A Comprehensive Analysis of Platform-Enhanced LLM Accuracy Across 20 Critical Due Diligence Use Cases

Authors
Maya Boeye, Steiner Williams, Aubrey Inmon, Jeondo Lee, and Jennifer Venus

ToltIQ

March 2026

Table of Contents

1. Executive Summary 3
2. Introduction 4
2.1 The AI Transformation in Private Equity
2.2 Gaps in Existing Research
2.3 Study Overview
2.4 Research Objectives
3. Methodology 6
3.1 Dataset Composition
3.2 Testing Framework
3.3 Models Evaluated
3.4 Accuracy Measurement
3.5 Why Vals AI?
4. Results 9
4.1 General Private Markets Use Cases
4.2 Debt-Focused Use Cases
4.3 Overall Platform Distribution
5. Analysis 14
5.1 Platform Lift
5.2 Use Case Patterns in Platform Lift
5.3 Limitations and Future Research
6. Conclusion 19
Appendix A: Use Case Definitions 20
Appendix B: Dataset and Confidentiality 22
Appendix C: Prior Benchmarking Results.. 23
Appendix D: Supplementary Visualizations... 25
Appendix E: Acknowledgments and Disclosure 27
References 28

1. Executive Summary

Despite accelerating LLM adoption in private equity, practitioners lack granular, reproducible benchmarks mapping model performance to specific due diligence workflows. Existing evaluations report aggregate capabilities without distinguishing between use cases, and most draw on publicly available datasets vulnerable to training-data contamination. This study addresses both gaps.

We evaluated GPT 5.1, GPT 5.2, Claude Sonnet 4.5, and Claude Opus 4.5 across 20 PE due diligence use cases, comparing standalone API accuracy against performance through ToltIQ's custom architecture. The evaluation used the Vals AI CorpFin V2 Public and Private Validation datasets, independent benchmarks comprising 360 prompts derived from 18 financial documents. Each prompt was replicated five times per model, yielding 14,400 scored responses.

The principal finding is that deployment architecture is at least as consequential as model selection. Three of four models exhibited statistically significant accuracy gains through ToltIQ’s custom architecture ($p<0.0125$), with improvements ranging from 6.63 to 9.63 percentage points. Opus 4.5 achieved the highest aggregate accuracy at 85.11%, surpassing 80% in 17 of 20 use cases. Sonnet 4.5 followed at 83.54%, GPT 5.1 at 80.45%, while GPT 5.2 registered no significant improvement.

Performance varied systematically by use case. Document-navigation workflows showed the largest custom architecture gains, with CIM Analysis and VDR Analysis each exceeding 13 percentage points across three models. Debt-focused workflows produced higher peak accuracy and cross-model consistency. Synthesis-intensive tasks, Quality of Earnings, Credit Term Benchmarking, and Management Team Questionnaires, remained below 81% across all configurations, indicating continued need for substantive human oversight. These findings provide PE firms with use-case-level data to inform model selection, automation boundaries, and deployment decisions.

2. Introduction

2.1 The AI Transformation in Private Equity

Private equity due diligence is undergoing a structural shift. The traditional model, in which analyst teams manually review thousands of pages of CIMs, credit agreements, financial models, and legal documents under compressed deal timelines, is increasingly untenable as deal volume grows and sponsor competition intensifies (Deloitte, 2025).

Advances in large language models have dramatically improved the ability of AI systems to understand complex financial and legal language, extract relevant information, and synthesize insights across multiple documents (Nie et al., 2024). As a result, dealmakers are facing compressed timelines, and Private Equity sponsors need to adopt AI capabilities across their entire portfolios within 3 to 5 years (PwC, 2025).

2.2 Gaps in Existing Research

The high-stakes nature of PE investment decisions demands rigorous validation of AI capabilities before deployment. A single missed covenant provision or misinterpreted financial term can have significant consequences for deal structuring and portfolio performance (Bellucci & McCluskey, 2017; Bain & Company, 2026). Yet despite growing adoption, the research available to practitioners making deployment decisions remains inadequate in three specific ways.

Most published research on AI in financial services characterizes performance in broad or relative terms, without providing reproducible, numerical results tied to specific workflows or platforms (Mohsin, 2025; Kong et al., 2025). Academic surveys of finance practitioners have found that nearly half feel current research provides meaningful guidance on LLM evaluation only to a small extent (Kong et al., 2025). For PE firms weighing whether to deploy a specific model or platform, the absence of concrete performance data presents a genuine obstacle to informed decision-making.

Existing evaluations tend to treat AI-assisted due diligence as a single capability rather than examining how performance varies across discrete use cases such as Covenant Analysis, Quality of Earnings Analysis, or Credit Agreement Term Extraction. Without this granularity, practitioners have limited ability to determine which workflows are candidates for automation and which require continued human oversight (Mohsin, 2025; Kong et al., 2025).

The last gap concerns the quality of evaluation data itself. Most financial benchmarks draw on publicly available sources, introducing the risk that models have been evaluated on data they were also trained on, undermining the reliability of reported accuracy figures (Vals AI, 2025).

The Vals AI CorpFin V2 validation set used in this study addresses this concern through a proprietary, non-public dataset developed with domain experts.

2.3 Study Overview

This paper presents a use-case-level benchmark evaluation of leading LLMs for private equity due diligence workflows. To our knowledge, it is one of the first studies to provide reproducible accuracy data mapped to the discrete tasks that define PE deal execution. We evaluated GPT 5.1, GPT 5.2, Claude Sonnet 4.5, and Claude Opus 4.5 across 20 PE-specific use cases using the Vals AI CorpFin V2 validation set, an independent benchmark comprising 360 prompts derived from 18 financial documents and developed in collaboration with domain experts. Each prompt was replicated five times per model, yielding 14,400 total evaluated responses.

Each model was tested under two conditions: as a standalone API implementation and as deployed through ToltIQ's custom ingestion process and retrieval-augmented generation (RAG) pipeline. ToltIQ's implementation of this architecture was designed specifically for the structural and linguistic characteristics of PE deal documents. This dual-condition design allows the study to isolate the contribution of deployment architecture to accuracy outcomes, independent of underlying model capability.

2.4 The Role of RAG Systems

Retrieval-augmented generation (RAG) is an architecture that combines a document retrieval system with a large language model, dynamically sourcing relevant content at inference time to ground model responses in source material rather than training data alone (Lewis et al., 2020). Rather than relying on parametric knowledge encoded during pretraining, RAG-enabled systems retrieve passage-level content from a document corpus at the time a query is issued, supplying the model with grounded context for each response (Gao et al., 2024).

For private equity applications, this architecture addresses three limitations inherent to standalone LLM deployment. First, it grounds responses in actual source documents, reducing the risk of hallucinated or outdated information in high-stakes analytical outputs. Second, it enables analysis across document sets that would exceed standard context window limits, a meaningful constraint in deal environments where analysts routinely work across hundreds of documents simultaneously. Third, it allows the document corpus to be updated continuously without retraining the underlying model, supporting the dynamic document environments typical of active deal processes (Gao et al., 2024; Sharma, 2025).

The effectiveness of a RAG system depends substantially on implementation choices: how documents are chunked, how embeddings are constructed, how retrieval is ranked, and how retrieved context is incorporated into prompts (Gao et al., 2024). ToltIQ's RAG pipeline is paired with a proprietary custom ingestion process. Together they were designed specifically for the structural and linguistic characteristics of PE deal documents, incorporating intelligent document chunking, semantic embedding, vector-based retrieval, and context-aware prompt construction. This study evaluates the accuracy contribution of that pipeline across 20 PE due diligence use cases, isolating deployment architecture as an independent variable in model performance.

2.5 Research Objectives

This research was designed to answer several critical questions facing PE firms considering AI-assisted due diligence:

1. How much does ToltIQ's deployment architecture contribute to accuracy outcomes?
2. Which LLMs deliver the highest accuracy when deployed through ToltIQ for PE due diligence tasks?
3. How does model performance vary across different use case categories of due diligence work, both within ToltIQ and when tested directly through the API?

3. Methodology

3.1 Dataset Composition

The evaluation utilized the Vals AI CorpFin V2 validation set, an independent benchmark comprising 360 prompts derived from 18 curated financial documents. The benchmark was selected for this study on the basis of its domain specificity and data integrity (Vals AI, 2025).

The CorpFin V2 benchmark offers three context configurations that vary how much of each source document is provided to the model. We selected the Max Fitting Context condition, which includes the largest possible document chunk, starting from the first page, that fits within each model's context window. This means longer-context models receive more information than those with shorter windows. This configuration was chosen because it most closely approximates ToltIQ client environments, where models receive substantial document context and must locate relevant information within larger passages

The 360 prompts were mapped to 20 PE due diligence use case categories, with individual questions assigned to up to 10 use cases based on their analytical scope. The benchmark questions span a diverse range of task complexity representative of real-world PE workflows. At the simpler end, questions test basic extraction of terms and numbers, for example, "Who is the borrower's legal counsel?" and summarization and interpretation, such as, "How will the loan proceeds be used?" More demanding questions require numeric reasoning and calculation, "How much initial debt capacity is available to the Borrower on day one?" and multi-section reference and definition chaining, where the model must trace definitions across sections to arrive at a correct answer. The most challenging questions test market standard judgment, requiring opinions on whether terms are unusual relative to comparable transactions, and industry jargon comprehension, using terms with commonly understood meanings that are rarely defined explicitly in agreements. For instance, "Does the contract contain a Chewy Blocker?" requires the model to recognize a specific clause type that prevents a subsidiary from being released from its debt obligations. This range ensures the benchmark evaluates not only basic information retrieval but also the higher-order reasoning and domain knowledge required for effective due diligence (Vals AI, 2025).

3.2 Testing Framework

To ensure a controlled comparison between ToltIQ-deployed and standalone model performance, each document in the benchmark set was evaluated in an isolated single-document session. Document-specific prompts were executed within that restricted context. All prompts were formatted identically across models and presented in the same fixed order to eliminate prompt-level variability as a confounding factor.

For the standalone baseline, models were accessed directly via API through the Vals AI platform using default inference parameters. ToltIQ-deployed models were tested using default platform settings with no manual parameter adjustment, ensuring the accuracy gains reported reflect standard deployment conditions rather than optimized configurations. Testing was conducted from January 15, 2026 to February 18, 2026.

