INS-ActBench: A Comprehensive Benchmark for Assessing Professional Actuarial Capability of Large Language Models

Anonymous Authors

Abstract

Large Language Models (LLMs) have shown strong potential in financial reasoning, but existing benchmarks often evaluate domain knowledge, numerical reasoning, long-context understanding, and tool use in separate settings. This limits their ability to assess realistic professional workflows that require auditable, context-grounded, and tool-executable decisions. We introduce **INS-ActBench**, a comprehensive benchmark for evaluating professional actuarial capability in LLMs. INS-ActBench contains 12,050 Q&A pairs from public exams and sample questions released by 16 actuarial associations. It covers three subsets: **INS-Act-Know** for standardized actuarial knowledge, **INS-Act-Case** for long-context insurance case reasoning, and **INS-Act-Practice** for spreadsheet and R-code tasks with verifiable numerical outputs. Experiments on nine representative LLMs and human actuarial experts reveal a clear capability boundary: frontier LLMs perform strongly on standardized knowledge, but remain much weaker in case reasoning, tool-based workflows, and jurisdiction-sensitive practice. INS-ActBench provides a reproducible foundation for developing actuarial LLMs toward reliable professional assistance. The code and data are available at https://anonymous.4open.science/r/ActuarialBench-425C.

1 Introduction

Large Language Models (LLMs) have achieved rapid progress in natural language processing and are increasingly applied to specialized domains that require expert knowledge and complex reasoning (Xie et al., 2024). Finance is one such high-stakes domain. Financial tasks often require models to interpret domain-specific documents, perform numerical reasoning, assess risks, support forecasting, and assist decision-making under uncertainty (Xie et al., 2023; Zhao et al., 2022; Li et al., 2025; Gan et al., 2025; Tang et al., 2025). As a result, finance has become an important testbed for evaluating whether LLMs can move beyond general language understanding toward professional analytical capability.

Despite this progress, existing financial benchmarks still provide limited coverage of practice-oriented numerical reasoning in realistic professional workflows. Prior benchmarks typically evaluate financial text understanding (Xie et al., 2023), question answering (Xie et al., 2024), investment decision-making (Li et al., 2025), table-text reasoning (Zhu et al., 2021), long-context document understanding (Reddy et al., 2024), spreadsheet reasoning (Ravnik et al., 2026), or code generation (Song et al., 2026) as separate capabilities. However, real financial practice often requires these abilities to be integrated into auditable workflows that combine domain knowledge, long-context evidence, regulatory constraints, numerical calculation, and tool-based implementation. This gap is critical in high-stakes settings, where verifiable and context-grounded decisions matter more than plausible standalone answers.

Actuarial science provides a rigorous domain for filling this gap. As a specialized subfield of finance, actuarial science applies mathematics, statistics, and financial theory to analyze the economic cost of risk and uncertainty, and provides professional support for pricing, reserving, capital management, and long-term financial forecasting in insurance and pensions (Miljkovic, 2017). Compared with general financial tasks, actuarial work is more calculation-intensive, more tightly connected to regulatory and business contexts, and more dependent on professional tools such as spreadsheets and statistical programming (Acharya et al., 2009; Espinosa and Zarruk, 2021; Campbell, 2010; Dutang et al., 2008). Therefore, actuarial work is especially suitable for evaluating professional calculation and application in complex practical scenarios.

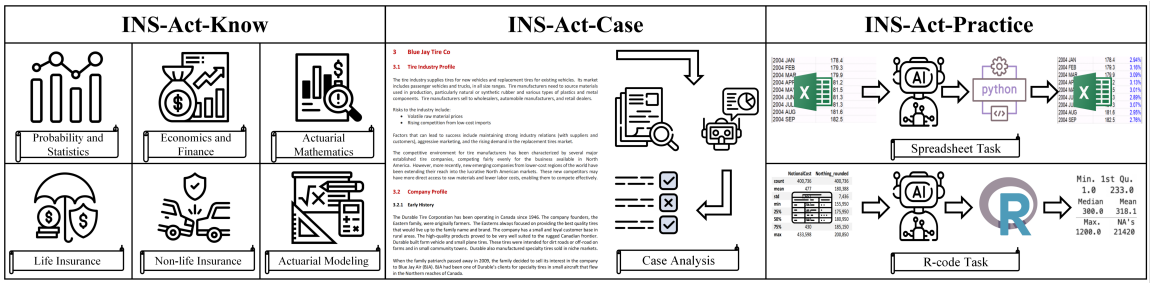


Figure 1: Evaluation dataset for INS-ActBench.

Existing research and benchmarks do not yet adequately evaluate this integrated actuarial capability. Current actuarial LLM research contains only a single-case analysis (Balona, 2024), and existing insurance benchmarks mainly cover general insurance knowledge, such as textbook test (Ding et al., 2025; Zhou et al., 2025), policy retrieval (Chen et al., 2025), or visual liability assessment (Lin et al., 2025), which do not evaluate numerical reasoning under complex contextual settings. A comprehensive actuarial benchmark is therefore needed. As a result, it remains unclear whether current LLMs can support actuarial tasks that resemble real professional practice rather than isolated financial knowledge tests.

To address this gap, we introduce **INS-ActBench**, a comprehensive benchmark for evaluating LLMs in actuarial science. INS-ActBench is constructed from publicly available past exams released by 16 actuarial associations, and contains 12,050 Q&A pairs. It is organized into three subsets corresponding to progressive actuarial capabilities: INS-Act-Know evaluates standardized actuarial knowledge across six subcategories; INS-Act-Case evaluates long-context actuarial reasoning in realistic insurance business cases; and INS-Act-Practice evaluates practice-oriented tool use through spreadsheet and R-code tasks. We illustrate INS-ActBench in Figure 1. We further annotate source jurisdictions and numerical question types to support fine-grained analysis of actuarial knowledge, regulatory context, and practical workflow competence.

Experiments on nine representative LLMs and human actuarial experts show a clear capability boundary. Frontier LLMs perform strongly on standardized actuarial knowledge and can exceed human experts on this subset, but their performance drops substantially on long-context case analysis and practice-oriented tool use. In contrast, human experts remain more balanced across dimensions. These results indicate that current LLMs can already conduct preliminary actuarial calculation and knowledge checking, while expert-level actuarial assistance still requires stronger long-context reasoning, reliable tool execution, and cross-jurisdictional adaptation.

Our contributions are summarized below:

- We introduce **INS-ActBench**, the first large-scale actuarial benchmark spanning knowledge, case reasoning, and tool-based practice.
- We curate 12,050 Q&A pairs from 16 actuarial associations, with coverage of long-context cases, spreadsheet/R-code tasks, jurisdictions, and numerical question types.
- We provide a systematic comparison between LLMs and human experts, showing that LLMs have made strong progress in actuarial knowledge but face clear limitations in case reasoning, tool-based workflows, and jurisdiction-sensitive practice.

2 Related Work

Existing financial benchmarks have developed along several relatively separate lines. General financial benchmarks mainly evaluate broad finance and decision-oriented tasks: PIXIU builds a financial LLM benchmark and instruction dataset for general finance (Xie et al., 2023); FinBen provides a holistic financial benchmark covering multiple types of tasks (Xie et al., 2024); CFinBench evaluates Chinese financial knowledge (Nie et al., 2025); InvestorBench focuses on LLM-based agents for investment decision-making (Li et al., 2025); and MME-Finance extends financial evaluation to multimodal reasoning (Gan et al., 2025). These benchmarks are mainly designed for general financial evaluation, with limited focus on a specific professional task type or a specialized financial sub-discipline. Specifically, Insurance-oriented benchmarks further move evaluation into the insurance domain: InsQABench evaluates Chinese

