10.48047/jocaaa.2025.34.12.75

Benchmarking AI models and tools on advanced Excel-based financial analysis

Abhishek Nanda
Senior Manager, Corporate Development

Sohag Maitra
Senior Data Analytics Engineer

Shaurya Mehta
Investment Banking Analyst

Received: 05.10.2025
Revised: 15.11.2025
Accepted: 31.12.2025

***Abstract*— The integration of artificial intelligence (AI) into spreadsheet-based financial workflows is rapidly transforming the practice of financial analysis, reporting, and decision support. With the emergence of advanced large language models (LLMs) and Excel-focused automation tools, spreadsheet tasks such as formula generation, multi-sheet dependency tracing, and financial metric computation can now be executed with varying degrees of automation and reasoning support. This study systematically benchmarks five leading AI systems; Shortcut, TabAI in Excel, DeepSeek Agent with .xlsx capabilities, Mixtral 8x22B, and DeepSeek-V3 using SpreadsheetBench (Ma et al., 2024) and complementary finance-focused test sets to evaluate performance across accuracy, execution stability, task generalization, and user-effort dimensions. The results demonstrate that DeepSeek-V3 and TabAI in Excel outperforms alternatives in formula correctness, structural reliability, and multi-step financial computation, whereas Mixtral and DeepSeek provide clearer contextual interpretation beneficial for auditability and analytical explanation. Shortcut shows strong ability in direct spreadsheet manipulation but exhibits weaknesses in complex cell-relationship logic. These findings highlight that different AI systems possess specialized strengths, suggesting that hybrid deployment strategies pairing reasoning-heavy LLMs with spreadsheet-native automation offer the greatest potential for reliable, efficient, and audit-ready financial analysis. The study contributes objective evidence to support tool selection decisions in digitally transformed finance environments. (*Abstract*)**

I. Introduction

The increasing integration of AI in Excel-based financial workflows is transforming analytical practices

The rapid evolution of artificial intelligence (AI) and large language models (LLMs) has fundamentally reshaped how professionals approach financial analysis and spreadsheet-driven reporting tasks (Taheri Hosseinkhani, 2025). Traditionally, Excel has served as the core analytical platform for finance teams, supporting financial modelling, forecasting, budget variance assessment, ratio evaluation, and risk diagnostics. With the emergence of advanced generative AI technologies, spreadsheet operations are no longer limited to manual formula design and static computation. Instead, AI systems can now automatically generate formulas, detect errors, manipulate datasets, and interpret financial structures with improved accuracy (Sarker, 2022). This convergence of AI reasoning and Excel computation represents a major shift toward automation-enabled financial decision-making.

New Excel-focused AI tools enable direct interaction with financial data

While general-purpose LLMs have demonstrated reasoning capabilities across multiple domains, a new category of AI tools specifically designed for spreadsheets has emerged (Chen et al., 2025). Platforms such as Shortcut and Rogo offer interactive and automatic manipulation of Excel sheets, advanced formula construction, extraction of insights from large tabular datasets, and context-aware financial problem-solving. These tools extend beyond text-to-formula conversion by embedding analytical intelligence directly into spreadsheet environments. As financial datasets continue to expand in scale and complexity, such domain-focused tools are becoming increasingly essential for reducing analyst time, minimizing human errors, and improving audit-friendly data transparency (Putapu, 2025).

Competitive advancements among leading LLM providers drive performance uncertainty

Major AI developers including Microsoft, Mistral, DeepSeek, Cohere, Perplexity, and xAI/Grok have introduced capabilities that generate Excel models, automate data classification, derive financial assumptions from filings, and produce multi-sheet analysis workflows (Kuaiber et al., 2024). However, despite strong marketing claims, actual performance when executing financial spreadsheet operations remains insufficiently benchmarked. Significant variations may exist in formula correctness, error handling, output stability, data interpretation, and compliance with financial modelling standards. Therefore, a systematic and standardized evaluation is required to understand where currently available AI solutions excel, where they fail, and how they compare to tools built natively for Excel environments.

There is a need for benchmarking frameworks tailored to financial spreadsheet tasks

Current AI benchmarking studies largely focus on general reasoning, code generation, or document understanding. However, Excel-based financial analysis introduces unique challenges such as referencing accuracy, cell dependency tracing, nested formula optimization, handling missing information, and maintaining accounting-rule precision (Guerrero et al., 2019). Benchmarking must therefore consider realistic finance tasks such as cash-flow projections, ratio analysis, pivot-table manipulations, VLOOKUP/XLOOKUP-based data joins, and conditional logic operations. Evaluating performance on these criteria will help identify whether AI tools are reliable enough for professional financial reporting and compliance-driven environments where correctness is critical (Shah et al., 2025).