All responses were evaluated using the standardized scoring framework developed by Vals AI, ensuring consistent criteria across both test conditions. Where automated evaluation produced ambiguous scores, responses were reviewed manually to ensure scoring accuracy. Each prompt was replicated five times per model, yielding 7,200 responses for the ToltIQ platform and 7,200 for standalone implementations, for a combined total of 14,400 evaluated responses.

To assess whether accuracy differences between platforms were statistically meaningful, a Student's t-test was conducted for each model comparing its standalone accuracy against its ToltIQ accuracy. To account for multiple comparisons across the four models tested, a Bonferroni correction was applied, yielding an adjusted significance threshold of $\alpha = 0.0125$.

3.3 Models Evaluated

Four models were selected as representing the leading commercial LLMs widely deployed in PE workflows at the time of testing: two from OpenAI (GPT 5.1 and GPT 5.2) and two from Anthropic (Claude Sonnet 4.5 and Claude Opus 4.5). Several models released during the preparation of this paper, including Claude Sonnet 4.6, Claude Opus 4.6, Gemini 3.1, GPT 5.3, and GPT 5.4 were not included in this analysis. Incorporating them was not feasible without compromising the integrity of the testing framework. These models will be assessed in a follow-up study using the same methodology.

GPT 5.1 was OpenAI’s flagship model at the time of evaluation (OpenAI, 2025a); GPT 5.2, released in December 2025, succeeded it with improvements to long-context reasoning and agentic tool use (OpenAI, 2025b). Both OpenAI models share a 400,000-token context window (OpenAI, 2025a, 2025b). On the Anthropic side, Sonnet 4.5, released in September 2025, is optimized for high-throughput production workloads and complex agentic coding tasks (Anthropic, 2025a). Opus 4.5, released in November 2025, was Anthropic's most capable model at the time this benchmark was conducted, designed for sustained multi-step reasoning and high-accuracy performance on complex professional workflows (Anthropic, 2025b). Both Claude models support a 200,000-token context window (Anthropic, 2025a, 2025b). This difference is relevant to interpreting the results under the Max Fitting Context condition, as OpenAI models received more raw document content than Claude models due to their larger context windows.

3.4 Accuracy Measurement

Each response was scored based on whether it matched the established ground truth, which was developed by Vals AI in collaboration with financial analysts, legal professionals, and academics, and validated using Vals AI's standardized evaluation framework. Responses were scored using this framework, employing Vals AI's proprietary LLM-based judge under deterministic conditions across both standalone and ToltIQ-deployed responses.

Accuracy was chosen as the sole evaluation metric for this study because in PE due diligence, the operational question is whether the model's answer is correct, not whether it is fluent, well-structured, or partially right. Binary scoring provides the clearest signal for deployment decisions, though it necessarily discards granularity on borderline or partially correct responses. This conservative approach means that a response containing the correct covenant but an incorrect threshold would be scored the same as a fully incorrect answer.

Because individual questions were assigned to up to 10 use cases based on their analytical scope, a single correct response contributes to the accuracy score of each use case to which that question is mapped. Use case accuracy figures are therefore not fully independent of one another. For each prompt, accuracy was calculated separately across the five replicates, and the final accuracy score represents the mean of those five replicate-level results.

3.5 Why Vals AI?

Why Vals AI?

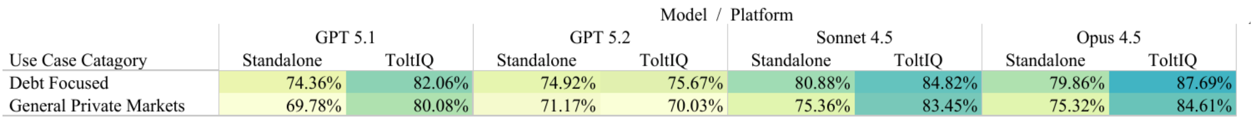
We selected the Vals AI CorpFin (v2) benchmark for several important reasons:

- Independent validation: As a third-party benchmark, Vals AI public and private validation benchmark that we performed provides objective evaluation criteria not influenced by any model provider.
- Domain specificity: Unlike general-purpose benchmarks, CorpFin V2 focuses specifically on corporate finance tasks relevant to PE due diligence workflows.
- Data integrity: The private validation set is not publicly available, reducing the risk that models have been trained on the test data.
- Industry credibility: Vals AI benchmarks are developed with input from practitioners at leading financial institutions and have been referenced in academic and industry publications.

4. Results

To provide actionable guidance for workflow design, the 20 use cases evaluated in this study were organized into two categories: General Private Markets and Debt-Focused. General private markets use cases cover workflows common across deal execution, portfolio monitoring, and investor reporting. Debt-focused use cases address credit-specific workflows including credit agreement analysis, covenant monitoring, and capital structure evaluation. The use cases within each category are ranked by priority based on input from ToltIQ's in-house private equity experts (Table 1).

Table 1. Aggregate accuracy (%) by use case category, model, and platform. Scores represent the mean accuracy across all use cases within each category, calculated from five replicates per prompt.



	Model / Platform							
	GPT 5.1		GPT 5.2		Sonnet 4.5		Opus 4.5	
Use Case Catagory	Standalone	ToltIQ	Standalone	ToltIQ	Standalone	ToltIQ	Standalone	ToltIQ
Debt Focused	74.36%	82.06%	74.92%	75.67%	80.88%	84.82%	79.86%	87.69%
General Private Markets	69.78%	80.08%	71.17%	70.03%	75.36%	83.45%	75.32%	84.61%

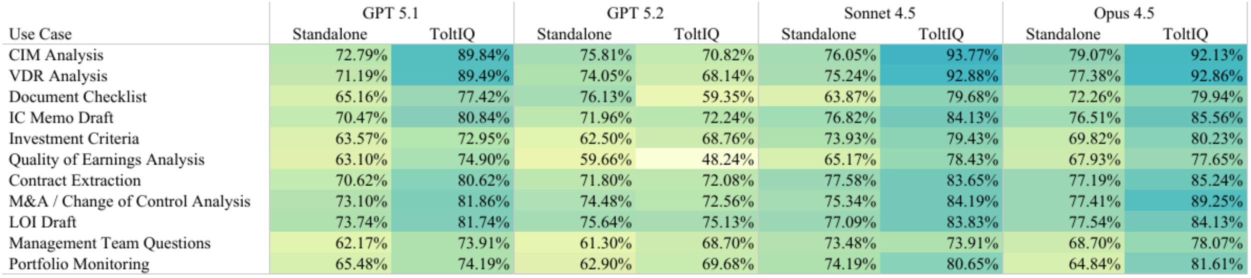
The sections that follow present detailed accuracy results by use case within each category (4.1 and 4.2), followed by an overall performance distribution across all 20 use cases (4.3).

4.1 General Private Markets Use Cases

This category encompasses 11 use cases covering workflows common across private markets deal execution, portfolio monitoring, valuation, and investor reporting.

Across these 11 use cases, accuracy on the ToltIQ platform ranged from 48.24% to 93.77%. The highest individual use case scores were achieved by Sonnet 4.5 on CIM Analysis (93.77%), Opus 4.5 on VDR Analysis (92.86%), GPT 5.1 on CIM Analysis (89.84%), and GPT 5.2 on LOI Draft (75.13%). Opus 4.5 demonstrated the narrowest performance range, with 8 of 11 use cases exceeding 80%. Sonnet 4.5 exceeded 80% in 7 of 11 use cases, GPT 5.1 in 6 of 11, and GPT 5.2 in none. The use cases with the lowest accuracy across all models were Quality of Earnings Analysis, Management Team Questionnaire, and Investment Criteria (Table 2).

Table 2. Accuracy (%) by general private markets use case, model, and platform. Values represent the mean accuracy across five replicates per prompt. Blue shading indicates higher accuracy; yellow indicates lower accuracy.



	GPT 5.1		GPT 5.2		Sonnet 4.5		Opus 4.5	
Use Case	Standalone	ToltIQ	Standalone	ToltIQ	Standalone	ToltIQ	Standalone	ToltIQ
CIM Analysis	72.79%	89.84%	75.81%	70.82%	76.05%	93.77%	79.07%	92.13%
VDR Analysis	71.19%	89.49%	74.05%	68.14%	75.24%	92.88%	77.38%	92.86%
Document Checklist	65.16%	77.42%	76.13%	59.35%	63.87%	79.68%	72.26%	79.94%
IC Memo Draft	70.47%	80.84%	71.96%	72.24%	76.82%	84.13%	76.51%	85.56%
Investment Criteria	63.57%	72.95%	62.50%	68.76%	73.93%	79.43%	69.82%	80.23%
Quality of Earnings Analysis	63.10%	74.90%	59.66%	48.24%	65.17%	78.43%	67.93%	77.65%
Contract Extraction	70.62%	80.62%	71.80%	72.08%	77.58%	83.65%	77.19%	85.24%
M&A / Change of Control Analysis	73.10%	81.86%	74.48%	72.56%	75.34%	84.19%	77.41%	89.25%
LOI Draft	73.74%	81.74%	75.64%	75.13%	77.09%	83.83%	77.54%	84.13%
Management Team Questions	62.17%	73.91%	61.30%	68.70%	73.48%	73.91%	68.70%	78.07%
Portfolio Monitoring	65.48%	74.19%	62.90%	69.68%	74.19%	80.65%	64.84%	81.61%

Brief descriptions of each use case are provided below; full definitions are available in Appendix A.

CIM Analysis - Extract and evaluate financial projections, competitive positioning, and EBITDA definitions from Confidential Information Memorandums.

VDR Analysis - Extract financial documentation including revenue, EBITDA, working capital, and CapEx data from Virtual Data Rooms.

Document Checklist - Identify conditions precedent, required consents, and closing requirements.

IC Memo Draft - Summarize ownership structure, financial metrics, and strategic rationale for Investment Committee preparation.

Investment Criteria - Evaluate targets against financial thresholds including revenue, EBITDA margins, and leverage limits.

Quality of Earnings Analysis - Analyze EBITDA definitions, permitted add-backs, and adjustment caps.

Contract Extraction - Extract standardized terms including duration, renewal conditions, and governing law across multiple agreements.