Benchmark	Task Type	Insurance	Numerical	Long Ctx.	Spreadsheet	Code
FinBen (Xie et al., 2024)	QA and reasoning	×	✓	×	×	×
CFinBench (Nie et al., 2025)	Chinese evaluation	×	✓	×	×	×
InvestorBench (Li et al., 2025)	Investment	×	✓	×	×	×
MME-Finance (Gan et al., 2025)	Multimodal	×	✓	×	×	×
InsQABench (Ding et al., 2025)	Insurance Q&A	✓	×	×	×	×
INSEva (Chen et al., 2025)	Insurance evaluation	✓	×	×	×	×
INS-MMBench (Lin et al., 2025)	Insurance multimodal	✓	×	×	×	×
TAT-QA (Zhu et al., 2021)	Table-text reasoning	×	✓	×	×	×
FinMathBench (He et al., 2026)	Math reasoning	×	✓	×	×	×
FinanceBench (Islam et al., 2023)	Document Q&A	×	×	✓	×	×
DocFinQA (Reddy et al., 2024)	Long-ctx. reasoning	×	✓	✓	×	×
FinSheet-Bench (Ravnik et al., 2026)	Spreadsheets	×	✓	×	✓	×
Finch (Dong et al., 2025)	Workflows	×	✓	×	✓	✓
StatLLM (Song et al., 2026)	Statistical coding	×	✓	×	×	✓
INS-ActBench (ours)	**Actuarial tasks**	✓	✓	✓	✓	✓

Table 1: Comparison with existing financial benchmarks based on insurance area, numerical calculation, long-context, spreadsheet, coding. INS-ActBench includes all of them.

insurance question answering over commonsense knowledge, structured databases, and unstructured documents (Ding et al., 2025), while INSEva provides a broader Chinese insurance benchmark covering business areas, task formats, difficulty levels, and cognitive-knowledge dimensions (Chen et al., 2025). However, existing insurance benchmarks have not yet focused on numerical reasoning tasks. For numerical reasoning in finance, TAT-QA and FinMathBench require reasoning over financial tables and associated texts (Zhu et al., 2021; He et al., 2026). Then, for long-context financial understanding, FinanceBench and DocFinQA evaluates open-book question answering over full long-document contexts (Islam et al., 2023; Reddy et al., 2024). For tool-oriented settings in finance, FinSheet-Bench and Finch evaluate financial spreadsheet question answering and numerical reasoning (Ravnik et al., 2026; Dong et al., 2025). StatLLM evaluates statistical analysis code generation and execution quality (Song et al., 2026). They are illustrated in Table 1.

However, this separation limits the ability of existing benchmarks to evaluate actuarial work, where these dimensions are practically coupled. Actuarial work requires an integrated evaluation setting because its core tasks jointly involve numerical calculation, long-context document analysis, regulatory interpretation, and tool-based implementation (Dutang et al., 2008). Besides, insurance is more regulation-sensitive than generic financial tasks, since solvency supervision and jurisdiction-specific rules directly shape actuarial judgment (Acharya et al., 2009). Insurance companies also operate as complex risk-bearing systems, where pricing and reserving interact under long-term uncertainty (Owadally et al., 2018; Espinosa and Zarruk, 2021). Therefore, actuarial evaluation requires LLMs to reason over long business contexts and integrate scattered case evidence before producing numerical decisions. In addition, actuarial numerical work is commonly implemented through spreadsheets and statistical programming: spreadsheets remain a major workhorse in actuarial practice (Campbell, 2010), while R has become a standard environment for computational actuarial modeling (Miljkovic, 2017). These requirements make it necessary to evaluate insurance-grounded numerical reasoning, long-context case analysis, spreadsheet operation, and code-based computation in a unified setting. INS-ActBench is designed for this purpose, providing a benchmark that reflects the professional workflow of actuarial work rather than isolated financial knowledge.

3 INS-ActBench

3.1 Data Sources

All Q&A pairs in INS-ActBench are collected from publicly available past exams or sample questions released by 16 actuarial associations. We use the Society of Actuaries (SOA) and the Institute and Faculty of Actuaries (IFoA) as the two major global actuarial sources, as they are among the actuarial as-

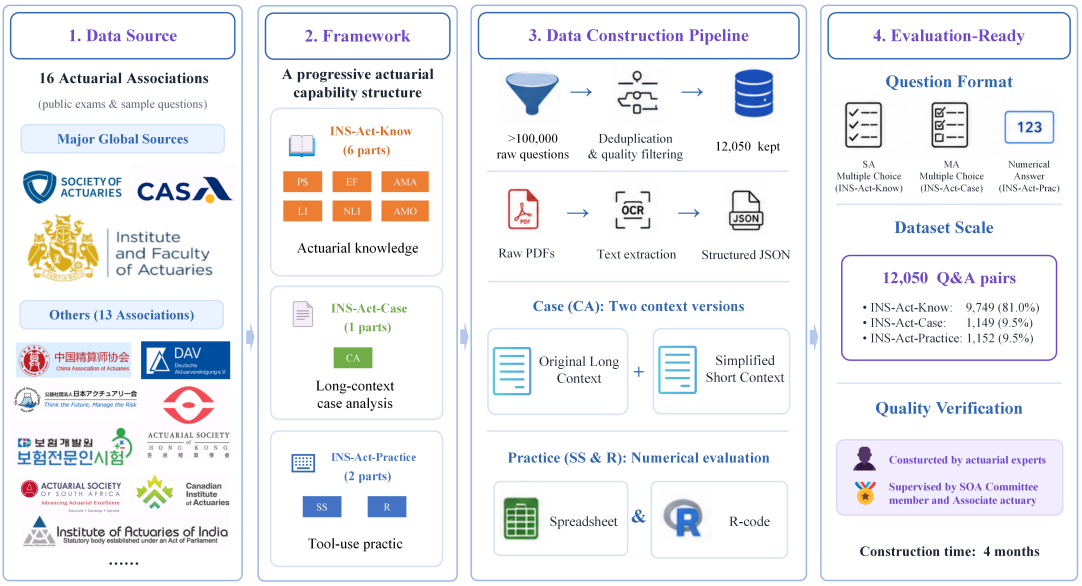


Figure 2: Overview of INS-ActBench pipeline.

sociations with the largest numbers of reported credentialed members worldwide; details are provided in Appendix A. The Casualty Actuarial Society (CAS) is grouped with SOA because both belong to the North American actuarial system, and we refer to this combined source category as 'SOA' in the rest of the paper. The remaining Q&A pairs of the other 13 actuarial associations are grouped as 'Others'. Their full names are listed in Appendix A. Since actuarial work is closely tied to insurance regulation, these source categories also correspond to different regulatory environments rather than only exam providers. This source design enables INS-ActBench to evaluate actuarial capability across major global actuarial systems and broader local regulatory contexts.

3.2 Framework

Based on our review of publicly available exams and sample questions from 16 actuarial associations, we observe that actuarial assessment follows a progressive capability structure. The first level focuses on standardized actuarial knowledge, where candidates are expected to understand actuarial concepts, principles, formulas, and routine computational methods. The second level focuses on integrated case-based reasoning, where candidates must combine actuarial knowledge with business background, financial information, and contextual evidence to solve complex insurance problems. The third level focuses on practice-oriented professional execution, where candidates need to translate actuarial assumptions and task requirements into verifiable computational workflows. Following this observed hierarchy, INS-ActBench is organized into three progressively structured subsets: INS-Act-Know, INS-Act-Case, and INS-Act-Practice.

INS-Act-Know evaluates standardized actuarial knowledge, including actuarial concepts, principles, formulas, computation, and modeling. It contains six subcategories: Probability and Statistics (PS), Economics and Finance (EF), Actuarial Mathematics (AMA), Life Insurance (LI), Non-life Insurance (NLI), and Actuarial Modeling (AMO).

INS-Act-Case evaluates long-context actuarial reasoning in realistic insurance business scenarios. It uses company background, financial statements, operating information, and related case materials to test whether models can locate relevant evidence, integrate scattered information, and make actuarial judgments under complex business conditions. It contains one subcategory: Case Analysis (CA).

INS-Act-Practice evaluates practice-oriented actuarial tool use. It requires LLMs to translate actuarial task requirements into executable spreadsheet formulas or R code, and to produce verifiable numerical results. It contains two subcategories: Spreadsheet (SS) and R-code (R).