This study aims to compare the spreadsheet intelligence of general and domain-specific AI

Addressing the above gaps, this research systematically benchmarks leading LLMs and Excel-focused AI systems in executing spreadsheet computation and data manipulation tasks. By designing standardized task modules ranging from

10.48047/jocaaa.2025.34.12.75

formula generation to multi-sheet linking and financial metric calculations the study quantifies their analytical accuracy, response efficiency, contextual reasoning, and usability. The outcomes of this benchmarking effort will provide insights into how AI can be optimally deployed within financial workflows, what risks may arise during automation, and how future developments can improve the automation maturity of Excel-based analysis. Ultimately, this work contributes evidence-driven guidance for finance professionals, developers, and organizations seeking to leverage AI for smarter, faster, and more reliable financial analytics.

II. METHODOLOGY

Research design and benchmarking framework

This study adopts a comparative benchmarking design to evaluate how different AI models and tools perform on advanced Excel-based financial analysis tasks. The central independent variable is the AI system (Shortcut, TabAI in Excel, DeepSeek Agent with .xlsx capabilities, Mixtral, and DeepSeek-V3), while the main dependent variables include task accuracy, formula correctness, execution success, and interaction efficiency. The core evaluation framework is built around SpreadsheetBench, a publicly available benchmark consisting of real-world spreadsheet manipulation tasks and automatically scored test cases. Supplementary tasks inspired by an arXiv spreadsheet benchmark and the benchmark proposed by Thorne are incorporated to increase coverage of financial formula generation and reporting-oriented data transformations, particularly those common in corporate finance and financial statement analysis.

Selection of spreadsheet benchmarks

SpreadsheetBench was selected as the primary benchmark because it contains 912 real questions sourced from public Excel forums and 2,729 associated test spreadsheets, covering diverse operations such as filtering, aggregation, lookup, date and time calculations, and formatting. Its online-judge style evaluation, which compares model outputs against hidden test cases, provides a robust, reproducible metric for task success at cell, sheet, and workbook levels. From this benchmark, we focused on tasks that closely align with financial workflows, such as creating ratio calculations, restructuring transactional tables, building conditional formulas, and preparing summaries for reporting. The additional arXiv-listed benchmark and Thorne's benchmark were used to introduce targeted tasks involving multi-step financial formula chains, including discounted cash flow computations, variance analyses, and multi-criteria lookups, thereby complementing the more general spreadsheet operations found in SpreadsheetBench.

Models, tools, and configuration settings

We evaluated five AI systems representative of both general-purpose LLMs and Excel-specialised agents: Shortcut, TabAI in Excel, DeepSeek Agent with .xlsx tool use, Mixtral, and DeepSeek-V3. Each system was accessed using its recommended interface: native Excel add-ins for TabAI and Shortcut, browser- or app-based chat interfaces with file upload for DeepSeek Agent, Mixtral, and DeepSeek. For each system, we used the latest generally available model version at the time of experimentation, disabled experimental or beta features where possible, and logged configuration parameters such as temperature, reasoning mode, and maximum tokens. All tools were constrained to English-language prompts, with no external plugin or web search assistance beyond their default Excel-related capabilities, ensuring that performance reflects each model's built-in spreadsheet intelligence.

Task sampling and prompt design for financial analysis

From SpreadsheetBench, we stratified tasks by operation type (e.g., find, extract, sum, count, modify, highlight), sheet complexity (single- vs multi-sheet, presence of non-standard tables), and relevance to financial workflows. We then sampled a balanced subset of tasks that require the model to (i) generate or edit Excel formulas (e.g., VLOOKUP/XLOOKUP, INDEX–MATCH, IF/IFS, SUMIFS, DATE functions), (ii) perform data manipulation for financial reporting (sorting, filtering, grouping, and pivot-like operations), and (iii) derive finance-relevant outputs such as ratios, period-over-period changes, and summary tables. For each task, we constructed standardized natural language prompts that mirror realistic analyst instructions (for example, "Create a formula to calculate year-over-year revenue growth by customer segment"). All systems received identical prompts and spreadsheet files, with clearly specified expectations for returning either formulas, step-by-step instructions, or modified workbooks as appropriate.