M&A / Change of Control Analysis - Analyze change of control definitions, triggers, and consequences within credit agreements.

LOI Draft - Extract key deal terms for Letter of Intent preparation.

Management Team Questionnaire - Generate questions for management on unusual provisions and areas requiring clarification.

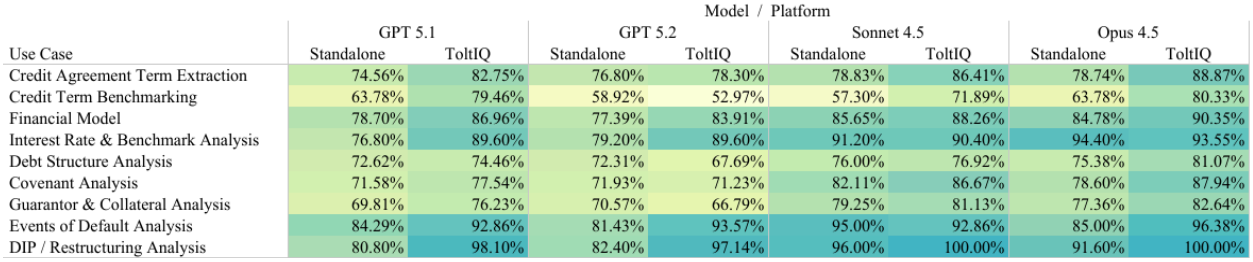
Portfolio Monitoring - Track performance metrics, covenant compliance, and early warning indicators across portfolio companies.

4.2 Debt-Focused Use Cases

This category encompasses 9 use cases covering workflows specific to credit and lending, including credit agreement analysis, covenant monitoring, collateral assessment, and capital structure evaluation.

Across these 9 use cases, accuracy on the ToltIQ platform ranged from 52.97% to 100.00%. Both Sonnet 4.5 and Opus 4.5 achieved 100.00% on DIP/Restructuring Analysis, followed by GPT 5.1 at 98.10% and GPT 5.2 at 97.14% on the same use case. Opus 4.5 was the only model to exceed 80% on all 9 debt-focused use cases, with 4 exceeding 90%. Sonnet 4.5 exceeded 80% in 7 of 9, GPT 5.1 in 5 of 9, and GPT 5.2 in 4 of 9. Credit Term Benchmarking was the lowest-performing use case across all models, with scores ranging from 52.97% (GPT 5.2) to 80.33% (Opus 4.5) (Table 3).

Table 3. Accuracy (%) by debt-focused use case, model, and platform. Values represent the mean accuracy across five replicates per prompt. Blue shading indicates higher accuracy; yellow indicates lower accuracy.



	Model / Platform							
	GPT 5.1		GPT 5.2		Sonnet 4.5		Opus 4.5	
Use Case	Standalone	ToltIQ	Standalone	ToltIQ	Standalone	ToltIQ	Standalone	ToltIQ
Credit Agreement Term Extraction	74.56%	82.75%	76.80%	78.30%	78.83%	86.41%	78.74%	88.87%
Credit Term Benchmarking	63.78%	79.46%	58.92%	52.97%	57.30%	71.89%	63.78%	80.33%
Financial Model	78.70%	86.96%	77.39%	83.91%	85.65%	88.26%	84.78%	90.35%
Interest Rate & Benchmark Analysis	76.80%	89.60%	79.20%	89.60%	91.20%	90.40%	94.40%	93.55%
Debt Structure Analysis	72.62%	74.46%	72.31%	67.69%	76.00%	76.92%	75.38%	81.07%
Covenant Analysis	71.58%	77.54%	71.93%	71.23%	82.11%	86.67%	78.60%	87.94%
Guarantor & Collateral Analysis	69.81%	76.23%	70.57%	66.79%	79.25%	81.13%	77.36%	82.64%
Events of Default Analysis	84.29%	92.86%	81.43%	93.57%	95.00%	92.86%	85.00%	96.38%
DIP / Restructuring Analysis	80.80%	98.10%	82.40%	97.14%	96.00%	100.00%	91.60%	100.00%

Brief descriptions of each use case are provided below; full definitions are available in Appendix A.

Credit Agreement Term Extraction - Identify and summarize credit facility types, amounts, purposes, key covenants, and lender protections.

Credit Term Benchmarking - Compare credit agreement terms against market standards and flag borrower-friendly, lender-friendly, or unusual provisions.

Financial Model Analysis - Extract interest rate mechanics, amortization schedules, prepayment percentages, and cash flow sweep parameters.

Interest Rate & Benchmark Analysis - Analyze SOFR/LIBOR benchmarks, applicable margins, floors, and benchmark transition provisions.

Debt Structure Analysis - Analyze lien priorities, incremental facility capacity, debt incurrence baskets, and the overall debt stack hierarchy.

Covenant Analysis - Analyze financial maintenance covenants including leverage ratios, coverage tests, cure rights, and covenant holiday periods.

Guarantor & Collateral Analysis - Review guarantor requirements, collateral packages, security interests, and release conditions.

Events of Default Analysis - Identify default triggers, cure periods, materiality thresholds, and acceleration rights.

DIP/Restructuring Analysis - Analyze debtor-in-possession financing provisions including priming liens, roll-up mechanics, and restructuring-specific terms.

4.3 Overall Performance Distribution

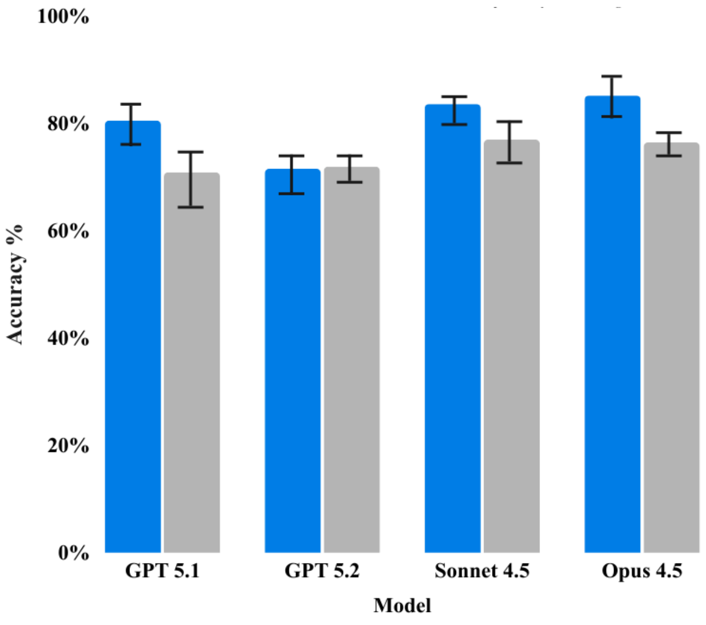


Figure 1. Mean accuracy (± SE) by model and platform. Grey bars represent standalone API performance; blue bars represent performance on the ToltIQ platform. Accuracy reflects the mean of five replicates per prompt. Platform differences were statistically significant for GPT 5.1, Sonnet 4.5, and Opus 4.5 ($p < 0.0125$, Bonferroni-corrected). The difference for GPT 5.2 was not significant.

Across all 20 use cases, Opus 4.5 achieved the highest aggregate accuracy on the ToltIQ platform at 85.11% (SE ± 0.59%), followed by Sonnet 4.5 at 83.54% (± 0.36%), GPT 5.1 at 80.45% (± 0.87%), and GPT 5.2 at 71.47% (± 0.69%). On standalone implementations, the performance hierarchy differed: Sonnet 4.5 led at 76.91% (± 0.55%), followed by Opus 4.5 at 76.38% (± 0.18%), GPT 5.2 at 71.89% (± 0.30%), and GPT 5.1 at 70.82% (± 0.84%).

The difference between standalone and ToltIQ accuracy was statistically significant for three of four models under the Bonferroni-corrected threshold (α = 0.0125): GPT 5.1 (+9.63 pp), Opus 4.5 (+8.73 pp), and Sonnet 4.5 (+6.63 pp). GPT 5.2 registered a non-significant decline of 0.42 pp (Figure 1).

On the ToltIQ platform, Opus 4.5 exceeded 80% accuracy in 17 of 20 use cases, Sonnet 4.5 in 14, GPT 5.1 in 11, and GPT 5.2 in 4. Under standalone conditions, the corresponding counts were 4, 5, 2, and 2 (Figure 2).

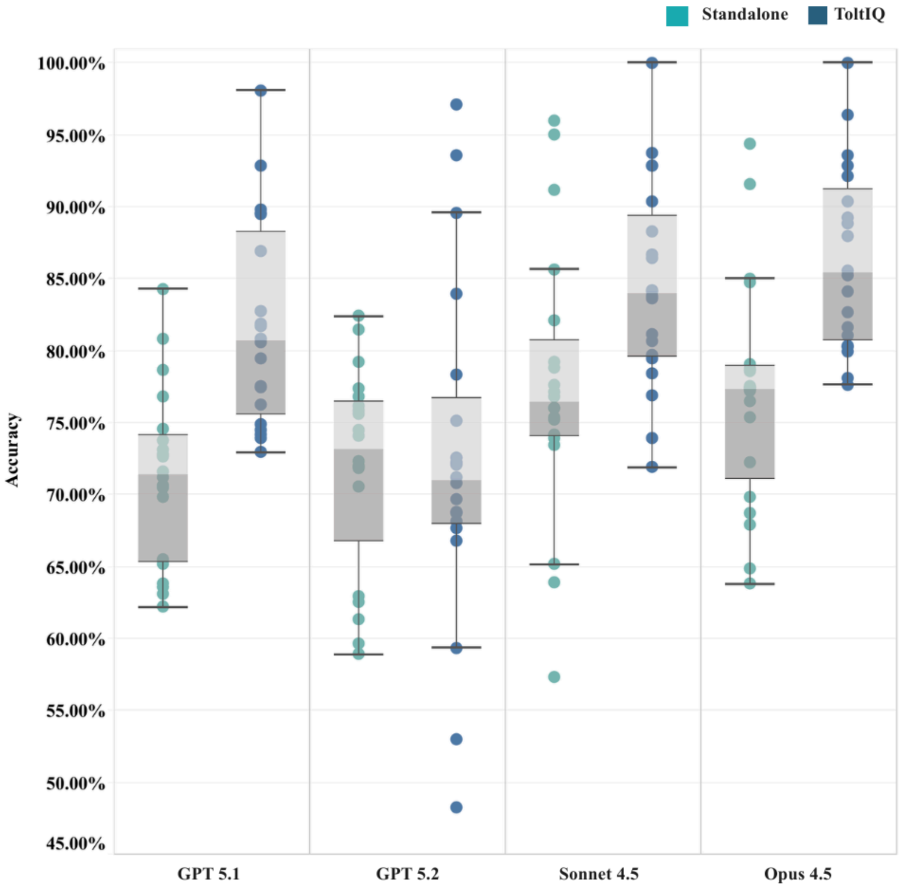


Figure 2. Distribution of use case accuracy scores by model and platform. Each box represents the interquartile range across all 20 use cases, combining both general private markets and debt-focused categories.