3.3 Data Construction Pipeline

3.3.1 Question Selection

We initially collected more than 100,000 raw questions from publicly available actuarial exams and sample questions released by 16 actuarial associations. Given the high similarity among some historical exam items and balancing the scale of the

Annotation	INS-Act-Know							INS-Act-Case	INS-Act-Practice			Total
	PS	EF	AMA	LI	NLI	AMO	Total	CA	SS	R	Total	
SOA	658	493	61	163	101	510	1986	560	146	36	182	2728
IFoA	214	438	146	205	265	246	1514	316	83	216	299	2129
Others	1292	1094	927	1481	802	653	6249	273	407	264	671	7193
Num	2108	1309	1058	1735	1018	1317	8545	–	–	–	–	–
Non-num	56	716	76	114	150	92	1204	–	–	–	–	–
Total	**2164**	**2025**	**1134**	**1849**	**1168**	**1409**	**9749**	**1149**	**636**	**516**	**1152**	**12050**

Table 2: Basic statistics of INS-ActBench.

benchmark, we applied a selective filtering process. First, we prioritized questions from the most recent five years for each association to improve freshness. Second, we focused on numerical and calculation-oriented actuarial questions, while retaining only a small number of representative non-numerical questions that test core actuarial concepts. In addition, we observed that actuarial knowledge, case analysis, and tool-use tasks appear in credentialing exams at an approximate ratio of 8:1:1, so we followed this structure when determining the relative sizes of the subsets. This process resulted in 12,050 selected questions. The original materials were primarily in PDF format, which we processed through text-layer extraction or OCR and organized into a standardized JSON format.

3.3.2 Task Standardization

To support standardized and reproducible evaluation, we further adapted the selected questions into task formats aligned with the three subsets.

For INS-Act-Know, we standardize questions into single-answer multiple-choice format. Subjective questions are adapted into independent objective items while preserving the information required for deriving the answer. For questions that already meet our format requirements, we still revise the wording and shuffle the options while preserving the original meaning. We further annotate each question as numerical(Num) or non-numerical(Non-num) to distinguish calculation actuarial knowledge from concept knowledge.

For INS-Act-Case, we convert subjective case questions into multiple-answer multiple-choice questions. This partially reduces the openness of the original answers, but makes the tasks more suitable for benchmark construction and automated evaluation. We preserve the original case materials as long-context inputs with a maximum length of 175K tokens and an average length of 58K tokens. Each case question is paired with an original long context and a simplified short context (reducing the context length to within 32K tokens) to support LLMs with different context lengths. The token distribution is shown in Figure 5 in Appendix B.

For INS-Act-Practice, we construct spreadsheet and R-code tasks. To ensure evaluation accuracy, we adapt all questions to require numerical answers. Each spreadsheet task contains an input spreadsheet for the LLM and a ground-truth spreadsheet as reference answer, which differ only in the answer cells for evaluation. We convert every cell in the input spreadsheet into textual input to avoid losing actuarial information. We require the LLM to write Python code that fills the predicted answers into predefined answer cells and regenerates the spreadsheet. Each R-code task require numerical outputs such as statistics, correlation coefficients, or model-selection criteria, with the original data provided in textual form. We require LLMs to generate R code that computes the corresponding numerical results.

This task standardization process adapts all selected questions into benchmark formats and reduces the risk of data contamination. Representative examples for each subset are provided in Appendix B. And Table 2 provides a basic statistics overview of INS-ActBench.

3.4 Quality Verification

INS-ActBench was constructed by a researcher with interdisciplinary training in actuarial science and computer science, and then verified by two actuarial graduate students. The verification process details are shown in Appendix B. The entire process was supervised and calibrated by a member of the SOA China Committee and an Associate of SOA. In total, 9.65% of the questions were revised during quality control, and the full construction

Model		INS-Act-Know							INS-Act-Case	INS-Act-Practice			Total
		PS	EF	AMA	LI	NLI	AMO	Total	CA	SS	R	Total	
Proprietary LLMs	**Gemini-3.1-Pro**	**97.5**	**95.4**	94.8	**90.6**	**94.0**	**94.0**	**94.5**	55.5	67.0	69.0	67.9	**88.3**
	GPT-5.5	97.0	**95.4**	93.1	87.2	93.2	92.5	93.3	56.3	67.7	71.1	69.2	87.4
	DeepSeek-V4-Pro	96.5	93.1	92.2	83.3	90.2	91.0	91.3	43.9	55.2	69.6	61.7	83.9
	Qwen3.6-Plus	95.1	90.7	89.9	73.5	89.2	88.7	87.8	44.6	47.4	69.8	57.4	80.8
	Claude-Opus-4.7	84.7	85.5	76.6	72.9	83.5	81.8	81.1	43.3	60.6	71.3	65.4	76.0
	Kimi-K2.6	78.7	75.8	64.7	57.2	76.2	70.5	70.9	24.8	29.5	69.0	47.2	64.3
Open-source LLMs	**Qwen3-14B**	42.2	44.3	38.2	34.6	40.9	41.0	40.4	21.0	2.7	54.8	26.0	37.2
	Qwen3.5-35B-A3B	43.4	43.5	33.8	29.9	37.9	36.8	38.1	20.2	16.3	64.0	37.7	36.4
	Gemma-3-12B-IT	33.6	42.0	32.2	27.3	36.4	36.1	34.7	16.7	1.0	44.4	20.4	31.6
Human Experts Baseline		83.3	91.7	**100.0**	80.6	80.6	91.7	88.0	**76.7**	**76.4**	**90.0**	**83.2**	85.6

Table 3: Evaluation results of different LLMs and human experts on each subset of INS-ActBench. The values in the table represent the average accuracy. The best performance is in **bold**.

process took four months.

4 Experiments

4.1 Experimental Setup

LLMs. We have evaluated 6 proprietary LLMs:GPT-5.5, Claude-Opus-4.7, Gemini-3.1-Pro, DeepSeek-V4-Pro, Kimi-K2.6, and Qwen3.6-Plus; and 3 open-source LLMs: Qwen3.5-35B-A3B, Qwen3-14B, and Gemma-3-12B-IT.

Parameters. INS-Act-Know and INS-Act-Case use a two-shot setting, while INS-Act-Practice use a zero-shot setting. Proprietary models are run with thinking mode enabled and set to high when supported. Open-source models are are evaluated using the vLLM framework (Kwon et al., 2023). Due to context-length limits, Qwen3-14B and Gemma-3-12B-IT use simplified case contexts, while the other models use original long-context cases in INS-Act-Case. The system prompts are provided in Figure 4 in Appendix B.

Human Baseline. We also conduct a human expert evaluation on a sampled subset of INS-ActBench. Three actuarial participants completed the test independently: one expert holding a PhD degree in actuarial science, one expert holding a master's degree in actuarial science, and one outstanding undergraduate student majoring in actuarial science. None of them participated in dataset construction. We sampled 102 questions: 72 from INS-Act-Know, covering four questions from each jurisdiction in each of the six subcategories; 10 from INS-Act-Case; 10 spreadsheet and 10 R-code tasks from INS-Act-Practice. Participants were allowed to consult textbooks but had no Internet access. The human evaluation subset corresponds to approximately 10–12 hours of examination time in real actuarial exams. To preserve the time pressure of professional testing, we required participants to complete the test within 10 hours, which also matched their actual completion time. Given the difficulty of the tasks and the long completion time, this subset provides a sufficient estimate of human performance on INS-ActBench.

Evaluation. (1) For single-answer multiple-choice questions, we report accuracy. (2) For multiple-answer multiple-choice questions, an answer is counted as correct only when all selected options exactly match the reference answer. (3) For Spreadsheet tasks, we execute the generated Python code to produce an output file, and then compare the numerical values in the predefined answer cells with the reference answers. If the response is a formula in answer cell, we execute it to obtain the numerical result. If multiple answer cells are required, we take the average accuracy of each cell. A relative error tolerance of 0.2% is allowed, which is acceptable in actuarial science. (4) For R-code tasks, we execute the generated R code in Docker and compare the extracted numerical outputs with the reference answers using the same 0.2% tolerance.

To further analyze the failure reasons in INS-Act-Practice, we conduct an error analysis for all incorrect Spreadsheet and R-code responses, and classify each error into three categories: (1) **Tool call failure** refers to cases where the model fails to transform the task into an executable tool operation, causing the workflow to remain unstarted, the output to be missing. This reflects a pure tool-use problem. (2) **Function usage error** refers to cases where the model has entered the tool operation pro-

cess, but makes operational mistakes in data access, worksheet organization, reference construction, or function usage. This reflects a mixed problem involving both tool use and actuarial knowledge. (3) **Modeling and calculation error** refers to cases where the model successfully completes the tool execution and produces an evaluable result, but obtains an answer inconsistent with the reference due to errors in actuarial model selection, formula specification, or numerical derivation; this reflects a pure actuarial knowledge problem.