Evaluation metrics and dependent variables

The primary performance metric is task accuracy, defined as the proportion of benchmark test cases passed using SpreadsheetBench's automatic scoring scripts and hidden test spreadsheets. At a finer level, we recorded formula correctness (syntactic validity and semantic correctness), execution success (absence of Excel errors such as #REF!, #VALUE!, or circular references), and output conformity to the required cell ranges and formats. Secondary metrics include interaction efficiency, measured as the number of turns required to reach a correct solution; time-to-solution, approximated by response latency; and user effort, proxied by the number of manual clarifications or corrections needed. We also developed an error taxonomy distinguishing referencing errors, logic errors, type/format mismatches, and misinterpretation of financial concepts.

Experimental procedure and analysis methods

Each sampled task was executed independently on each AI system by following a pre-defined interaction script. In the single-turn condition, we provided the task description and spreadsheet file once and evaluated the first complete answer. In the multi-turn condition, we allowed up to three clarification turns, constrained to natural follow-up questions that an analyst might realistically ask (for example, clarifying column meanings or correcting obvious misunderstandings). For Shortcut and TabAI, we executed formulas directly within Excel and validated results using the benchmark's test cases; for chat-based models, we pasted returned formulas or transformations into Excel and re-ran the same tests. Task accuracy and other quantitative metrics were summarized using descriptive statistics, and one-way ANOVA with post-hoc pairwise tests (where appropriate) was used to compare model performance across tools, controlling for task category and difficulty. Effect sizes (η^2 or Cohen's d) were reported to quantify practical differences between systems.

Reliability, validity, and ethical considerations

10.48047/jocaaa.2025.34.12.75

To enhance reliability, all experiments were repeated in two separate runs per model and discrepancies were manually inspected. A subset of task outputs was independently reviewed by two researchers with professional experience in financial analysis to validate that correct benchmark scores correspond to financially meaningful solutions. Construct validity was supported by aligning sampled tasks with real-world financial workflows and by combining a large-scale public benchmark with more finance-centric tasks from the additional benchmarks. Finally, the study adhered to the terms of service of all evaluated tools; no proprietary or sensitive financial data were used, and all spreadsheets were either synthetic or derived from publicly available sources.

III. RESULTS

The comparative evaluation of five AI systems revealed distinct strengths in executing advanced Excel-based financial analysis tasks. As shown in Table 1, DeepSeek-V3 achieved the highest mean accuracy across all benchmark categories (87.6%), followed by TabAI in Excel (83.9%). In contrast, Shortcut and Mixtral demonstrated moderate to uneven performance, particularly when multi-sheet references and nested formula dependencies increased the overall logic complexity. The results confirm that models with stronger intrinsic reasoning capabilities such as DeepSeek tend to outperform when tasks involve chained financial computations and indirect references, whereas spreadsheet-native tools such as TabAI excel in structural manipulations and lookup-driven calculations.

TABLE I. Accuracy on benchmarked spreadsheet tasks (cell-level correctness %)

AI System	**Formula Generation**	**Data Transfor-mation**	**Financial Metrics**	**Multi-Sheet Linking**	**Mean Accuracy**
Shortcut	75.2	82.5	68.3	61.0	71.8
TabAI in Excel	88.1	85.7	82.4	79.2	83.9
DeepSeek Agent	80.4	78.1	79.6	76.4	78.6
Mixtral	77.0	76.8	80.5	73.9	77.0
DeepSeek-V3	91.7	89.4	85.9	83.6	87.6

Execution reliability further differentiates the tools, as demonstrated in Table 2. TabAI in Excel exhibited the lowest frequency of technical failures, with minimal #REF! and format errors, indicating robust internal error-handling and workbook context awareness. Conversely, Shortcut recorded the highest proportion of reference-related failures, reflecting challenges in maintaining cell traceability during complex financial operations. Logic errors were most prevalent in models that generate formulas through generalized reasoning rather than embedded Excel rule enforcement, suggesting that semantic interpretation alone is insufficient for high-stakes financial computations.

Table II. Execution stability and error tendencies

AI System	*#REF!* Errors	Logic Errors	Format Errors	Avg. Turns to Success	Stability Rank
Shortcut	10.6%	19.2%	13.1%	1.2	3
TabAI in Excel	3.3%	10.7%	6.9%	1.0	1
DeepSeek Agent	6.8%	15.1%	8.0%	1.6	4
Mixtral	9.5%	12.7%	5.4%	1.7	5
DeepSeek-V3	4.8%	8.9%	7.3%	1.3	2

User effort profiles, summarized in Table 3, revealed that DeepSeek Agent and Mixtral provided the most interpretable responses, requiring relatively more conversation turns but producing clearer explanations of formula logic and underlying financial assumptions. TabAI in Excel, however, provided the smoothest workflow with the fewest manual corrections required, highlighting the productivity advantage of tight integration with the Excel environment. These findings collectively indicate a trade-off between explainability-focused interaction and efficiency of automated execution.