5. Analysis

5.1 Platform Lift

The results presented in Section 4.3 demonstrate that deployment through ToltIQ's purpose-built document ingestion and RAG pipeline produced statistically significant accuracy improvements for three of four models tested, with gains ranging from 6.63 to 9.63 pp. The magnitude and consistency of these improvements, which were observed across 7,200 evaluated responses per platform, with five independent replicates per prompt, indicate that the accuracy gains are consistent with structural characteristics of the retrieval and context construction pipeline rather than stochastic output variability. If the gains were driven by randomness, they would not replicate consistently across models with fundamentally different architectures.

The performance hierarchy also shifted between platforms. Under standalone conditions, Sonnet 4.5 and Opus 4.5 were nearly tied (76.91% and 76.38% respectively), while GPT 5.1 trailed at 70.82%. On the ToltIQ platform, Opus 4.5 pulled ahead decisively at 85.11%, opening a 1.57 percentage point gap over Sonnet 4.5 and a 4.66 point gap over GPT 5.1. Notably, Sonnet 4.5 and Opus 4.5 achieved higher platform accuracy despite receiving less input context than OpenAI models under the Max Fitting Context condition. This may reflect the retrieval architecture's ability to compensate for window size differences, though the two factors cannot be fully disentangled within the current study design.This suggests that certain models benefit more from RAG-enhanced retrieval than others, and that standalone performance is an incomplete predictor of deployed performance.

GPT 5.2 is the sole outlier, registering a non-significant decline of 0.42 pp on the ToltIQ platform despite standalone accuracy (71.89%) broadly comparable to GPT 5.1 (70.82%). OpenAI’s own system card documents performance regressions in GPT 5.2 relative to GPT 5.1 across multiple evaluation categories, including the illicit, harassment, and hate benchmarks, as well as reduced robustness to jailbreak attacks (OpenAI, 2025c). The system card further notes elevated deception rates in specific domains, including a greater willingness to hallucinate answers when faced with missing inputs (OpenAI, 2025c). These documented regressions are consistent with the underperformance observed in this study. Based on these results, ToltIQ reverted to GPT 5.1 as its default OpenAI model.

Table 4. Mean accuracy (± SE) by model and platform, with net improvement and statistical significance. Platform differences were assessed using a Student's t-test with Bonferroni-corrected threshold (α = 0.0125).

Model	Standalone LLM (± SE)	LLM within ToltIQ (± SE)	Net Improvement	P-Value
GPT 5.1	70.82 ± 0.84%	80.45 ± 0.87%	+9.63 ± 1.21 pp	< 0.001
GPT 5.2	71.89 ± 0.30%	71.47 ± 0.69%	-0.42 ± 0.75 pp	0.5521
Sonnet 4.5	76.91 ± 0.55%	83.54 ± 0.36%	+6.63 ± 0.66 pp	< 0.01
Opus 4.5	76.38 ± 0.18%	85.11 ± 0.59%	+8.73 ± 0.61 pp	< 0.001

5.2 Use Case Patterns in Platform Lift

The accuracy improvements observed in Section 4.3 were not uniformly distributed across use cases. The largest gains were concentrated in document-navigation workflows: tasks requiring models to locate, synthesize, and reason across information distributed throughout long, structurally heterogeneous documents. CIM Analysis and VDR Analysis each showed improvements exceeding 13 pp across GPT 5.1, Sonnet 4.5, and Opus 4.5. These workflows place the highest demand on retrieval precision, and the consistency of lift across three architecturally distinct models suggests that the gains are driven by ToltIQ's retrieval and context construction pipeline rather than by any model-specific interaction. Conversely, use cases involving standardized, uniformly formatted provisions such as DIP/Restructuring Analysis and Events of Default Analysis showed minimal platform lift, as standalone models already performed at or near ceiling on these tasks. This pattern indicates that custom architecture contributes its greatest marginal value precisely where standalone models are most constrained: heterogeneous, multi-section documents where relevant information is distributed unpredictably.

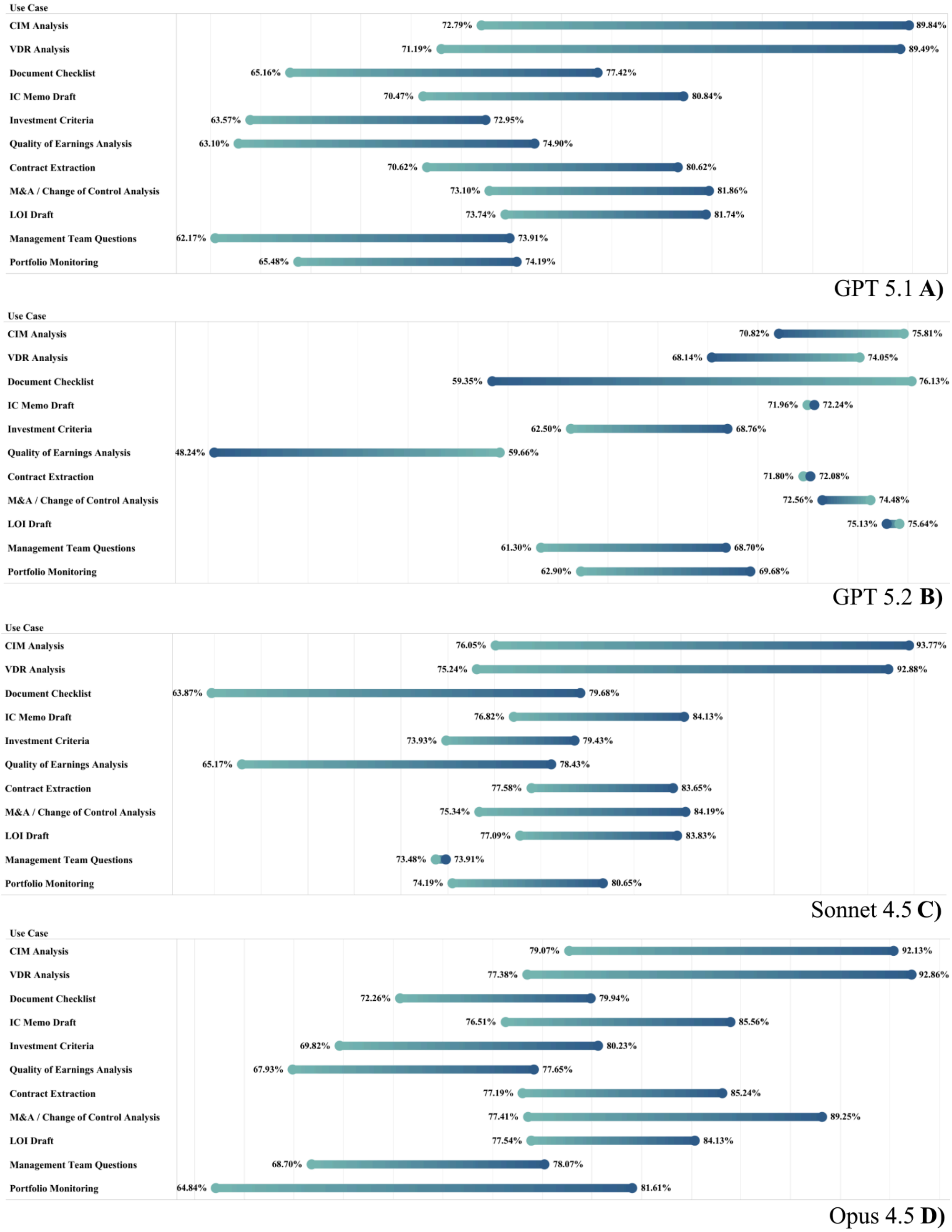


Figure 3. Accuracy (%) by general private markets use case, comparing standalone and ToltIQ platform performance for **a)** GPT 5.1, **b)** GPT 5.2, **c)** Sonnet 4.5, and **d)** Opus 4.5. The standalone score is in Teal while ToltIQ score is in dark blue.