4.2 Main Results

Current LLMs show strong actuarial knowledge mastery but still lack reliable actuarial work execution. Table 3 shows that proprietary LLMs perform strongly on standardized actuarial knowledge, but their scores decline sharply in case analysis and practice-oriented tool use. Gemini-3.1-Pro achieves the best total score and the highest INS-Act-Know score; Open-source models remain far behind, especially on CA and SS tasks. The results suggest that actuarial knowledge questions mainly test rule-based concepts and short-chain numerical reasoning, where frontier LLMs can rely on stable domain regularities and calculation schemas. Case and tool-use tasks require evidence localization, assumption tracking, cross-table computation, and executable workflow construction. INS-ActBench therefore identifies the boundary between actuarial knowledge mastery and actuarial work competence, providing a reproducible benchmark for developing LLMs that can support realistic actuarial workflows with verifiable outputs. This finding provides a practical roadmap for future actuarial LLM development: moving from strong exam-style knowledge performance toward trustworthy decision support across real insurance workflows.

Compared with human experts, LLMs lead in standardized actuarial knowledge but lag behind in case reasoning and practical tool use. Proprietary LLMs have exceeded actuarial experts on actuarial knowledge, but fall below on case and practice. Human experts are expected to remain stable across dimensions because actuarial training emphasizes applying principles under business context and practical calculation procedures. LLMs show more uneven results because their strengths are concentrated in standardized textual and numerical patterns, while long-context integration and executable tool-based workflows demand greater reliability. Recent studies show that although LLMs

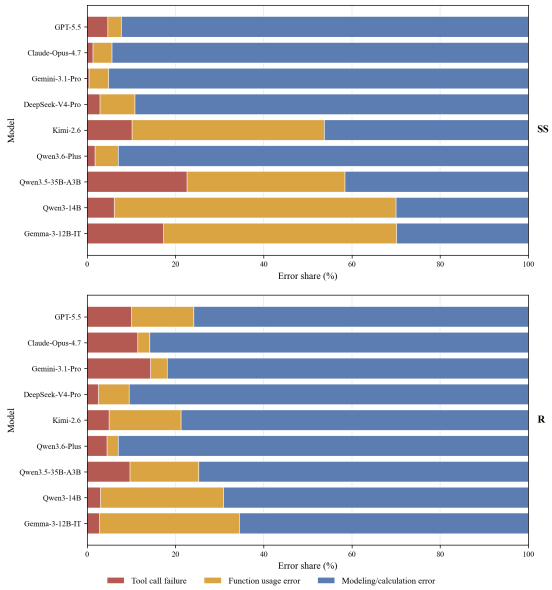


Figure 3: Error Analysis in INS-Act-Practice. Spreadsheet Tasks Upper and R-code Tasks Lower.

have achieved high accuracy on standard mathematical benchmarks, this capability does not directly translate into real-world applications(Cao et al., 2026). Our finding indicates that current LLMs have already demonstrated strong capabilities in actuarial calculation and knowledge checking, and future actuarial LLM development should prioritize context-grounded reasoning, reliable tool execution, and expert-level consistency across the full actuarial workflow. The results from each human expert are presented in Table 7 in Appendix C.1.

To further interpret main results, we conduct three additional analyses in next three subsections.

4.3 Actuarial Computation Results

Frontier LLMs have developed strong actuarial numerical reasoning, while weaker math reasoning LLMs still struggle with calculation actuarial tasks. Table 4 shows that frontier LLMs perform strongly on both numerical and non-numerical questions in INS-Act-Know based on Section 3.3.2, with numerical questions even yielding higher scores in most categories. This suggests that leading LLMs can already handle formula-based actuarial computation when the problem provides standard models, and well-defined calculation steps. However, LLMs with weaker numerical reasoning ability show much lower accuracy on numerical questions, indicating that actuarial numerical reasoning remains a major capability barrier for many LLMs. This decomposition therefore helps

	Model	PS		EF		AMA		LI		NLI		AMO		Total		
		Num	Non-num	Num	Non-num	Num	Non-num	Num	Non-num	Num	Non-num	Num	Non-num	Num	Non-num	Gap
Proprietary LLMs	Gemini-3.1-Pro	**97.6**	**94.6**	**96.0**	94.4	**95.1**	**90.8**	**90.5**	**93.0**	**94.0**	**94.0**	**94.1**	**93.5**	**94.6**	93.9	+0.7
	GPT-5.5	97.2	92.9	95.0	**96.1**	93.3	**90.8**	86.9	92.1	93.4	92.0	92.5	92.4	93.1	**94.4**	-1.3
	DeepSeek-V4-Pro	96.6	92.9	93.6	92.2	92.5	88.2	83.1	87.7	90.7	87.3	91.1	89.1	91.3	90.7	+0.6
	Qwen3.6-Plus	95.1	92.9	91.1	89.9	89.8	**90.8**	73.3	77.2	88.8	92.0	88.9	85.9	87.7	88.9	-1.2
	Claude-Opus-4.7	84.4	92.9	82.3	91.3	75.9	86.8	71.9	87.7	82.8	88.0	81.2	90.2	79.8	90.3	-10.5
	Kimi-K2.6	78.7	82.1	72.4	82.0	63.6	80.3	56.9	60.5	75.3	82.0	69.2	90.2	69.6	80.5	-10.9
Open-source LLMs	Qwen3-14B	41.5	69.6	28.8	72.6	35.5	75.0	32.3	68.4	36.3	72.0	38.6	76.1	35.9	72.4	-36.5
	Qwen3.5-35B-A3B	42.5	76.8	25.7	76.0	31.0	72.4	27.4	67.5	32.4	75.3	34.0	76.1	32.9	74.9	-42.0
	Gemma-3-12B-IT	32.8	64.3	26.8	69.8	30.5	55.3	25.0	62.3	32.7	61.3	33.9	67.4	30.2	66.7	-36.5

Table 4: Performance on numerical and non-numerical questions in INS-Act-Know.

	Model	INS-Act-Know			INS-Act-Case			INS-Act-Practice			Total		
		SOA	IFoA	Others	SOA	IFoA	Others	SOA	IFoA	Others	SOA	IFoA	Others
Proprietary LLMs	GPT-5.5	97.7	95.2	91.4	**55.0**	43.4	74.0	80.3	**80.0**	**61.4**	**87.8**	85.4	87.9
	Claude-Opus-4.7	85.1	91.7	77.2	36.4	32.3	70.3	79.0	77.1	56.5	74.7	80.9	75.0
	Gemini-3.1-Pro	**98.6**	**95.6**	**93.0**	42.3	**57.6**	**80.2**	**86.8**	78.0	58.3	86.3	**87.5**	**89.3**
	DeepSeek-V4-Pro	97.7	91.7	89.1	37.1	32.6	70.7	72.2	79.3	50.9	83.6	81.2	84.8
	Kimi-K2.6	75.7	81.9	66.7	28.9	16.5	26.0	37.7	71.2	39.1	63.6	70.7	62.6
	Qwen3.6-Plus	95.7	93.5	84.0	36.1	36.1	71.8	56.8	74.7	49.9	80.9	82.3	80.3
Open-source LLMs	Qwen3.5-35B-A3B	30.5	68.6	33.1	12.3	12.0	45.8	29.5	61.3	29.4	26.7	59.2	33.3
	Qwen3-14B	32.4	72.0	35.3	18.9	11.4	36.3	20.9	50.0	16.7	28.9	59.9	33.6
	Gemma-3-12B-IT	29.6	62.9	29.5	15.4	7.0	30.8	17.6	41.3	11.9	25.9	51.6	27.9

Table 5: Cross-jurisdictional performance on INS-ActBench.

INS-ActBench identify whether a model's weakness comes from actuarial calculation, conceptual understanding, or both, and highlights actuarial numerical reasoning as an important direction for future LLM improvement.

