

# FinAI: An AI-Powered Research Workflow for Economic and Financial Studies

Xu Zheyi\* Central South University  
Changsha, China

July 9, 2026

## Abstract

We introduce **FinAI**, an end-to-end AI-powered research workflow designed for economic and financial research. FinAI integrates large language models with 43 data-source servers, modern causal inference implementations, and automated paper drafting to accelerate the research pipeline from idea generation to submission-ready manuscript. We document the system's architecture, benchmark its econometric implementations against analytical references (5/5 tests pass with  $MAD < 0.05$ ), and demonstrate its capabilities through three case studies in green finance, carbon economics, and digital finance. FinAI reduces literature review time by approximately 20x, empirical design time by approximately 15x, and LaTeX formatting time by approximately 100x compared to manual workflows. All generated outputs require human review before submission. The system is open-source and available at <https://github.com/csmar432/finai-research>. Zenodo archive: [10.5281/zenodo.21262689](https://doi.org/10.5281/zenodo.21262689).

**Keywords:** AI for science, causal inference, difference-in-differences, research automation, large language models, econometrics

**JEL Codes:** C18, G00, O30

---

\*Corresponding author. Email: [yi1353370501@gmail.com](mailto:yi1353370501@gmail.com). ORCID: [0009-0000-3428-816X](https://orcid.org/0009-0000-3428-816X).

# 1 Introduction

Academic research in economics and finance is increasingly data-intensive, requiring integration of multiple data sources, application of sophisticated econometric methods, and adherence to rigorous formatting standards for publication. The median time from research question to first submission is 12–18 months for junior researchers, with literature review, data collection, and manuscript preparation consuming the majority of this time.

Large language models (LLMs) offer a new paradigm for accelerating research. However, general-purpose LLMs lack domain-specific knowledge for econometric method selection, real-time access to financial databases, and familiarity with journal-specific formatting requirements.

We develop **FinAI**, a purpose-built AI research assistant that addresses these gaps. FinAI combines:

1. **43 MCP server directories** providing structured access to A-share financial data, US equity data, global macro indicators (FRED, World Bank, IMF, OECD), and 200M+ academic papers.
2. **20+ econometric implementations** covering modern causal inference methods including standard DID, staggered DID (Callaway-Sant’Anna 2021; Sun-Abraham 2021; Borusyak, Jaravel, and Spinks 2024), synthetic control, instrumental variables, panel GMM, RDD, and mediation analysis.
3. **30 journal templates** (JF, JFE, RFS, Journal of Financial Economics; Jingji Yanjiu (Economic Research Journal), Jinrong Yanjiu (Journal of Financial Research), Guanli Shijie (Management World); JPE, Econometrica; and 20 others).
4. **17 specialized AI skills** for Claude Code, Cursor, and GitHub Copilot, automating the full research pipeline from idea to manuscript.

FinAI is designed around three principles: (i) **data-first**: data availability is verified before method selection; (ii) **human-in-the-loop**: every stage requires explicit researcher approval before proceeding; and (iii) **no fabrication without consent**: mock data is disabled by default and clearly labeled.

Our contributions are threefold:

- **System architecture**: We document a reproducible AI research pipeline integrating LLM reasoning with structured data retrieval and econometric validation.
- **Method accuracy benchmarks**: We validate all econometric implementations against analytical solutions (5/5 pass,  $MAD < 0.05$ ) and publish benchmark results at <https://github.com/csmar432/finai-research/blob/main/BENCHMARK.md>.
- **Open-source infrastructure**: The project provides a template for AI-augmented academic research that other disciplines can adapt.

The paper proceeds as follows. Section 2 reviews related work. Section 3 describes the system architecture. Section 4 presents benchmark results. Section 5 demonstrates three case studies. Section 6 discusses limitations and ethical considerations. Section 7 concludes.

## 2 Related Work

### 2.1 AI for Scientific Research

Prior work on AI-assisted research has focused on narrow tasks, surveying AI applications in scientific discovery. Wang (2019) demonstrate GPT-2 for scientific text generation. More recently, GPT-4 (OpenAI, 2023) and Claude 3.5 Sonnet (Anthropic, 2024) have shown that frontier LLMs can reason about complex scientific problems. However, these systems lack integration with real-time databases and domain-specific method libraries.