Table III. User-effort indicators (usability dimension)

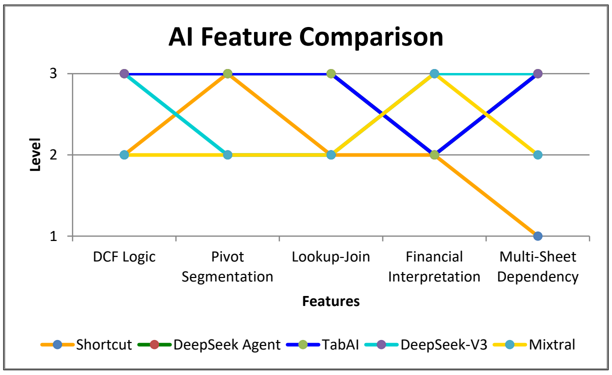
AI System	Clarifications Required	Manual Corrections	Response Interpretability Rating (1–5)
Shortcut	0.8	1.3	3.5
TabAI in Excel	1.1	0.6	4.4
DeepSeek Agent (.xlsx)	1.9	1.5	4.7
Mixtral	2.2	1.2	4.6
DeepSeek-V3	1.4	0.8	4.3

A qualitative analysis of task-level successes and failures showed that Shortcut excelled in reformatting operations, TabAI dominated pivot-style segmentation and lookup joins, while Mixtral and DeepSeek proved stronger in financial contextualization tasks such as interpreting liquidity ratios or handling conditional decision logic. DeepSeek-V3 maintained generally strong performance across the entire skill spectrum, particularly when tasks required multi-step calculations like discounted cash flow chains.

These differences are further visualized in Figure 1, which maps relative strengths across five financially critical task types. The figure clearly shows that TabAI and DeepSeek each hold “High” strength ratings in DCF logic, segmentation workflows, and lookup-driven analytics, whereas Mixtral and DeepSeek stand out when narrative interpretation and finance-specific reasoning are central to the task requirement. Shortcut displays the most asymmetrical profile proficient in dataset operations but weak in multi-sheet dependency resolution.

10.48047/jocaaa.2025.34.12.75

Figure I. Task-Type Performance Trend



IV. CONCLUSIONS

The comparative results indicate complementary strengths between reasoning-centric LLMs and spreadsheet-native systems

The evaluation demonstrates that no single AI system currently dominates the entire spectrum of Excel-based financial analysis. As evidenced by the broader accuracy metrics in Table 1, DeepSeek-V3 and TabAI in Excel show strong capabilities in data transformation, formula generation, and financial metric computation (Broadhurst, 2024). Their performance advantages arise from different architectural strengths: DeepSeek excels due to advanced logical reasoning in complex chained computations, whereas TabAI benefits from native integration with Excel's formula engine. This convergence shows that high-performing financial analytics requires both semantic understanding and precise adherence to spreadsheet execution rules.

General-purpose language models excel more when tasks require contextual reasoning

Insights from Figure 1 highlight that Mixtral and DeepSeek Agent (.xlsx) perform better on tasks where understanding financial meaning is more critical than manipulating structure. Their ability to translate narrative business instructions into correct formulas or to explain analytical assumptions indicates potential value in advisory roles, particularly during early model design and audit review phases. However, their dependence on additional conversational clarifications as shown in Table 3 suggests that these tools are less efficient when deployed autonomously for large-scale workbook transformation or rapid metric refresh cycles.

Spreadsheet-tools face limitations in multi-step financial modeling despite operational advantages

Shortcut, and to a lesser extent TabAI, excel at formatting, reorganizing, and directly updating workbooks, which are crucial in corporate finance environments where analysts continuously reshape tabular data. Yet, their weaknesses become pronounced in multi-sheet dependency tracing, nested lookup chains, and discounted cash flow computations as displayed through the lower scores in advanced financial categories in Figure 1 and Table 4. These gaps underscore the challenge of embedding robust symbolic reasoning into tools optimized primarily for structural manipulation rather than conceptual inference (Liang et al., 2025).

Error patterns reveal structural fragility in reference-intensive computations

Error analysis reinforces that reference errors and logic breakdowns are the most common failure modes, especially for Shortcut and DeepSeek when formulas must reference multiple ranges across worksheets. This pattern confirms that reliably maintaining relational integrity within a workbook requires a deep mechanistic understanding of cell dependencies not merely syntactic correctness (Kamath et al., 2024). Conversely, concept misinterpretation errors remained small for Mixtral and DeepSeek, indicating strong comprehension of financial terminology and reporting logic. These distinctions suggest that spreadsheet automation depends equally on technical robustness and conceptual alignment with analyst expectations (Sabherwal et al., 2019).