4.4 Practice Error Analysis

Failures in INS-Act-Practice are mainly concentrated in knowledge-related practice capabilities. Figure 3 shows that across both SS and R tasks, the tool call failures account for only a small share for LLMs. This pattern indicates that current LLMs can usually enter the tool-use workflow and produce executable outputs, but still struggle to transform spreadsheet/R operations into correct actuarial procedures. These results show that INS-Act-Practice evaluates a deeper form of actuarial competence: future actuarial LLMs need stronger integration between tool operation and actuarial reasoning, rather than only better tool invocation. More results are provided in Appendix C.3.

4.5 Cross-Jurisdictional Performance

Actuarial reasoning by LLMs remains sensitive to regulatory context, especially when tasks move beyond standardized knowledge into case analysis and practice. Table 5 shows clear performance variation across SOA, IFoA, and Other sources, and this variation becomes more visible in case analysis and practice-oriented tasks than in standardized knowledge questions. This suggests that LLMs do not handle actuarial tasks uniformly across regulatory environments, even when the tasks belong to the same professional domain. The reason is that insurance is strongly shaped by jurisdiction-specific regulation, so actuarial reasoning often depends on local institutional context rather than general abstract technical knowledge alone. These results indicate that actuarial LLMs should be evaluated and improved under diverse global regulatory settings instead of being optimized for a single actuarial environment.

5 Conclusion

In this paper, we introduced INS-ActBench, a large-scale benchmark for evaluating LLMs in actuarial science. INS-ActBench covers actuarial knowledge, long-context case analysis, and practice-oriented tool use with spreadsheets and R code. Experiments on nine leading LLMs and human experts show that current models perform strongly on standardized actuarial calculations, yet still lag behind experts in long-context case reasoning and tool-based numerical workflows. These findings highlight actuarial science as a rigorous domain for testing financial reasoning and provide a benchmark for future actuarial LLM research.

Limitations

First, although INS-ActBench covers actuarial knowledge, long-context case reasoning, and practice-oriented tool use, its tasks are constructed from static public exams, sample questions, and case materials. This setting may not fully capture dynamic actuarial workflows that involve client interaction, iterative assumption updates, and long-horizon monitoring. Second, some subjective actuarial questions are converted into objective formats for reproducible evaluation, which improves scoring consistency but may reduce the openness of real professional judgment. Third, INS-Act-Practice focuses on spreadsheet and R-code tasks; proprietary actuarial systems, database workflows, and interactive spreadsheet editing are not yet included. Finally, although the benchmark covers major actuarial source categories and broader local associations, its results may still reflect the distribution of available public materials and may not represent every regulatory environment, language context, or firm-specific actuarial practice.

Ethical Considerations

INS-ActBench is constructed from publicly available professional materials and is intended for research and educational evaluation. The benchmark does not contain private personal information, confidential policyholder records, or proprietary corporate data. Because actuarial work is closely related to insurance pricing, reserving, solvency assessment, and regulatory reporting, we emphasize that benchmark performance should not be interpreted as professional actuarial qualification or as a basis for real financial decisions. LLM outputs evaluated by INS-ActBench require review by certified actuarial professionals before any practical use. The dataset is designed to measure model capability and identify failure modes in actuarial reasoning, tool use, and jurisdiction-sensitive interpretation, and the authors do not assume responsibility for downstream decisions made using models evaluated on this benchmark.

Acknowledgements

We used generative AI tools only for language polishing and short-form input assistance during the writing process. This assistance was limited to language-level refinement.

References

Viral V Acharya, John Biggs, Matthew Richardson, and Stephen Ryan. 2009. On the financial regulation of insurance companies. *NYU Stern School of Business*.

Caesar Balona. 2024. Actuarygpt: Applications of large language models to insurance and actuarial work. *British Actuarial Journal*, 29:e15.

Mary Pat Campbell. 2010. Spreadsheet issues: pitfalls, best practices, and practical tips. In *Actuarial Practice Forum*, pages 1–39.

Bowen Cao, Dongdong Zhang, Yixia Li, Junpeng Liu, Shijue Huang, Chufan Shi, Hongyuan Lu, Yaokang Wu, Guanhua Chen, Wai Lam, and 1 others. 2026. From abstract to contextual: What llms still cannot do in mathematics. *arXiv preprint arXiv:2601.23048*.

Shisong Chen, Qian Zhu, Wenyan Yang, Chengyi Yang, Zhong Wang, Ping Wang, Xuan Lin, Bo Xu, Daqian Li, Chao Yuan, and 1 others. 2025. Inseva: A comprehensive chinese benchmark for large language models in insurance. *arXiv preprint arXiv:2509.04455*.

Jing Ding, Kai Feng, Binbin Lin, Jiarui Cai, Qiushi Wang, Yu Xie, Xiaojin Zhang, Zhongyu Wei, and Wei Chen. 2025. Insqabench: Benchmarking chinese insurance domain question answering with large language models. *arXiv preprint arXiv:2501.10943*.

Haoyu Dong, Pengkun Zhang, Yan Gao, Xuanyu Dong, Yilin Cheng, Mingzhe Lu, Zikun Zhu, Adina Yakefu, and Shuxin Zheng. 2025. Finch: Benchmarking finance & accounting across spreadsheet-centric enterprise workflows. *arXiv preprint arXiv:2512.13168*.

Christophe Dutang, Vincent Goulet, and Mathieu Pigeon. 2008. actuar: An r package for actuarial science. *Journal of Statistical software*, 25:1–37.

Oscar Espinosa and Armando Zarruk. 2021. The importance of actuarial management in insurance business decision-making in the twenty-first century. *British Actuarial Journal*, 26:e14.

Ziliang Gan, Dong Zhang, Haohan Li, Yang Wu, Xueyuan Lin, Ji Liu, Haipang Wu, Chaoyou Fu, Zenglin Xu, Rongjunchen Zhang, and 1 others. 2025. Mme-finance: A multimodal finance benchmark for expert-level understanding and reasoning. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 12867–12874.

Yi He, Ping Wang, Shiqiang Xiong, Chao Chen, and Haixiang Hu. 2026. Finmathbench: A formula-driven benchmark for evaluating llms' math reasoning capabilities in finance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 30987–30995.

Pranab Islam, Anand Kannappan, Douwe Kiela, Rebecca Qian, Nino Scherrer, and Bertie Vidgen. 2023. Financebench: A new benchmark for financial question answering. *arXiv preprint arXiv:2311.11944*.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th symposium on operating systems principles*, pages 611–626.

Haohang Li, Yupeng Cao, Yangyang Yu, Shashidhar Reddy Javaji, Zhiyang Deng, Yueru He, Yuechen Jiang, Zining Zhu, Kp Subbalakshmi, Jimin Huang, and 1 others. 2025. Investorbench: A benchmark for financial decision-making tasks with llm-based agent. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2509–2525.

Chenwei Lin, Hanjia Lyu, Xian Xu, and Jiebo Luo. 2025. Ins-mmbench: A comprehensive benchmark for evaluating lvlms' performance in insurance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9036–9047.

Tatjana Miljkovic. 2017. Computational actuarial science with r.

Ying Nie, Binwei Yan, Tianyu Guo, Hao Liu, Haoyu Wang, Wei He, Binfan Zheng, Weihao Wang, Qiang Li, Weijian Sun, and 1 others. 2025. Cfinbench: A comprehensive chinese financial benchmark for large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 876–891.

MI Owadally, Feng Zhou, and ID Wright. 2018. The insurance industry as a complex social system: Competition, cycles, and crises. *Journal of Artificial Societies and Social Simulation*, 21(4).

Jan Ravnik, Matjaž Ličen, Felix Bührmann, Bithiah Yuan, Felix Stinson, and Tanvi Singh. 2026. Finsheetbench: From simple lookups to complex reasoning, where llms break on financial spreadsheets. *arXiv preprint arXiv:2603.07316*.

Varshini Reddy, Rik Koncel-Kedziorski, Viet Dac Lai, Michael Krumdick, Charles Lovering, and Chris Tanner. 2024. Docfinqa: A long-context financial reasoning dataset. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 445–458.

Xinyi Song, Lina Lee, Kexin Xie, Xueying Liu, Xinwei Deng, and Yili Hong. 2026. Statllm: A dataset for evaluating the performance of large language models in statistical analysis. *Scientific Data*.