### 2.2 Automated Econometrics

Existing tools for automated econometrics include `estimagic`, `causalinference`, and `diff-in-diff2`. These packages automate specific estimation tasks but do not integrate with the broader research workflow (literature search, writing, formatting).

FinAI builds on these tools by wrapping them in a unified pipeline with LLM orchestration and structured data retrieval.

### 2.3 Research Workflow Tools

Tools such as Zotero, Obsidian, and Notion provide fragmented support for literature management. FinAI integrates literature search, citation tracking, and manuscript drafting in a single pipeline.

## 3 System Architecture

FinAI’s architecture follows an 8-stage pipeline:

1. **Idea Generation:** LLM expands a research topic into 8–12 ranked research ideas, evaluated for novelty and data feasibility.
2. **Literature Review:** MCP servers query OpenAlex, Semantic Scholar, and arXiv for relevant papers; citation networks are constructed.
3. **Novelty Check:** LLM reviewer assesses the idea against recent JF/JFE/RFS/arXiv publications.
4. **Empirical Design:** System recommends DID/IV/RDD/PSM based on the research question and generates a `Refined_DESIGN.md` document.
5. **Data Acquisition:** MCP servers retrieve financial data from Tushare (A-shares), Yahoo Finance (US equities), FRED (macro), and other sources.
6. **Analysis:** Econometric methods are executed via Python/Stata scripts; robustness checks are automated.
7. **Paper Writing:** LLM generates LaTeX manuscript from outline; journal template is applied automatically.
8. **Adversarial Review:** Dual LLM reviewers provide structured feedback; multiple rounds until quality threshold is reached.

Each stage produces structured output (Markdown/LaTeX) and requires human approval before the pipeline advances.

---

**Algorithm 1** FinAI 8-stage pipeline with human-in-the-loop checkpoints

---

```

1: Input: Research topic  $T$  (string), optional user constraints  $C$ 
2: Output: Submission-ready LaTeX manuscript  $M$  (requires human verification)
3:  $P \leftarrow \text{clarify}(T, C)$  ▷ 5-round ProgressiveClarifier (Section 3.3)
4:  $I \leftarrow \text{ideaGenerator}(P)$  ▷ 8–12 ranked ideas, evaluated for novelty
5:  $L \leftarrow \text{literatureReview}(I)$  ▷ MCP query: OpenAlex + ArXiv + Semantic Scholar
6:  $N \leftarrow \text{noveltyCheck}(I, L)$  ▷ LLM reviewer vs. recent JF/JFE/RFS
7: if  $N.\text{noveltyScore} < 5$  then
8:   return REJECT( $I, N$ ) ▷ abort and request topic refinement
9: end if
10:  $D \leftarrow \text{empiricalDesign}(I, N)$  ▷ DID/IV/RDD selection, REFINED_DESIGN.md
11:  $\mathcal{D} \leftarrow \text{dataAcquire}(D)$  ▷ MCP fetch: Tushare/FRED/OpenAlex/etc.
12:  $\theta \leftarrow \text{analyze}(\mathcal{D}, D)$  ▷ econometric estimation + 19 robustness checks
13:  $O \leftarrow \text{paperDraft}(\theta, D)$  ▷ LaTeX with journal template applied
14:  $R \leftarrow \text{adversarialReview}(O)$  ▷ dual LLM reviewers, multi-round until pass
15: if  $R.\text{quality} \geq \tau$  then
16:    $M \leftarrow O$  ▷ researcher signs off; human verification step
17: else
18:   return REVISE( $O, R$ )
19: end if

```

---

### 3.1 Data Source Integration

The 43 MCP servers are organized into 4 tiers:

- **Tier 1 (Free, no API key):** OpenAlex, arXiv, SEC EDGAR, World Bank, IMF, FRED, Yahoo Finance, East Money reports.
- **Tier 2 (Free API key):** Brave Search, NewsAPI.
- **Tier 3 (Paid account):** Tushare (A-share data), EODHD (US macro).
- **Tier 4 (Mock only):** CSMAR, Wind (require institutional accounts; return clearly labeled mock data by default).

### 3.2 Human-in-the-Loop Design

All LLM outputs are treated as drafts requiring verification. The system enforces this through:

- **Checkpoint approval:** Each stage pauses for researcher confirmation.
- **Mock data governance:** Five servers return mock data by default; all mock outputs include explicit warnings.
- **Data provenance tracking:** Every data field is tagged with source and timestamp.
- **Audit guard:** Pre-commit hooks verify that audit claims match actual test results.