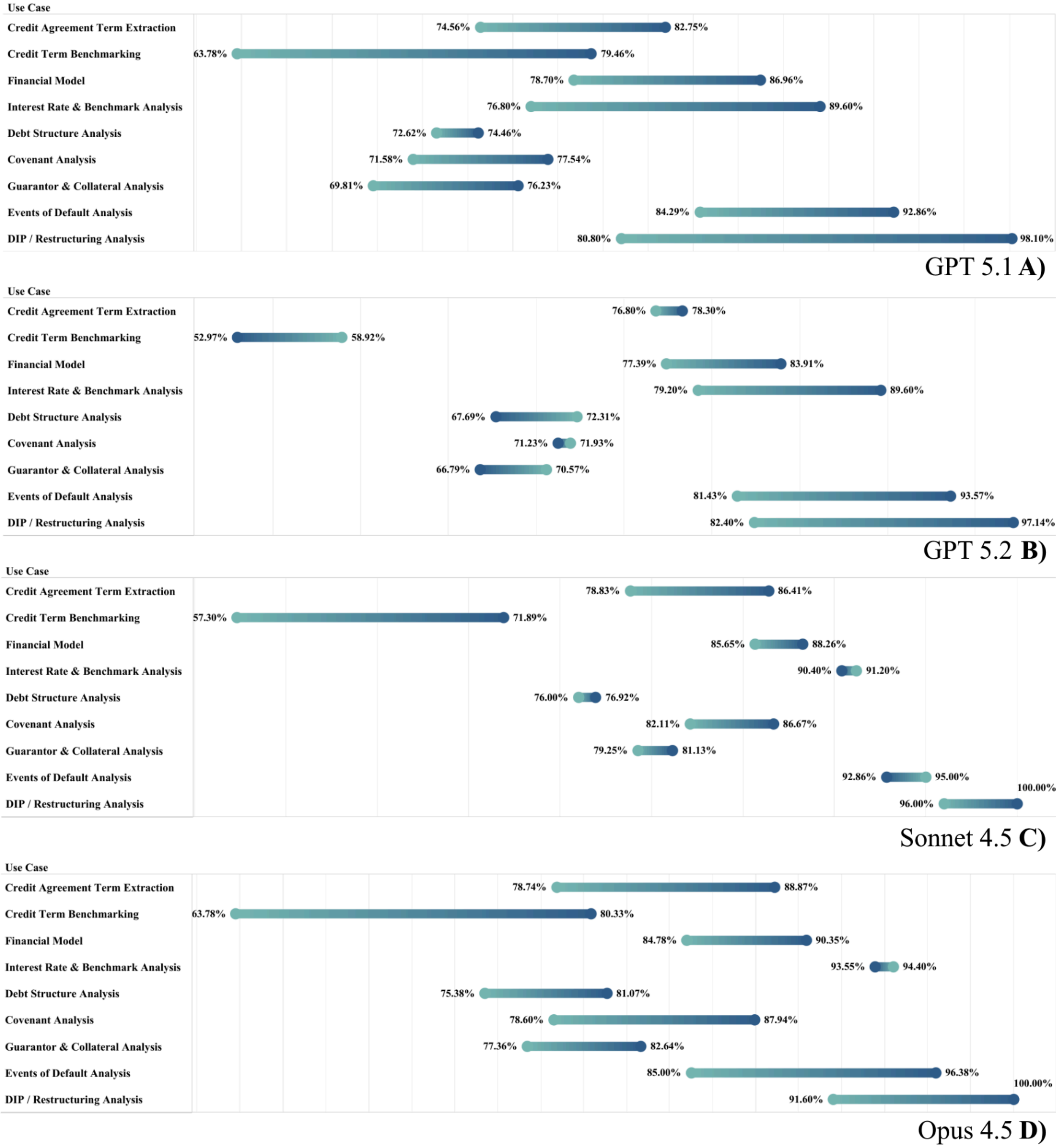


Figure 4. Accuracy (%) by debt-focused use case, comparing standalone and ToltIQ platform performance for **a)** GPT 5.1, **b)** GPT 5.2, **c)** Sonnet 4.5, and **d)** Opus 4.5.

A structural divide between the two use case categories is evident in the results. Debt-focused workflows produced higher peak accuracy and greater cross-model consistency than general private markets tasks. Three debt-focused use cases — DIP/Restructuring Analysis, Events of Default Analysis, and Interest Rate & Benchmark Analysis — exceeded 90% accuracy across multiple models, while only two general private markets use cases (CIM Analysis and VDR

Analysis) reached comparable levels. The most likely explanation is the nature of the underlying source material. Credit agreements follow relatively standardized drafting conventions, with key provisions appearing in predictable locations and using consistent terminology (Bellucci & McCluskey, 2017). This structural regularity benefits both the retrieval system, which can locate relevant passages more reliably, and the language model, which encounters less ambiguity in extraction and interpretation. General private markets documents such as CIMs, management presentations, and financial models are more variable in structure, formatting, and analytical framing, placing greater demand on both retrieval precision and inferential reasoning.

The most practically significant finding may be the persistence of low accuracy on synthesis-intensive tasks. Quality of Earnings Analysis, Credit Term Benchmarking, and Management Team Questionnaire remained below 81% across all models and all platform configurations, with standalone performance on these use cases clustering in the low-to-mid 60% range. These tasks share a common characteristic: they require the model to go beyond locating and extracting information and instead make inferential judgments: whether an EBITDA adjustment is reasonable, whether a provision is market-standard, whether a term warrants further diligence questioning. These are fundamentally different cognitive demands than extraction or summarization, and neither model capability nor RAG-enhanced retrieval appears sufficient to bridge the gap reliably. For PE firms designing AI-augmented workflows, these results suggest that synthesis-heavy tasks should retain substantive human oversight regardless of model or deployment choice, while extraction and document-navigation tasks represent the strongest candidates for automation.

5.3 Limitations and Future Research

The benchmark documents are exclusively credit agreements, a single document type within the broader universe of PE deal documentation (Vals AI, 2025). While credit agreements are central to leveraged finance workflows, due diligence also requires analysis of CIMs, management presentations, financial models, shareholder agreements, and regulatory filings; document types that differ substantially in structure, formatting, and analytical demands.

The evaluation was conducted in a single-document context, with each benchmark document analyzed in an isolated session. This represents the most favorable possible condition for standalone models, because a single document is the scenario in which context window limitations are least likely to constrain performance. In production PE environments, analysts routinely work across hundreds or thousands of documents in a virtual data room simultaneously (PwC, 2025; Bain & Company, 2026). Standalone LLMs cannot ingest document sets of this scale, while RAG architectures are specifically designed to handle multi-document retrieval. The accuracy gains reported in this study (6.63 to 9.63 pp) should therefore be understood as conservative estimates of the platform's real-world contribution. Testing under multi-document conditions is a priority for future research.

All models were tested using default inference parameters on both platforms, with no optimization of temperature, top-p, or other generation settings. It is possible that tuning these parameters for each model could improve standalone accuracy. However, parameter optimization requires significant technical expertise and iterative experimentation, time that most PE practitioners do not have. ToltIQ's architecture handles context construction and retrieval optimization at the platform level, which is part of the practical value proposition the benchmark was designed to evaluate.

LLM capabilities evolve rapidly, and the results reported here reflect model performance during the testing period of January-February 2026. Several models released during the preparation of this paper, including Claude Sonnet 4.6, Claude Opus 4.6, Gemini 3.1, and GPT 5.3, were not included in this evaluation. Incorporating them was not feasible without compromising the integrity of the testing framework. ToltIQ intends to evaluate these models in a follow-up study using the same methodology, alongside longitudinal tracking of model performance across version updates and expanded metadata analysis including response latency and in-text citation frequency.

6. Conclusion

This study evaluated four leading LLMs across 20 PE due diligence use cases under two deployment conditions, producing 14,400 scored responses and establishing the first use-case-level accuracy benchmark for AI-assisted due diligence workflows.

The three research objectives posed in Section 2.3 can be answered directly. First, ToltIQ's custom deployment architecture produced statistically significant accuracy gains for three of four models, ranging from 6.63 to 9.63 pp under Bonferroni-corrected thresholds. The consistency of these gains across architecturally distinct models indicates that deployment architecture is at least as consequential as model selection in determining accuracy outcomes. Second, Opus 4.5 achieved the highest aggregate accuracy at 85.11%, followed by Sonnet 4.5 at 83.54% and GPT 5.1 at 80.45%. GPT 5.2 was the sole underperformer, registering no significant gain from the ToltIQ platform. Third, performance varied substantially by use case category. Debt-focused workflows produced higher peak accuracy and greater cross-model consistency than general private markets tasks, with three debt-focused use cases exceeding 90% across multiple models. Synthesis-intensive tasks, Quality of Earnings Analysis, Credit Term Benchmarking, and Management Team Questionnaire, remained below 81% across all models and platforms.

These findings carry direct implications for PE firms designing AI-augmented workflows. For comprehensive due diligence requiring maximum accuracy, Opus 4.5 deployed through a RAG-enhanced platform offers the strongest performance profile. Sonnet 4.5 provides a favorable balance of accuracy and computational efficiency for high-volume processing.

Extraction and document-navigation tasks, particularly CIM Analysis and VDR Analysis, represent the strongest candidates for automation, where custom architecture delivered its largest marginal gains. Synthesis-heavy tasks requiring inferential judgment should retain substantive human oversight regardless of model or deployment choice, as neither model capability nor retrieval enhancement proved sufficient to achieve reliable accuracy on these workflows.

The pace of model development means that any benchmark is a snapshot. The data presented here reflects performance as of January 2026, and the competitive landscape will continue to shift. The value of this study lies not only in its current findings but in the methodology it establishes a reproducible, use-case-level evaluation framework that can be applied as new models emerge and as ToltIQ's custom architecture continues to evolve.

Appendix

Appendix A: Use Case Definitions

The 20 use cases evaluated in this study are organized into two categories: General Private Markets (11 use cases) and Debt-Focused (9 use cases). The descriptions below define the analytical scope of each use case as applied in the benchmark evaluation.

General Private Markets Use Cases

CIM Analysis - Analyze Confidential Information Memorandums for financial projections, market size claims, competitive positioning, EBITDA definitions and adjustments, and revenue metrics. Identify inconsistencies, gaps, or areas where data appears overly optimistic or lacks supporting evidence.

Virtual Data Room (VDR) Analysis - Extract financial documentation from Virtual Data Rooms, including monthly/quarterly revenue and EBITDA, working capital trends, CapEx requirements, debt service obligations, and specific EBITDA adjustments with supporting evidence.

Document Checklist - Identify conditions precedent, required consents, deliverables, and closing requirements. Track what documents, approvals, and actions are needed to satisfy deal or amendment conditions.

IC Memo Draft - Summarize company formation, ownership structure, key brands, top-line financial metrics, and strategic rationale for proposed acquisitions, including deal timeline and purchase price. Supports Investment Committee memorandum preparation with key deal terms and financial highlights.

Investment Criteria - Evaluate targets against investment criteria using financial metrics: revenue size thresholds, EBITDA margins, debt-to-EBITDA ratios, leverage limits, and growth rates. Identify specific numerical thresholds, caps, and permitted amounts that inform investment screening.

Quality of Earnings (QoE) Analysis - Analyze EBITDA definitions, including permitted add-backs (cost savings, synergies, pro forma adjustments), caps on adjustments, run-rate calculations, and how Consolidated Net Income flows into the EBITDA calculation.

Contract Extraction - Extract key contract terms, including contract duration, initial term length, start/end dates, renewal terms, extension conditions, governing law, and parties involved. Used for bulk document review to systematically pull standardized information across multiple agreements.