Zichen Tang, E Haihong, Ziyan Ma, Haoyang He, Jiacheng Liu, Zhongjun Yang, Zihua Rong, Rongjin Li, Kun Ji, Qing Huang, and 1 others. 2025. Financereasoning: Benchmarking financial numerical reasoning more credible, comprehensive and challenging. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15721–15749.

Qianqian Xie, Weiguang Han, Zhengyu Chen, Ruoyu Xiang, Xiao Zhang, Yueru He, Mengxi Xiao, Dong Li, Yongfu Dai, Duanyu Feng, and 1 others. 2024. Finben: A holistic financial benchmark for large language models. *Advances in Neural Information Processing Systems*, 37:95716–95743.

Qianqian Xie, Weiguang Han, Xiao Zhang, Yanzhao Lai, Min Peng, Alejandro Lopez-Lira, and Jimin Huang. 2023. Pixiu: A comprehensive benchmark, instruction dataset and large language model for finance. *Advances in Neural Information Processing Systems*, 36:33469–33484.

Yilun Zhao, Yunxiang Li, Chenying Li, and Rui Zhang. 2022. Multihiertt: Numerical reasoning over multi hierarchical tabular and textual data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6588–6600.

Hua Zhou, Bing Ma, Yufei Zhang, and Yi Zhao. 2025. Design, results and industry implications of the world's first insurance large language model evaluation benchmark. *arXiv preprint arXiv:2511.07794*.

Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. 2021. Tat-qa: A question answering benchmark on a hybrid of tabular and textual content in finance. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3277–3287.

A Global Actuarial Association

According to the Members with Actuarial Credentials (MWACs) statistics released by the International Actuarial Association (IAA) in March 2026[1], SOA/IFoA/CAS are the three largest actuarial associations by reported credentialed members, together accounting for 57.84% of global MWACs. This indicates that the North American actuarial system, represented by SOA and CAS, and the British actuarial system, represented by IFoA, are the two most influential global actuarial credentialing systems. The remaining 71 actuarial associations account for 42.16% of MWACs and reflect diverse local regulatory, examination, and professional practice environments. Details are shown in Table 6. Therefore, INS-ActBench uses SOA/CAS and IFoA as the two major global sources, while further incorporating 13 other actuarial associations to capture broader jurisdiction-specific actuarial characteristics. They are:

China Association of Actuaries (CAA); Institute of Actuaries of Japan (IAJ); Institute of Actuaries of India (IAI); Persatuan Aktuaris Indonesia / Society of Actuaries of Indonesia (PAI); Institute of Actuaries of Korea (IAK); Deutsche Aktuarvereinigung / German Association of Actuaries (DAV); Instituto Brasileiro de Atuária / Brazilian Institute of Actuaries (IBA); Actuarial Society of Hong Kong (ASHK); Canadian Institute of Actuaries (CIA); Actuarial Society of South Africa (ASSA); Japanese Society of Certified Pension Actuaries (JSCPA); Actuaries Institute Australia (AIA); Israel Association of Actuaries (ISOA).

No.	Association	Members	Share
1	SOA	34,466	30.99%
2	IFoA	18,406	16.55%
3	CAS	11,451	10.30%
4	Others	46,890	42.16%
–	Total	111,213	100.00%

Table 6: Top actuarial associations by reported MWACs.

B Dataset Details

Figure 4 reports the system prompts used for these four question types. Figure 5 shows the context-length distribution of the case-analysis questions. Figures 6 - 9 present representative examples of the four question types in INS-ActBench.

[1] https://actuaries.org/committees/membership-assistance/membership-assistance-reference-materials/

In quality verification, the process checked: (1) whether each question is fluent and contains sufficient information for deriving the answer; (2) whether the final answer is consistent with the original reference; (3) whether the case background is complete and the simplified version preserves necessary information; (4) whether spreadsheet answer cells are correctly annotated; and (5) whether R-code reference outputs are correct.

C Additional Results

C.1 Human Expert

The three human experts show broadly consistent performance on INS-ActBench. Their total scores fall within a narrow range, indicating limited variation across individual experts. Their performance is also relatively stable across the three subsets.

Regarding compensation, the human experts were compensated at a rate equivalent to one day of their regular pay.

C.2 Different INS-Act-Case Scoring

When we apply a partial-credit rule to the multiple-answer questions in INS-Act-Case, where partially selected correct options receive half credit, all models show a clear score increase compared with the original exact-match scoring scheme. The comparison is reported in Table 8. This result indicates that LLMs have already captured part of the relevant information in actuarial cases, but often fail to recover the complete set of correct options. From the perspective of error analysis, the large improvement under partial credit suggests that their case-analysis errors are more often caused by under-selection than by selecting incorrect options. In other words, models tend to miss some necessary case conditions, overlook part of the evidence, or incompletely integrate multiple pieces of case information, rather than systematically misidentifying the meaning of the case materials.

C.3 More INS-Act-Practice Results

Regarding code efficiency, we have recorded the formula execution time of Python code answers in spreadsheet tasks and the execution time of R code in Docker in R tasks. The differences among models are minimal. Additionally, actuarial science places primary emphasis on numerical correctness. Therefore, efficiency is not used as a core scoring criterion in our benchmark, and is reported only as a reference result, details are shown in Table 9.

System Prompt: INS-Act-Know	System Prompt: INS-Act-Case
You are an expert actuarial analyst solving the multiple-choice question. Output format rules: - Respond with exactly ONE option letter - Do not output any other text.	You are an expert actuarial analyst solving case-based actuarial multiple-select questions. You will be given case materials and, when relevant, supporting tables that may be useful for solving the question. For each item, use the provided stem as the background and answer the question under that stem. This is a multiple-select question. At least TWO options must be selected. Scoring rule: credit is awarded only if all and only the correct options are selected; any missing option, extra option, incorrect option, or empty response receives zero credit. Output format rules: - Respond with the selected option letters only - Do not output any explanation, reasoning, words, punctuation, or extra text - Write the capital letters together in alphabetical order, for example: AB
System Prompt: Spreadsheet Task of INS-Act-Practice	**System Prompt: R-code Task of INS-Act-Practice**
You are solving an actuarial spreadsheet task. Write one executable Python code block using openpyxl. Load the input workbook from INPUT_WORKBOOK, or input.xlsx if the variable is missing. Use the workbook snapshot and task text to understand the sheets, cells, and formulas. Fill the final answer cells in ANSWER_POSITION; you may use other existing cells for intermediate calculations. Prefer spreadsheet formulas over hard-coded final values when formulas can be built from the workbook data. Do not rename sheets, create extra sheets, or overwrite unrelated content. Save the completed workbook to OUTPUT_WORKBOOK, or output.xlsx if the variable is missing. Output only the Python code block, with no explanation.	You are solving an actuarial R coding test. Your task is to write accurate R code to compute and solve the given actuarial question. CRITICAL Output Format: - Output ONLY valid, executable R code enclosed in a single R code block. - Do NOT output any standard text explanations, preambles, or concluding remarks. CRITICAL Code Rules: - Assume a fresh base R session. Do not install external packages. - Recreate any inline vectors/matrices exactly. - If datasets like .RData or .csv are provided, assume they are available in the working directory and load them directly. - ALL final numerical answers, statistics, or metrics MUST be explicitly printed to the console so they can be captured by standard output.

Figure 4: System prompts for the four task types in INS-ActBench.

Human	INS-Act-Know							INS-Act-Case	INS-Act-Practice			Total
	PS	EF	AMA	LI	NLI	AMO	Total	CA	SS	R	Total	
Human Expert 1	91.67	91.67	100.00	75.00	91.67	91.67	90.28	80.00	77.88	90.00	83.94	87.73
Human Expert 2	83.33	91.67	100.00	91.67	83.33	83.33	88.89	70.00	64.03	80.00	72.02	83.04
Human Expert 3	75.00	91.67	100.00	75.00	66.67	100.00	84.72	80.00	87.13	100.00	93.57	86.16

Table 7: Individual human expert performance on INS-ActBench.