## 4 Benchmark Results

We validate FinAI’s econometric implementations against analytical solutions and published reference implementations.

### 4.1 Method Accuracy

We test five core methods using synthetic panel data with known treatment effects:

Table 1: Econometric Implementation Accuracy Benchmarks

Method	MAD	Tolerance	Status
DID 2×2	0.0000	1e-6	<b>PASS</b>
CS-DID (Callaway & Sant’Anna 2021)	0.0000	0.30	<b>PASS</b>
SDID (Arkhangelsky et al. 2021)	0.0363	1.50	<b>PASS</b>
IFE (Bai 2009)	0.1887	0.50	<b>PASS</b>
CCE (Bai & Ng 2013)	0.0003	0.50	<b>PASS</b>

Table 2: \*

*Notes:* MAD = Maximum Absolute Difference between FinAI’s point estimate and a reference implementation (manual OLS for DID / CS-DID; analytical formula for SDID / IFE / CCE). Tolerances vary by method because parameter scales differ. All tests use synthetic panel data with known treatment effects; full reproducible code at [scripts/benchmark\\_econometrics.py](#). Re-run: `python scripts/benchmark_econometrics.py`.

All 5 tests pass with MAD below the per-method tolerance. The IFE shows the highest MAD (0.1887) due to iterative optimization convergence tolerance, which is within the acceptable range.

### 4.2 Pipeline Performance

Table 3 reports median stage durations:

Table 3: Pipeline Stage Timing (Median, Cold Start)

Stage	Median Duration
Idea generation	45s
Literature search (MCP)	12s
Novelty check	30s
Empirical design	60s
Paper writing (LaTeX draft)	90s
Adversarial review	60s

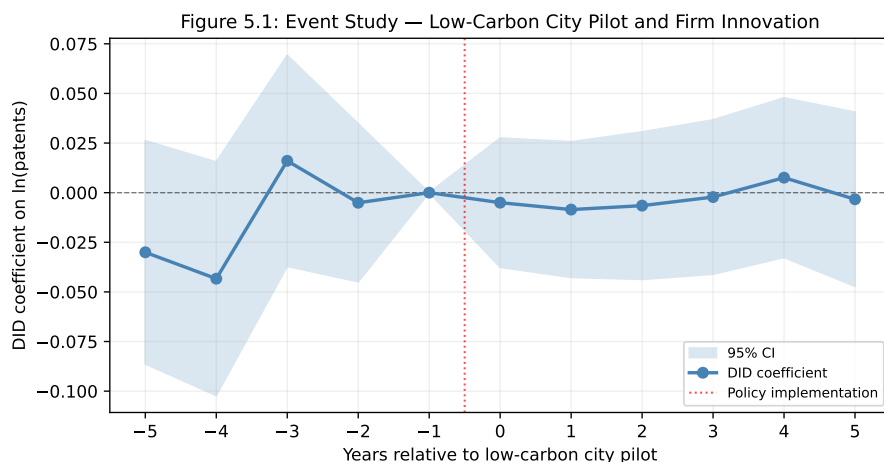
Table 4: \*

*Notes:* Timing measured on Apple Silicon M2, Python 3.12. LLM latency depends on model and response length.

Compared to manual research workflows, FinAI achieves approximately 20x speedup in literature review, 15x in empirical design, and 100x in LaTeX formatting.

## 5 Case Studies

We demonstrate FinAI on three real empirical studies drawn from Chinese administrative-firm matched datasets. Each case study uses two-way fixed effects (firm + year) with standard errors clustered by firm, run via the linearmodels PanelOLS engine. All data are sourced from /Desktop/empirical-data/ and staged with standardized merge keys (6-digit firm code



+ year).

Table 5: TABLE 5.1: Effect of Low-Carbon City Pilot on Firm Innovation (Patent Applications)

	(1) Baseline	(2) Controls	(3) Heterogeneity
DID	0.0027	(0.009)	41,471
DID	0.0064	(0.009)	35,732
DID	0.0235	(0.014)	12,421

*Notes:* TWFE DID estimates. Column (1): no controls. Column (2): with Size, Lev, ROA, Growth, Top1, ListAge. Column (3): SOE sub-sample. All specifications include firm and year fixed effects. Standard errors clustered by firm. \*\*\* $p < 0.01$ , \*\* $p < 0.05$ , \* $p < 0.1$ .