M&A / Change of Control Analysis - Analyze change of control definitions, triggers, and consequences, including put rights, mandatory prepayment, consent requirements, and permitted acquisition baskets. Review how M&A activity affects the credit agreement.

LOI Draft - Extract key deal terms needed for Letter of Intent drafting, including maturity dates, interest rates, principal amounts, prepayment terms, and other material transaction parameters that would be referenced in preliminary deal documentation.

Management Team Questionnaire - Generate questions for management regarding unusual provisions, operational flexibility, exceptions to standard terms, key employee arrangements, and areas requiring clarification. Identify topics that warrant deeper diligence discussion.

Portfolio Monitoring - Track portfolio company performance metrics over time, including EBITDA, revenue, leverage ratios, covenant compliance status, and default/breach indicators. Identify concerning trends, inflection points, and early warning signals across the portfolio.

Debt-Focused Use Cases

Credit Agreement Term Extraction - Identify and summarize types and amounts of credit facilities (term loans, revolving credits), their designated purposes (refinancing, general corporate use), key covenants, baskets, permitted actions, and lender protections. Core use case for analyzing loan documentation.

Credit Term Benchmarking - Compare credit agreement terms against market standards. Identify whether provisions are borrower-friendly, lender-friendly, or market-standard. Flag aggressive or unusual terms relative to comparable transactions.

Financial Model Analysis - Extract inputs for financial modeling, including interest rate mechanics (base rates, margins, floors), amortization schedules, mandatory prepayment percentages, fee structures, and cash flow sweep parameters.

Interest Rate & Benchmark Analysis - Analyze interest rate mechanics, including SOFR/LIBOR benchmarks, applicable margins, interest rate floors, benchmark transition provisions (hardwired vs. amendment approach), and spread adjustments.

Debt Structure Analysis - Analyze the capital structure, including lien priorities (first lien, second lien, unsecured), incremental facility capacity, debt incurrence baskets, and junior debt arrangements, and the overall debt stack hierarchy.

Covenant Analysis - Analyze financial maintenance covenants, including leverage ratios, interest coverage tests, fixed charge coverage, and other ratio-based requirements. Identify testing frequency, cure rights, equity cure provisions, and covenant holiday periods.

Guarantor & Collateral Analysis - Review guarantor requirements, collateral packages, security interests, pledge agreements, and release conditions. Analyze which entities provide guarantees, what assets secure the obligations, and conditions for collateral release.

Events of Default Analysis - Identify events of default triggers, including payment defaults, covenant breaches, cross-defaults, judgment thresholds, ERISA events, and change-of-control events. Analyze cure periods, materiality thresholds, and acceleration rights.

DIP/Restructuring Analysis - Analyze debtor-in-possession financing provisions, including priming liens, roll-up mechanics, adequate protection, milestones, and restructuring-specific terms. Review bankruptcy-related protections and carve-outs.

Appendix B: Vals AI Dataset and Confidentiality

The Vals AI CorpFin V2 benchmark used in this study is organized into three distinct context configurations, each of which uses the full set of 360 prompts but varies how much of each source document is provided to the model.

Exact Pages provides only the specific pages required to answer each question, typically resulting in a small context of only a few pages. This configuration tests the model's ability to extract and reason over a minimal, targeted passage.

Shared Max Context provides a subset of approximately 80 pages that is guaranteed to fit within the context window of all models evaluated and to contain the information needed for a correct answer. The subselection does not necessarily start at the first page, which can make document structure comprehension challenging for models that rely on positional cues.

Max Fitting Context provides the largest possible chunk of the document, starting from the first page, that fits within each model's context window. This means models with larger context windows receive more document content than those with shorter windows. This configuration was selected for the current study, as described in Section 3.1, because it most closely approximates production environments where models receive substantial document context and must locate relevant information within larger passages.

Per the licensing terms of the Vals AI CorpFin V2 private validation set, the specific documents and prompts included in the benchmark cannot be disclosed. The dataset is proprietary and not publicly available. For additional information about the CorpFin V2 benchmark methodology, document composition, and task design, please refer to the Vals AI website (Vals AI, 2025).

Appendix C: Prior Benchmarking Results

Prior to the controlled evaluation presented in this paper, ToltIQ published a preliminary benchmarking analysis using the same Vals AI CorpFin V2 benchmark and the same four models (Boeye & Williams, 2026). That analysis compared ToltIQ platform accuracy against standalone scores reported externally by Vals AI, rather than standalone tests conducted by the ToltIQ research team under matched conditions, yielding net improvements ranging from 7.6 to 22.4 pp across the four models.

The higher lift figures in the preliminary analysis reflect a difference in baseline methodology, not platform performance. ToltIQ's platform scores are identical across both analyses. The standalone baseline in the preliminary analysis used Vals AI's internally reported figures, which were generated under different testing conditions than the independently conducted standalone evaluation presented in Section 4. The current study was designed specifically to establish a more rigorous and reproducible baseline, and the lift figures reported here (6.63 to 9.63 pp) should be understood as the more methodologically sound estimate.

The results of the prior analysis are summarized below. ToltIQ platform scores are identical to those reported in the current study, as the same platform tests were used. Standalone scores differ because they reflect Vals AI's internally reported figures rather than the independently conducted standalone evaluation presented in Section 4.

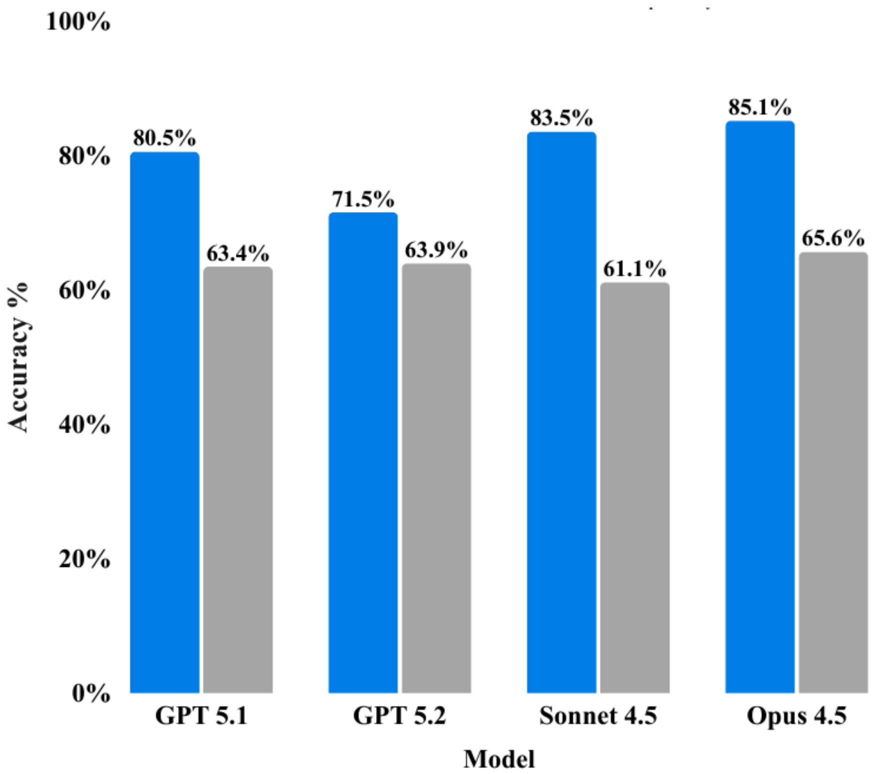


Figure 5. Mean accuracy (%) by model, comparing ToltIQ platform performance against standalone scores reported internally by Vals AI. ToltIQ platform scores are identical to those reported in the current study. ToltIQ scores are in blue and Standalone scores are in grey. Standalone scores reflect Vals AI's internally reported figures and differ from the independently conducted standalone evaluation presented in Section 4.

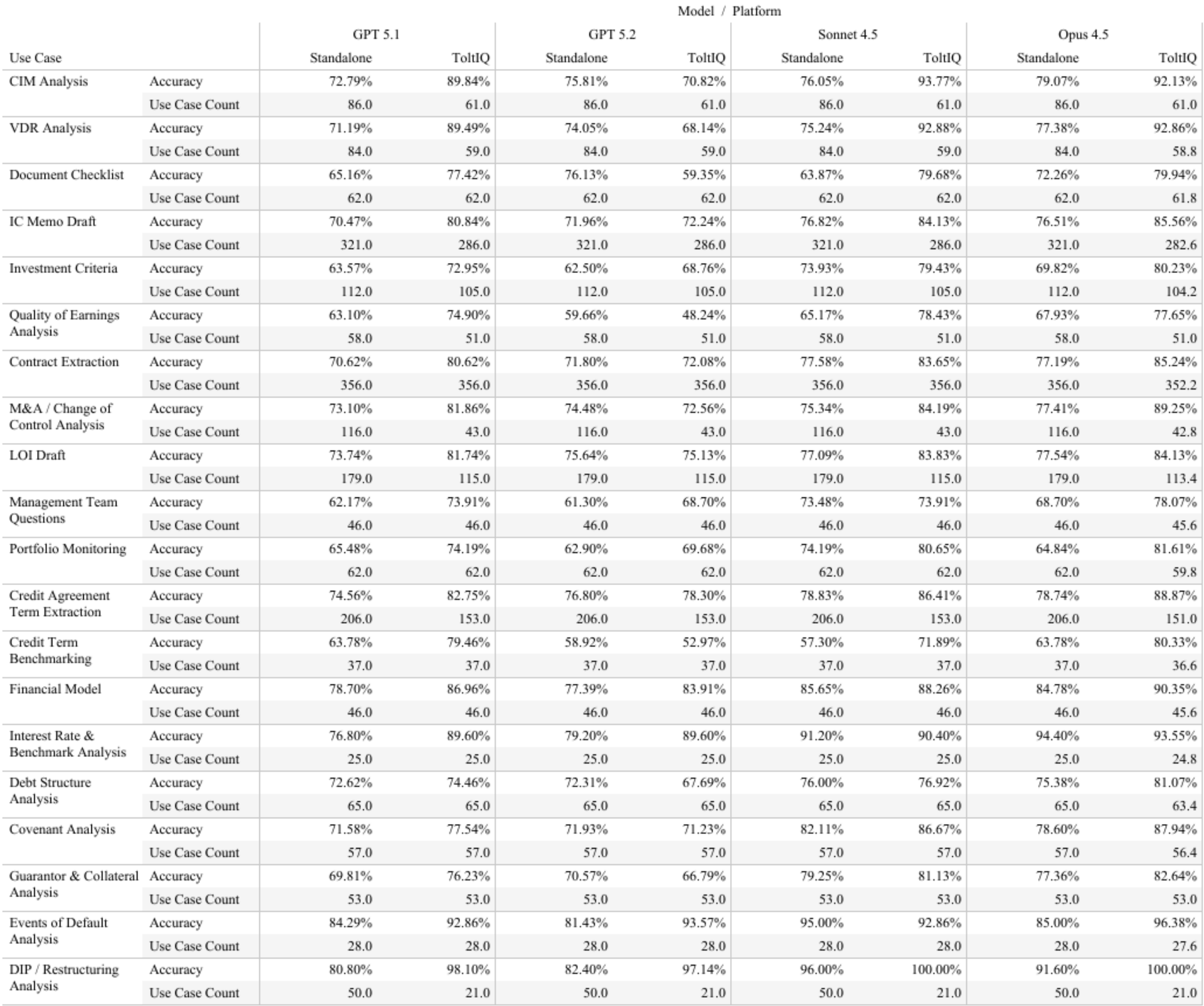
GPT 5.1 achieved 80.45% on the ToltIQ platform compared to a standalone average of 63.40%, with particular strength in structured extraction tasks. GPT 5.2 achieved 71.47% compared to 63.87% standalone, with notable weaknesses in complex analytical tasks such as Quality of Earnings Analysis and Credit Term Benchmarking. Sonnet 4.5 achieved 83.54% compared to 61.07% standalone, offering a balance of accuracy and computational efficiency for high-volume processing. Opus 4.5 achieved the highest accuracy at 85.11% compared to 65.60% standalone, with consistent performance across all use case categories.