Model	INS-Act-Case	INS-Act-Case (partial credit)
Proprietary LLMs		
GPT-5.5	**56.3**	64.3
Claude-Opus-4.7	43.3	57.5
Gemini-3.1-Pro	55.5	**69.5**
DeepSeek-V4-Pro	43.9	58.1
Kimi-K2.6	24.8	37.7
Qwen3.6-Plus	44.6	62.1
Open-source LLMs		
Qwen3.5-35B-A3B	20.2	40.0
Qwen3-14B	21.0	31.2
Gemma-3-12B-IT	16.7	30.6

Table 8: Partial-credit re-scoring for INS-Act-Case.

Model	Spreadsheet				R Code			
	Min	Max	Median	Mean	Min	Max	Median	Mean
Proprietary LLMs								
GPT-5.5	1.02	7.83	1.58	1.83	0.43	1.63	0.54	0.57
Claude-Opus-4.7	0.97	6.07	1.52	1.64	0.44	1.93	0.59	0.64
Gemini-3.1-Pro	1.03	9.01	1.57	1.79	0.42	3.00	0.55	0.61
DeepSeek-V4-Pro	1.05	8.72	1.78	2.04	0.43	1.57	0.55	0.59
Kimi-K2.6	1.10	3.80	1.75	1.80	0.46	2.03	0.58	0.64
Qwen3.6-Plus	0.93	10.00	1.67	1.90	0.43	3.03	0.57	0.64
Open-source LLMs								
Qwen3.5-35B-A3B	0.97	7.62	1.59	1.68	0.43	1.47	0.51	0.54
Qwen3-14B	1.32	4.86	2.29	2.41	0.48	2.52	0.59	0.67
Gemma-3-12B-IT	1.25	3.56	2.01	2.06	0.45	1.40	0.55	0.59

Table 9: Execution time statistics for INS-Act-Practice (Unit: second).

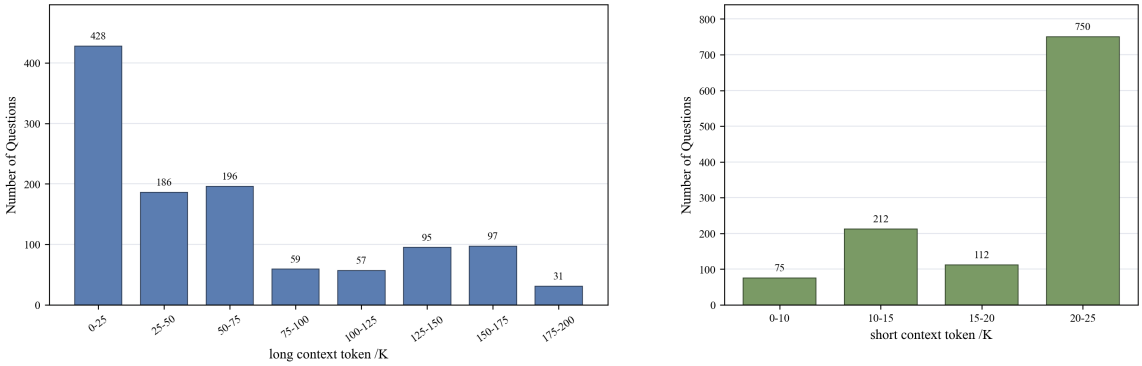


Figure 5: Context-token distribution.

Sample Data Instance: Annotated Multiple-Choice Question

Annotation:

Dataset: INS-Act-Know	**Datasubset:** Life Insurance (LI)	**Source:** SOA	**Numerical:** yes	**id:** 58

Question:
For an annuity-due that pays 100 at the beginning of each year that (45) is alive, you are given:

(i) Mortality for standard lives follows the Standard Ultimate Life Table.
(ii) The force of mortality for standard lives age $45 + t$ is represented as μ_{45+t}^{SULT}.
(iii) The force of mortality for substandard lives age $45 + t$, μ_{45+t}^{S}, is defined as:

$$\mu_{45+t}^{S} = \begin{cases} \mu_{45+t}^{SULT} + 0.05, & \text{for } 0 \leq t < 1 \\ \mu_{45+t}^{SULT}, & \text{for } t \geq 1 \end{cases}$$

(iv) $i = 0.05$

Calculate the actuarial present value of this annuity for a substandard life age 45.

Options: (A) 1700 (B) 1710 (C) 1720 (D) 1730 (E) 1740

Answer: A

Figure 6: A representative example of a Life Insurance question from the INS-Act-Know subset.

Sample Data Instance: Case-based Multiple-select Question

Metadata Annotation:

Dataset: INS-Act-Case	**Source:** SOA	**id:** 113

Case-long:
1.1 RPPC Corporation History
RPPC was established in 2005 with head offices in Luxembourg by four founding partners...(132K)

Case-short:
1.1 RPPC Corporation History
The business roots began with the coffee shop, owned by the Ruiz family since 1995...(24K)

Stem:
You are engaged by Blue Jay Air (BJA) to provide advice on its international expansion strategy. Since being separated from RPPC, BJA has continued pursuing ambitious growth opportunities. (Case Study Section 2.7) Research by your team shows that the median debt-to-equity ratio in the airline industry is 5.5x.
The BJA Finance Team updated the analysis of two alternatives for the international expansion project: one is to purchase the international plane fleet, and the other is to upgrade the existing plane fleet for international travel. This capital budgeting model is in tab Q2_b and uses the companywide cost of capital and expected free cash flows to calculate net present value. BJA needs an independent review of this model in order to obtain a fixed debt financing agreement.

Question:
Based on the Excel file and the financial statements, which of the following statements about BJA's net debt-to-value ratio and capital structure evolution over the past three years are correct?

Options:
(A) the 2020 net debt-to-value ratio was approximately 96%
(B) the 2021 net debt-to-value ratio was approximately 86%
(C) the 2022 net debt-to-value ratio was approximately 64%
(D) BJA's leverage decreased over the past three years
(E) BJA's debt-to-equity ratio has fallen below the industry median of 5.5x
(F) retained earnings over the past three years were all converted into cash dividends, thereby depressing equity
(G) net debt in the past three years was 349, 469, and 538 respectively
(H) shareholder equity in the past three years was 199, 89, and 22 respectively

---

Answer: ABCDEG

Figure 7: A representative example of a case-based multiple-select question from the INS-Act-Case subset.

Sample Data Instance: Spreadsheet Question

Metadata Annotation:

Dataset: INS-Act-Practice	**Datasubset:** Spreadsheet	**Source:** IFoA	**id:** 2

Question:
Calculate the annual effective inflation rate over the previous 12 months for each month from January 2004 to December 2019 using the index values provided.

Template: See Figure 10. **Answer Position:** `'G2 (i)'!D17:D208`

Answer: See Figure 11.

Figure 8: A representative example of a spreadsheet question from the INS-Act-Practice subset.

Sample Data Instance: Annotated R-Code Question

Metadata Annotation:

Dataset: INS-Act-Practice	**Datasubset:** R-Code	**Source:** IFoA	**id:** 39

Question:
An analysis was carried out to investigate the fairness of two exam markers. They both marked the same 150 exam papers, with 10 questions and total possible marks of 100 for each exam paper. The data were collected and arranged into 10 equally spaced groups, with marks rounded to the nearest whole number.

Below are the frequencies of the marks given by each of the exam markers:

Exam marks	**0–10**	**11–20**	**21–30**	**31–40**	**41–50**	**51–60**	**61–70**	**71–80**	**81–90**	**91–100**
Marker 1	1	8	14	22	33	34	21	9	6	2
Marker 2	0	4	16	25	27	42	23	4	9	0