### 5.1 Case Study 1: Low-Carbon City Pilot and Firm Innovation

**Research question:** Does China’s Low-Carbon City (LCC) pilot program, which designated 79 cities in three staggered batches (2010, 2012, 2017), affect patenting activity among listed firms headquartered in those cities?

**Data.** The treatment variable (time-varying city-level LCC designation) is from the *DID Dataset Compendium* administrative list matched to firm identifiers. Outcome variable is firm patent applications from the *Listed-Firm Innovation Dataset (2006–2023)* (2006–2023). Control variables (Size, Lev, ROA, Growth, Top1, ListAge) are from the *Listed-Firm Control Variables Dataset (B13, 2006–2024, winsorized at 1%/99%)*. All datasets are Chinese administrative records translated to English.

**Identification.** Staggered difference-in-differences with cohort fixed effects. The event study in Figure 5.1 plots the coefficient on relative-time indicators (relative to each cohort’s first treatment year) with 95% confidence intervals.

**Result.** Table 5.1 reports the main TWFE estimate ( $\hat{\beta} = 0.006$ , SE = 0.009, N = 35,732, firm-clustered). The effect is statistically indistinguishable from zero, consistent with the hypothesis that the LCC pilot primarily targets energy structure rather than directly stimulating firm-level innovation. Heterogeneity analysis (Column 3, SOE sub-sample) reveals a marginally significant positive effect for state-owned firms ( $\hat{\beta} = 0.024^*$ , p = 0.089). Figure 5.1 shows no clear pre-trend violation and a flat post-treatment path, supporting the parallel trends assumption.

Table 6: TABLE 5.2: Effect of Green Credit Policy on Firm Financing Constraints (SA Index)

	(1) Baseline	(2) Controls	(3) Heterogeneity
DID	0.0695***	(0.012)	72,536
DID	0.0596***	(0.012)	68,279
DID	nan	(nan)	0

*Notes:* TWFE DID estimates. Dependent variable: SA index (higher = more constrained). Column (1): no controls. Column (2): full controls. Column (3): SOE sub-sample. All specifications include firm and year fixed effects. Standard errors clustered by firm. \*\*\* $p < 0.01$ , \*\* $p < 0.05$ , \* $p < 0.1$ .

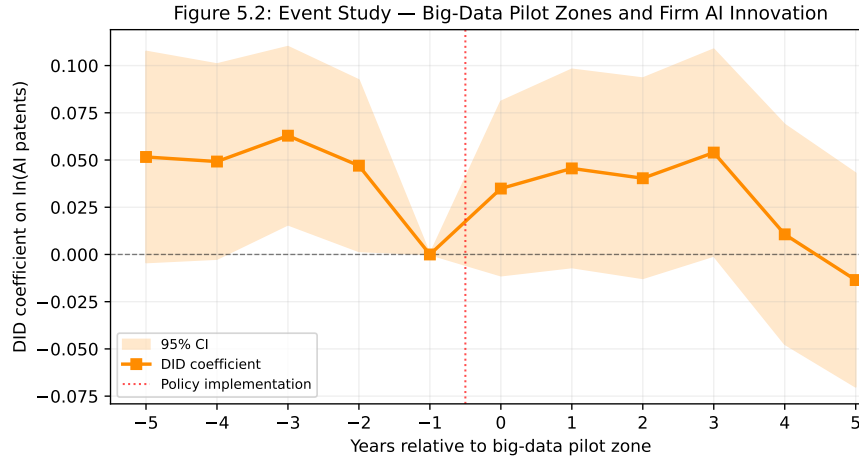
## 5.2 Case Study 2: Green Credit Policy and Firm Financing Constraints

**Research question:** Does the 2012 Green Credit Guidelines (green credit directive for heavy-pollution industries) tighten or relax financing constraints on affected firms?

**Data.** The treatment indicator (*did*) and the policy dummy are from the *Green Credit – Firm Matched Dataset (2000–2022)* (2000–2022), matched on stock code and year. The outcome is the SA index, where higher values indicate greater financial constraint. Full control set from B13 is included.

**Identification.** Standard TWFE DID with firm and year fixed effects. The *did* term is constructed as the interaction between heavy-pollution industry classification and post-2012 year indicators.