These preliminary findings motivated the current study's more rigorous design, which introduced independently conducted standalone testing, five-replicate consistency checks, and statistical significance testing to isolate the contribution of deployment architecture from baseline model variability.

Appendix D: Supplementary Visualizations

The following figures provide additional perspectives on the performance data presented in Sections 4 and 5. These visualizations offer alternative views of model accuracy distributions and use case patterns that supplement the primary figures in the body of the paper.

Table 5. Complete accuracy matrix by use case, model, and platform. Accuracy (%) represents the mean proportion of correct responses across five replicates. Use Case Count represents the mean number of evaluated responses per replicate for each use case, reflecting the denominator of the accuracy calculation. Variations in use case count across models indicate instances where a model failed to generate a scorable response.



		Model / Platform							
		GPT 5.1		GPT 5.2		Sonnet 4.5		Opus 4.5	
Use Case		Standalone	ToltIQ	Standalone	ToltIQ	Standalone	ToltIQ	Standalone	ToltIQ
CIM Analysis	Accuracy	72.79%	89.84%	75.81%	70.82%	76.05%	93.77%	79.07%	92.13%
	Use Case Count	86.0	61.0	86.0	61.0	86.0	61.0	86.0	61.0
VDR Analysis	Accuracy	71.19%	89.49%	74.05%	68.14%	75.24%	92.88%	77.38%	92.86%
	Use Case Count	84.0	59.0	84.0	59.0	84.0	59.0	84.0	58.8
Document Checklist	Accuracy	65.16%	77.42%	76.13%	59.35%	63.87%	79.68%	72.26%	79.94%
	Use Case Count	62.0	62.0	62.0	62.0	62.0	62.0	62.0	61.8
IC Memo Draft	Accuracy	70.47%	80.84%	71.96%	72.24%	76.82%	84.13%	76.51%	85.56%
	Use Case Count	321.0	286.0	321.0	286.0	321.0	286.0	321.0	282.6
Investment Criteria	Accuracy	63.57%	72.95%	62.50%	68.76%	73.93%	79.43%	69.82%	80.23%
	Use Case Count	112.0	105.0	112.0	105.0	112.0	105.0	112.0	104.2
Quality of Earnings Analysis	Accuracy	63.10%	74.90%	59.66%	48.24%	65.17%	78.43%	67.93%	77.65%
	Use Case Count	58.0	51.0	58.0	51.0	58.0	51.0	58.0	51.0
Contract Extraction	Accuracy	70.62%	80.62%	71.80%	72.08%	77.58%	83.65%	77.19%	85.24%
	Use Case Count	356.0	356.0	356.0	356.0	356.0	356.0	356.0	352.2
M&A / Change of Control Analysis	Accuracy	73.10%	81.86%	74.48%	72.56%	75.34%	84.19%	77.41%	89.25%
	Use Case Count	116.0	43.0	116.0	43.0	116.0	43.0	116.0	42.8
LOI Draft	Accuracy	73.74%	81.74%	75.64%	75.13%	77.09%	83.83%	77.54%	84.13%
	Use Case Count	179.0	115.0	179.0	115.0	179.0	115.0	179.0	113.4
Management Team Questions	Accuracy	62.17%	73.91%	61.30%	68.70%	73.48%	73.91%	68.70%	78.07%
	Use Case Count	46.0	46.0	46.0	46.0	46.0	46.0	46.0	45.6
Portfolio Monitoring	Accuracy	65.48%	74.19%	62.90%	69.68%	74.19%	80.65%	64.84%	81.61%
	Use Case Count	62.0	62.0	62.0	62.0	62.0	62.0	62.0	59.8
Credit Agreement Term Extraction	Accuracy	74.56%	82.75%	76.80%	78.30%	78.83%	86.41%	78.74%	88.87%
	Use Case Count	206.0	153.0	206.0	153.0	206.0	153.0	206.0	151.0
Credit Term Benchmarking	Accuracy	63.78%	79.46%	58.92%	52.97%	57.30%	71.89%	63.78%	80.33%
	Use Case Count	37.0	37.0	37.0	37.0	37.0	37.0	37.0	36.6
Financial Model	Accuracy	78.70%	86.96%	77.39%	83.91%	85.65%	88.26%	84.78%	90.35%
	Use Case Count	46.0	46.0	46.0	46.0	46.0	46.0	46.0	45.6
Interest Rate & Benchmark Analysis	Accuracy	76.80%	89.60%	79.20%	89.60%	91.20%	90.40%	94.40%	93.55%
	Use Case Count	25.0	25.0	25.0	25.0	25.0	25.0	25.0	24.8
Debt Structure Analysis	Accuracy	72.62%	74.46%	72.31%	67.69%	76.00%	76.92%	75.38%	81.07%
	Use Case Count	65.0	65.0	65.0	65.0	65.0	65.0	65.0	63.4
Covenant Analysis	Accuracy	71.58%	77.54%	71.93%	71.23%	82.11%	86.67%	78.60%	87.94%
	Use Case Count	57.0	57.0	57.0	57.0	57.0	57.0	57.0	56.4
Guarantor & Collateral Analysis	Accuracy	69.81%	76.23%	70.57%	66.79%	79.25%	81.13%	77.36%	82.64%
	Use Case Count	53.0	53.0	53.0	53.0	53.0	53.0	53.0	53.0
Events of Default Analysis	Accuracy	84.29%	92.86%	81.43%	93.57%	95.00%	92.86%	85.00%	96.38%
	Use Case Count	28.0	28.0	28.0	28.0	28.0	28.0	28.0	27.6
DIP / Restructuring Analysis	Accuracy	80.80%	98.10%	82.40%	97.14%	96.00%	100.00%	91.60%	100.00%
	Use Case Count	50.0	21.0	50.0	21.0	50.0	21.0	50.0	21.0