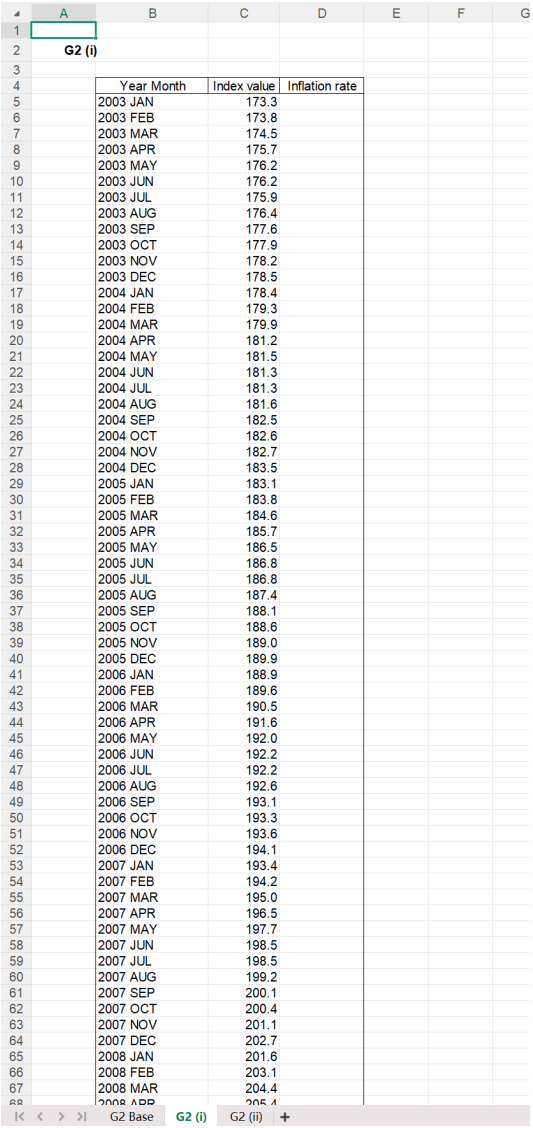
One of the marked exam papers is selected at random and the scores given by each of the markers are analysed further by question. Below are the scores given by each of the exam markers for this selected exam paper for each of the 10 questions:

Question	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Marker 1	4	1	5	1	4	6	4	5	3	6
Marker 2	3	2	4	0	3	4	2	3	3	6

Perform a suitable test to determine whether the difference in the mean scores of the two markers is zero or not, at the 5% confidence level, taking into account that the two markers have marked the same exam paper.

Ground Truth (Reference Output): `statistic = 2.86199428; p_value = 0.01871942`
Scoring: `[2.861994, 0.018719]`

Figure 9: A representative example of an R-code question from the INS-Act-Practice subset.

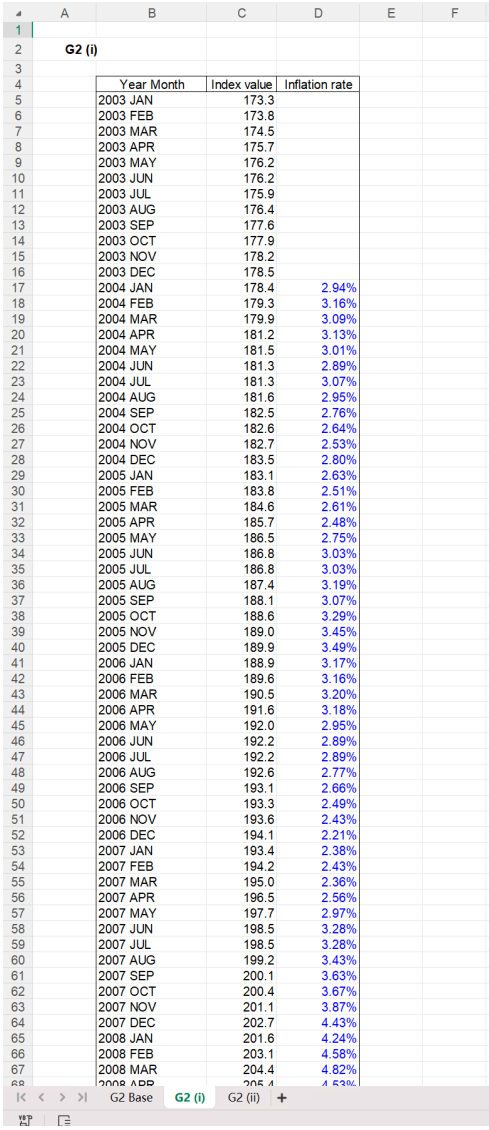


G2 (i)

Year Month	Index value	Inflation rate
2003 JAN	173.3	
2003 FEB	173.8	
2003 MAR	174.5	
2003 APR	175.7	
2003 MAY	176.2	
2003 JUN	176.2	
2003 JUL	175.9	
2003 AUG	176.4	
2003 SEP	177.6	
2003 OCT	177.9	
2003 NOV	178.2	
2003 DEC	178.5	
2004 JAN	178.4	
2004 FEB	179.3	
2004 MAR	179.9	
2004 APR	181.2	
2004 MAY	181.5	
2004 JUN	181.3	
2004 JUL	181.3	
2004 AUG	181.6	
2004 SEP	182.5	
2004 OCT	182.6	
2004 NOV	182.7	
2004 DEC	183.5	
2005 JAN	183.1	
2005 FEB	183.8	
2005 MAR	184.6	
2005 APR	185.7	
2005 MAY	186.5	
2005 JUN	186.8	
2005 JUL	186.8	
2005 AUG	187.4	
2005 SEP	188.1	
2005 OCT	188.6	
2005 NOV	189.0	
2005 DEC	189.9	
2006 JAN	188.9	
2006 FEB	189.6	
2006 MAR	190.5	
2006 APR	191.6	
2006 MAY	192.0	
2006 JUN	192.2	
2006 JUL	192.2	
2006 AUG	192.6	
2006 SEP	193.1	
2006 OCT	193.3	
2006 NOV	193.6	
2006 DEC	194.1	
2007 JAN	193.4	
2007 FEB	194.2	
2007 MAR	195.0	
2007 APR	196.5	
2007 MAY	197.7	
2007 JUN	198.5	
2007 JUL	198.5	
2007 AUG	199.2	
2007 SEP	200.1	
2007 OCT	200.4	
2007 NOV	201.1	
2007 DEC	202.7	
2008 JAN	201.6	
2008 FEB	203.1	
2008 MAR	204.4	
2008 APR	205.4	

G2 Base | G2 (i) | G2 (ii)

Figure 10: Template workbook example.



G2 (i)

Year Month	Index value	Inflation rate
2003 JAN	173.3	
2003 FEB	173.8	
2003 MAR	174.5	
2003 APR	175.7	
2003 MAY	176.2	
2003 JUN	176.2	
2003 JUL	175.9	
2003 AUG	176.4	
2003 SEP	177.6	
2003 OCT	177.9	
2003 NOV	178.2	
2003 DEC	178.5	
2004 JAN	178.4	2.94%
2004 FEB	179.3	3.16%
2004 MAR	179.9	3.09%
2004 APR	181.2	3.13%
2004 MAY	181.5	3.01%
2004 JUN	181.3	2.89%
2004 JUL	181.3	3.07%
2004 AUG	181.6	2.95%
2004 SEP	182.5	2.76%
2004 OCT	182.6	2.64%
2004 NOV	182.7	2.53%
2004 DEC	183.5	2.80%
2005 JAN	183.1	2.63%
2005 FEB	183.8	2.51%
2005 MAR	184.6	2.61%
2005 APR	185.7	2.48%
2005 MAY	186.5	2.75%
2005 JUN	186.8	3.03%
2005 JUL	186.8	3.03%
2005 AUG	187.4	3.19%
2005 SEP	188.1	3.07%
2005 OCT	188.6	3.29%
2005 NOV	189.0	3.45%
2005 DEC	189.9	3.49%
2006 JAN	188.9	3.17%
2006 FEB	189.6	3.16%
2006 MAR	190.5	3.20%
2006 APR	191.6	3.18%
2006 MAY	192.0	2.95%
2006 JUN	192.2	2.89%
2006 JUL	192.2	2.89%
2006 AUG	192.6	2.77%
2006 SEP	193.1	2.66%
2006 OCT	193.3	2.49%
2006 NOV	193.6	2.43%
2006 DEC	194.1	2.21%
2007 JAN	193.4	2.38%
2007 FEB	194.2	2.43%
2007 MAR	195.0	2.36%
2007 APR	196.5	2.56%
2007 MAY	197.7	2.97%
2007 JUN	198.5	3.28%
2007 JUL	198.5	3.28%
2007 AUG	199.2	3.43%
2007 SEP	200.1	3.63%
2007 OCT	200.4	3.67%
2007 NOV	201.1	3.87%
2007 DEC	202.7	4.43%
2008 JAN	201.6	4.24%
2008 FEB	203.1	4.58%
2008 MAR	204.4	4.82%
2008 APR	205.4	4.53%

G2 Base | G2 (i) | G2 (ii)

Figure 11: Ground-truth answer example.