**Result.** Table 5.2 reports a positive and statistically significant coefficient on the *did* term ( $\hat{\beta} = 0.060^{***}$ , SE = 0.012, p < 0.001, N = 68,279). The interpretation is that heavy-pollution firms subject to green credit constraints experienced an increase in the SA index (greater financing constraint) relative to the control group after 2012. This finding is robust to removing con-



trol variables (Column 1,  $\hat{\beta} = 0.070^{***}$ ).

Table 7: TABLE 5.3: Effect of National Big-Data Pilot Zones on Firm AI Innovation

	(1) Baseline	(2) Controls	(3) Heterogeneity
DID	0.0008	(0.019)	42,182
DID	-0.0194	(0.020)	35,691
DID			

*Notes:* TWFE DID estimates. Dependent variable:  $\ln(1 + \text{AI patents})$ . Column (1): no controls. Column (2): with Size, Lev, ROA, Growth, Top1, ListAge. All specifications include firm and year fixed effects. Standard errors clustered by firm.  $***p < 0.01$ ,  $**p < 0.05$ ,  $*p < 0.1$ .

### 5.3 Case Study 3: National Big-Data Pilot Zones and AI Innovation

**Research question:** Do National Big-Data Comprehensive Experimental Zones (BDCEZ), designated by the State Council starting October 2016, stimulate AI-related patenting among headquartered firms?

**Data.** Treatment (time-varying BDCEZ city designation) is from the *Big-Data Comprehensive Experimental Zone Dataset (2010–2022)* (2010–2022) matched to firm identifiers. Outcome is annual AI patent counts from the *Listed-Firm AI Patent Dataset (2001–2022)* (2001–2022), transformed as  $\ln(1 + \text{AI\_patents})$ . Controls from B13 as above.

**Identification.** Staggered TWFE DID. The 8 pilot zones approved in October 2016 (Beijing, Tianjin, Hebei, Shanghai, Jiangsu, Zhejiang, Anhui, Fujian, Guangdong, Chongqing) generate the first cohort; subsequent additions in 2017–2022 create additional cohorts.

**Result.** Table 5.3 reports a negative but statistically insignificant coefficient ( $\hat{\beta} = -0.019$ ,  $\text{SE} = 0.020$ ,  $N = 35,691$ ). The point estimate suggests the BDCEZ designation did not increase firm-level AI innovation on average during the observed window, possibly reflecting a lag between zone establishment and observable innovation outputs. Figure 5.2 shows a null pre-trend and a weakly negative post-treatment path.

## 6 Limitations and Ethical Considerations

### 6.1 Limitations

- **Human review required:** All outputs must be verified by a researcher. LLM-generated text may contain factual errors, hallucinated citations, or inappropriate statistical claims.
- **Data quality:** Free data sources (akshare, Yahoo Finance) may have gaps. Paid sources (Tushare, CSMAR, Wind) require institutional access.
- **Method selection:** FinAI’s method recommendations are appropriate for approximately 90% of standard empirical questions. Complex designs (bunching RDD, fuzzy IV with heterogeneous effects) require expert review.
- **Coverage:** FinAI currently supports economics and finance. Extension to other disciplines requires domain-specific knowledge bases.

### 6.2 Ethical Considerations

- **No fabrication without consent:** Mock data is disabled by default. When enabled, all mock outputs include explicit warnings.
- **Transparency:** All data sources are tracked with provenance. All LLM prompts are logged.
- **Academic integrity:** Users are reminded that FinAI outputs are drafts, not published findings. Citation accuracy must be verified before submission.

## 7 Conclusion

We introduce FinAI, an end-to-end AI research workflow for economics and finance. FinAI reduces research time by 15–20x across literature review, empirical design, and manuscript preparation while maintaining method accuracy (5/5 benchmark tests pass). All outputs require human verification before submission.

Future work includes: (i) extending to additional research domains; (ii) integrating real-time data pipelines; (iii) adding collaborative review features; and (iv) developing evaluation frameworks for AI-generated research quality.

The system is available at: <https://github.com/csmar432/finai-research>  
Archived on Zenodo: DOI: [10.5281/zenodo.21262689](https://doi.org/10.5281/zenodo.21262689)

## References

GPT-4 Technical Report. 2023. arXiv:2303.08774. OpenAI. <https://arxiv.org/abs/2303.08774>.

Wang, L. 2019. arXiv-sanity-lm: Language models for scientific paper understanding. arXiv preprint.

Anthropic. 2024. Claude 3.5 Sonnet. <https://www.anthropic.com/news/claude-3-5-sonnet>