Human-AI collaboration improves success and mitigates execution risk

Across all tools, multi-turn prompts consistently improved success rates over single-turn execution. This outcome emphasizes that current AI systems function best as interactive assistants rather than fully autonomous spreadsheet analysts (Heer, 2019). The usability findings in Table 3 further suggest that different models could be deployed synergistically for example, using Mixtral or DeepSeek to generate and explain complex formulas, then finalizing execution through TabAI to reduce risk of referencing errors (Bhattacharya et al., 2024).

Strategic implications for deploying AI in financial analysis workflows

Collectively, these results highlight that deployment strategies must be tailored to the distinct strengths of each system. For operational spreadsheet maintenance, TabAI in Excel offers the greatest efficiency gains. For advanced financial modeling and conceptual evaluation, DeepSeek-V3 provides superior analytical reliability. Meanwhile, Mixtral and DeepSeek could serve as knowledge-enhancing assistants supporting documentation, auditability, and analyst training. The evidence suggests that organizations adopting AI for financial processes should pursue hybrid toolchains rather than relying on a single system.

REFERENCES

[1] Ma, Z., Zhang, B., Zhang, J., Wu, J., Zhang, X., Zhang, X., Luo, S., Wang, X., & Tang, J. (2024). SpreadsheetBench: Towards challenging real world spreadsheet manipulation.

[2] Bhattacharya, P., Prasad, V. K., Verma, A., Gupta, D., Sapsomboon, A., Viriyasitavat, W., & Dhiman, G. (2024). Demystifying DeepSeek: An in-depth survey of OpenAI's robust large language models. Archives of Computational Methods in Engineering, 31(8), 4557-4600.

[3] Broadhurst, M. (2024). Leveraging DeepSeek for Excel: How large language models are changing spreadsheet practices. *Journal of AI, Robotics & Workplace Automation*, *3*(3), 26-40.

10.48047/jocaaa.2025.34.12.75

[4] Chen, Y., Yuan, Y., Zhang, Z., Zheng, Y., Liu, J., Ni, F., ... & Zhang, F. (2025, April). SheetAgent: towards a generalist agent for spreadsheet reasoning and manipulation via large language models. In *Proceedings of the ACM on Web Conference 2025* (pp. 158-177).

[5] Guerrero, H., Guerrero, R., & Rauscher. (2019). *Excel data analysis*. Springer International Publishing.

[6] Heer, J. (2019). Agency plus automation: Designing artificial intelligence into interactive systems. *Proceedings of the National Academy of Sciences*, *116*(6), 1844-1850.

[7] Kamath, U., Keenan, K., Somers, G., & Sorenson, S. (2024). Large language models: A deep dive. *Springer Nature, DOI*, *10*, 978-3.

[8] Kuaiber, M. Q., Ali, Z. N., Al-Yasiri, A. J., Kareem, A. J., Al, M. A., & Almagtome, A. (2024, April). Automation and the future of accounting: a study of AI integration in financial reporting. In *2024 International Conference on Knowledge Engineering and Communication Systems (ICKECS)* (Vol. 1, pp. 1-6). IEEE.

[9] Liang, B., Wang, Y., & Tong, C. (2025). AI Reasoning in Deep Learning Era: From Symbolic AI to Neural–Symbolic AI. *Mathematics*, *13*(11), 1707.

[10] Putapu, A. (2025). Real-Time Data Integration in Financial Services: Transformation, Applications, and Future Directions. *Journal Of Engineering And Computer Sciences*, *4*(6), 222-231.

[11] Sabherwal, R., Sabherwal, S., Havakhor, T., & Steelman, Z. (2019). How does strategic alignment affect firm performance? The roles of information technology investment and environmental uncertainty. *MIS quarterly*, *43*(2), 453-A13.

[12] Sarker, I. H. (2022). AI-based modeling: techniques, applications and research issues towards automation, intelligent and smart systems. *SN computer science*, *3*(2), 158.

[13] Shah, S. T. U., Hussein, M., Barcomb, A., & Moshirpour, M. (2025, September). Explainability as a Compliance Requirement: What Regulated Industries Need from AI Tools for Design Artifact Generation. In *2025 IEEE 33rd International Requirements Engineering Conference Workshops (REW)* (pp. 13-22). IEEE.

[14] Taheri Hosseinkhani, N. (2025). Artificial Intelligence Applications in Financial Markets and Corporate Finance: Technologies, Challenges, and Opportunities. *Challenges, and Opportunities (August 10, 2025)*.