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

Anonymous ACL submission

Abstract

While Large Language Models (LLMs) have shown strong performance in general financial tasks, their capabilities in **actuarial science**—the quantitative foundation of the insurance industry—remain insufficiently evaluated. Existing benchmarks are largely limited to knowledge-oriented question answering or capital market–focused tasks, and fail to assess practical actuarial modeling and execution skills. To bridge this gap, we introduce **INS-ActBench**, a comprehensive benchmark engineered to shift the evaluation paradigm from "declarative knowledge" to "professional capability." Grounded in the intersection of cross-jurisdictional competency actuarial frameworks and Bloom's Taxonomy, we construct a rigorous four-tier benchmark comprising 6,514 authentic tasks. By integrating six innovative task type, our pipeline validates models against real-world professional standards. Extensive evaluation reveals a distinct "*Strong Theory, Weak Practice*" phenomenon: while models exhibit proficiency in conceptual calculation, their performance deteriorates significantly in tasks requiring precise tool manipulation and multi-step logical derivation. These findings suggest that current LLMs are best positioned as assistants rather than autonomous actuarial agents, providing a critical quantitative baseline for the responsible deployment of LLMs in high-stakes financial risk management. The codes and data are available at https://anonymous.4open.science/r/ActuarialBench-3B5D.

1 Introduction

Recent advancements in Large Language Models (LLMs) have signaled a paradigm shift in artificial intelligence, demonstrating remarkable proficiency across a broad spectrum of general tasks (Naveed et al., 2025). Particularly in the finance sector, LLMs have shown impressive utility in market analysis, investment advisory, and report interpretation (Xie et al., 2024). However, within this broad "pan-financial" landscape, the capabilities of LLMs in actuarial science remain largely unexplored. Far from simple financial arithmetic, actuarial science is a highly complex interdisciplinary field that fuses mathematics, finance, economics to quantify and price future uncertainties. As a domain characterized by its stringent regulatory frameworks and a zero-tolerance policy for numerical inaccuracy, actuarial work directly determines insurance pricing strategies and the solvency of financial institutions(Artzner et al., 1999). Consequently, while LLMs are skilled in general financial tasks, their ability to navigate this specialized domain, such as high mathematical density, rigid compliance constraints, and high-stakes financial impact, remains an underexplored "black box."

However, existing benchmarks are ill-suited for evaluating LLM capabilities in actuarial science. Current financial benchmarks predominantly focus on capital market–oriented tasks, such as investment decision-making, information extraction, and return prediction (*e.g.,* TAT-QA (Zhu et al., 2021), MultiHiertt (Zhao et al., 2022), PIXIU (Xie et al., 2023), FinBen (Xie et al., 2024), InvestorBench (Li et al., 2025a)). Similarly, evaluations in the insurance domain remain largely restricted to "*Insurance Knowledge Service*" scenarios, targeting customer consultation, clause retrieval, and compliance checking (e.g., InsuranceQA (Feng et al., 2015), InsQABench (Ding et al., 2025), INSEva (Chen et al., 2025), CUFEInse (Zhou et al., 2025), INS-MMBench (Lin et al., 2025)). While isolated studies like ActuaryGPT (Balona, 2024) have hinted at the computational potential of LLMs, the industry still lacks a systematic framework to quantify professional capability in core actuarial practices—specifically, the execution of complex mathematical modeling to solve business problems.

This absence of evaluation dimensions conceals the profound challenges LLMs face when address-

ing actuarial tasks. Unlike general mathematical benchmarks such as GSM8K (Cobbe et al., 2021), actuarial problems are not merely logical deductions within a closed environment but rather involve the deep integration of heterogeneous knowledge. A qualified actuarial response often requires extracting dispersed business assumptions from long contexts, constructing models using actuarial theory, and strictly adhering to the regulatory statutes and ethical codes of specific jurisdictions. Current general-purpose models, while fluent in explaining insurance terminology, often suffer from hallucinations or logical disconnects when engaging in such comprehensive tasks that require long-chain reasoning, domain knowledge retrieval, and zero-error computation. Therefore, constructing a benchmark dedicated to "Actuarial Practical Capability" is critical for delineating the boundaries of LLMs in high-stakes financial scenarios.

