] LI Z, KOVACHKI N, AZIZZADENESHELI K, et al. Fourier Neural Operator for Parametric Partial Differential Equations[R]. 2021.

[8] TAN S. Stata-MCP: Let LLM Help You Achieve Your Regression Analysis with Stata[EB/OL]. https://github.com/sepinetam/stata-mcp. Shanghai, China, 2025.

[9] VASWANI A, SHAZEER N, PARMAR N, et al. Attention Is All You Need[R]. 2017.

[10] ZHAO W X, ZHOU K, LI J, et al. A Survey of Large Language Models[R]. 2025.

[11] FLAVIO C, JELMER R, LEA S. Unlocking Productivity with Generative AI: Evidence from Experimental Studies[M]. OECD, 2025.

[12] BRYNJOLFSSON E, LI D, RAYMOND L. Generative AI at Work[J]. The Quarterly Journal of Economics, 2025, 140(2):889-942.

[13] LEE D, ARNOLD M, SRIVASTAVA A, et al. The Impact of Generative AI on Higher Education Learning and Teaching: A Study of Educators' Perspectives[J]. Computers and Education: Artificial Intelligence, 2024,

6:100221.
[14] KORINEK A. Generative AI for Economic Research: Use Cases and Implications for Economists[J]. Journal of Economic Literature, 2023, 61(4):1281-1317.
[15] BROSE N, SPIELMANN C, TODE C. ChatGPT as Economics Tutor: Capabilities and Limitations[M]. School of Economics, University of Bristol, UK, 2025.
[16] COMANICI G, BIEBER E, SCHAEKERMANN M, et al. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities[R]. 2025.
[17] OPENAI, ACHIAM J, ADLER S, et al. GPT-4 Technical Report[R]. 2024.
[18] TEAM G, GEORGIEV P, LEI V I, et al. Gemini 1.5: Unlocking Multimodal Understanding across Millions of Tokens of Context[R]. 2024.
[19] TEAM G, ANIL R, BORGEAUD S, et al. Gemini: A Family of Highly Capable Multimodal Models[R]. 2025.
[20] ANDREA S. Will AI Replace Programmers and Software Engineers?[M]. San Diego: UC San Diego, 2024.
[21] RACHEL G. Can AI Really Code? Study Maps the Roadblocks to Autonomous Software Engineering[M]. MIT CSAIL, 2025.
[22] SUSAN A. The Impact of Machine Learning on Economics[A]. The Economics of Artificial Intelligence: An Agenda[C]. University of Chicago Press, 2018: 507-547.
[23] SUSAN A, GUIDO W I. Machine Learning Methods That Economists Should Know About[J]. Annual Review of Economics, 2019, 11:685-725.
[24] CHERNOZHUKOV V, CHETVERIKOV D, DEMIRER M, et al. Double/Debiased Machine Learning for Treatment and Structural Parameters[J]. The Econometrics Journal, 2018, 21(1):C1-C68.
[25] ATHEY S, IMBENS G W. The State of Applied Econometrics: Causality and Policy Evaluation[J]. Journal of Economic Perspectives, 2017, 31(2):3-32.
[26] DONG M M, STRATOPOULOS T C, WANG V X. A Scoping Review of ChatGPT Research in Accounting and Finance[J]. International Journal of Accounting Information Systems, 2024, 55:100715.
[27] CILLO P, RUBERA G. Generative AI in Innovation and Marketing Processes: A Roadmap of Research Opportunities[J]. Journal of the Academy of Marketing Science, 2025, 53(3):684-701.
[28] CHEN Q, HAN T, LI J, et al. Can AI Master Econometrics? Evidence from Econometrics AI Agent on Expert-Level Tasks[R]. 2025.
[29] WOOLDRIDGE J M. Introductory Econometrics: A Modern Approach[M]. Cengage Learning India, 2020.
[30] YAO S, ZHAO J, YU D, et al. ReAct: Synergizing Reasoning and Acting in Language Models[R]. 2023.
[31] GAO X, PEI Q, TANG Z, et al. A Strategic Coordination Framework of Small LLMs Matches Large LLMs in Data Synthesis[R]. 2025.
[32] YE Z, LI X, LI Q, et al. Beyond Scalar Reward Model: Learning Generative Judge from Preference Data[R]. 2024.