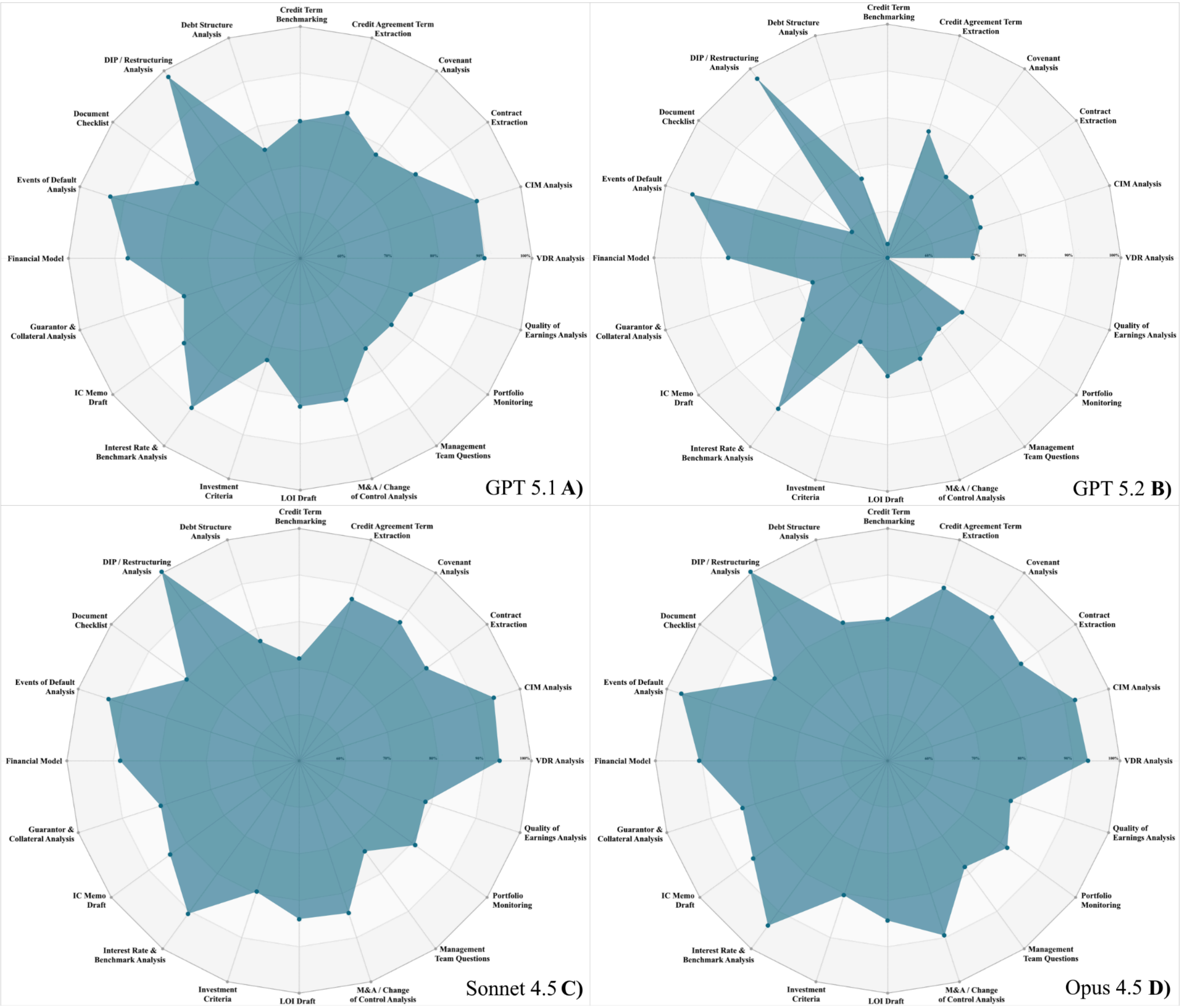


Figure 6. Radar charts showing ToltIQ platform accuracy (%) by use case for each model: **a)** GPT 5.1, **b)** GPT 5.2, **c)** Sonnet 4.5, and **d)** Opus 4.5. All 20 use cases are represented across both general private markets and debt-focused categories. Radial axis begins at 50%; visual differences between use cases are therefore magnified relative to absolute scale.

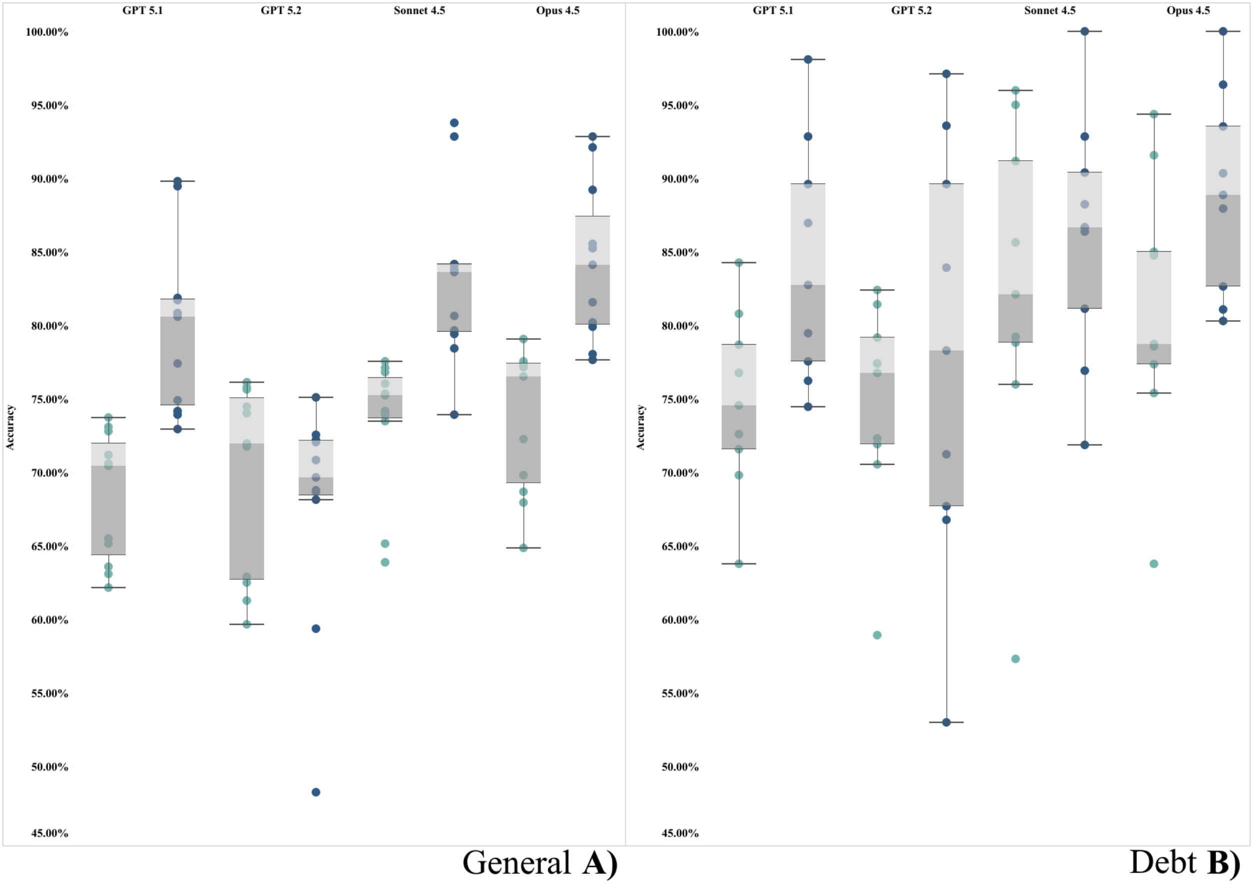


Figure 7. Distribution of use case accuracy scores by model and platform, separated by category: **a)** general private markets use cases (left) and **b)** debt-focused use cases (right). Each box represents the interquartile range of accuracy scores across use cases within that category. Blue points represent ToltIQ platform performance; grey points represent standalone API performance. This figure supplements Figure 2 in the body of the paper, which presents the combined distribution across both categories.

Appendix E: Acknowledgments and Disclosure

Acknowledgments

The authors would like to thank Ruban Selvakumar (Chief Client Officer) and the ToltIQ client team for their ongoing feedback and domain expertise, which helped shape the use case prioritization and real-world framing throughout this study. We are also grateful to David Schraeder, Head of Communications, for his support in the preparation and publication of this white paper.

We extend our appreciation to the team at Vals AI for developing the CorpFin V2 benchmark and for their responsiveness throughout the evaluation process. The integrity of this study depends significantly on the rigor of their evaluation infrastructure.

Finally, we thank the broader ToltIQ engineering, product, and leadership teams, whose work made this research possible.

Disclosure

This research was conducted by ToltIQ and evaluates, in part, the performance of ToltIQ's own platform. The authors are employees of ToltIQ. No external funding was received for this study.

To mitigate potential bias arising from this relationship, several design choices were made. The benchmark dataset was developed by Vals AI with input from independent domain experts. All responses — both standalone and ToltIQ-deployed — were scored using Vals AI's standardized evaluation framework rather than an internally developed evaluator. Each prompt was replicated five times to test for consistency, and statistical significance was assessed using a Bonferroni-corrected threshold. The standalone baseline was tested via direct API access through the Vals AI platform, not through ToltIQ's infrastructure.

These measures do not eliminate the inherent conflict of interest in self-evaluation, and readers should interpret the results with this context in mind. The authors welcome independent replication of this benchmark by third parties using the Vals AI CorpFin V2 dataset.

References

Anthropic. (2025a, September 29). Introducing Claude Sonnet 4.5. Anthropic. https://www.anthropic.com/news/claude-sonnet-4-5

Anthropic. (2025b, November 24). Introducing Claude Opus 4.5. Anthropic. https://www.anthropic.com/news/claude-opus-4-5

Bain & Company. (2026). Global private equity report 2026. Bain & Company. https://www.bain.com/insights/topics/global-private-equity-report/

Bellucci, M., & McCluskey, J. (2017). *The LSTA's complete credit agreement guide* (2nd ed.). McGraw-Hill.

Boeye, M., & Williams, S. (2026, February 2). Independent LLM benchmark validates ToltIQ performance advantages over ChatGPT and Claude for private equity due diligence. ToltIQ. https://www.toltiq.com/insights/toltiq-advantages-over-chatgpt-and-claude-for-private-equity-due-diligence

Deloitte. (2025). 2025 M&A generative AI study. Deloitte US. https://www.deloitte.com/us/en/what-we-do/capabilities/mergers-acquisitions-restructuring/articles/m-and-a-generative-ai-study.html

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., & Wang, H. (2024). Retrieval-augmented generation for large language models: A survey. arXiv. https://arxiv.org/abs/2312.10997

Kong, Y., Lee, H., Hwang, Y., Lopez-Lira, A., Levy, B., Mehta, D., Wen, Q., Choi, C., Lee, Y., & Zohren, S. (2025). Evaluating LLMs in finance requires explicit bias consideration. *arXiv*. https://arxiv.org/abs/2602.14233

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems, 33*, 9459–9474.

Mohsin, M. T. (2025). Evaluating large language models (LLMs) in financial NLP: A comparative study on financial report analysis. *arXiv*. https://arxiv.org/abs/2507.22936

Nie, Y., Kong, Y., Dong, X., Mulvey, J. M., Poor, H. V., Wen, Q., & Zohren, S. (2024). A survey of large language models for financial applications: Progress, prospects and challenges. arXiv. https://arxiv.org/abs/2406.11903

OpenAI. (2025a, August 13). GPT-5 system card. OpenAI. https://openai.com/index/gpt-5-system-card/

OpenAI. (2025b, December 11). Introducing GPT-5.2. OpenAI. https://openai.com/index/introducing-gpt-5-2/

OpenAI. (2025c, December 11). Update to GPT-5 system card: GPT-5.2. OpenAI. https://cdn.openai.com/pdf/3a4153c8-c748-4b71-8e31-aecbde944f8d/oai_5_2_system-card.pdf

PwC. (2025). Deals in the age of AI: How dealmakers turn human capability into AI-driven returns. PwC US. https://www.pwc.com/us/en/services/consulting/deals/library/ai-private-equity-corporate-deals-people-challenges.html

Sharma, C. (2025). Retrieval-augmented generation: A comprehensive survey of architectures, enhancements, and robustness frontiers. arXiv. https://arxiv.org/abs/2506.00054

Vals AI. (2025). CorpFin (v2) benchmark. Vals AI. https://www.vals.ai/benchmarks/corp_fin_v2