To bridge this critical technical gap, we propose **INS-ActBench**—the first comprehensive benchmark designed to assess the Professional Actuarial Capability of LLMs. Moving beyond simple knowledge retention, we push the evaluation boundary from "knowing insurance" to "practicing actuarial science." Based on this, we make the following primary contributions:

Establishment of a cross-jurisdictional industry standard. By synthesizing certification frameworks from different Actuarial Association, we translate human professional entry standards into a unified AI evaluation metric. This multi-jurisdictional design provides the first benchmark capable of stress-testing model robustness across diverse regulatory environments in insurance.

Bridging the gap between theoretical cognition and professional execution. We introduce a hierarchical architecture that mirrors the actuarial career path, incorporating novel Excel modeling and R programming tasks. This design compels models to move beyond static knowledge retrieval, validating their ability to navigate complex toolchains and stochastic simulations required in authentic business workflows.

Delineation of capability boundaries for responsible deployment. Our extensive evaluation reveals a distinct "Strong Theory, Weak Practice" pattern in current LLMs. This finding quantifies the performance disparity between compliance interpretation and autonomous modeling, providing financial institutions with experimental grounds to position LLMs as assistive knowledge engines rather than independent actuarial agents.

Figure 1: INS-ActBench Task Hierarchy and Theory.

2 Methodology

2.1 Task Classification and Benchmark Design

INS-ActBench adopts a four-tier competency assessment architecture mirroring the professional development path from theoretical foundations to advanced practice, as illustrated in Figure 1. The design of this architecture is grounded in the intersection of established actuarial standards. We first draw upon the Casualty Actuarial Society (CAS) three-level proficiency model, which establishes a vertical progression axis capturing the deepening of expertise from entry-level knowledge to strategic decision-making[1]. Synergizing with this vertical structure, we incorporate the eight capability dimensions defined by the Actuaries Institute Australia to form a horizontal coverage framework[2]. This multidimensional integration ensures that INS-ActBench evaluates not merely computational accuracy, but the full breadth of technical skills, business acumen, and professional judgment required in authentic practice. Furthermore, we anchor our task taxonomy in Bloom's Taxonomy (Bloom et al., 1956), a validated cognitive hierarchy distinguishing between lower-order

[1] https://www.casact.org/professional-education/cas-capability-model

[2] https://www.actuaries.asn.au/professional-standards-and-regulation/actuarial-capabilities-framework

skills (remembering, understanding) and higher-order reasoning (analyzing, evaluating, creating). This alignment enables fine-grained diagnosis of model capabilities, identifying whether failures stem from knowledge gaps, application errors, or deficiencies in complex synthesis. Collectively, these three frameworks—vertical professional progression, horizontal capability breadth, and cognitive depth—provide INS-ActBench with theoretical rigor, ensuring the evaluation is both professionally grounded and pedagogically interpretable.

2.1.1 Level 1: Foundational Knowledge

This level evaluates the mastery of mathematical and economic fundamentals required for all actuarial work. It consists entirely of Multiple Choice Questions (MCQ) assessing computational accuracy and theoretical understanding.

Probability (MCQ), LV1.1. Questions cover probability fundamentals, including conditional probability, Bayes' theorem, common distribution families, multivariate random variables, moment generating functions, limit theorems, among others. These concepts constitute the mathematical basis for all actuarial models.

Economics & Finance (MCQ), LV1.2. Items involve macroeconomics and microeconomics, time value of money, loan amortization schedules, bond pricing and yield calculations, duration and convexity, and immunization strategies. These skills are foundational for all financial experts.

Actuarial Mathematics (MCQ), LV1.3. Topics cover interest theory, survival model basics, risk measures. This category bridges pure mathematics and actuarial application, serving as entry-level knowledge for actuaries.

2.1.2 Level 2: Core Actuarial Models

Level 2 focuses on professional actuarial knowledge within specific insurance business contexts. Items remain MCQs but involve complex scenarios requiring the application of multiple concepts.

Life Insurance (MCQ), LV2.1. Covers life actuarial modeling, including life table construction, survival annuities, life product pricing, policy reserve valuation (net/gross premium reserves), and cash value calculations. It assesses expertise in long-term mortality risk assessment and pricing.

Non-Life Insurance (MCQ), LV2.2. Involves property and casualty practice, including loss distribution modeling, ratemaking procedures, loss reserving (Chain-Ladder, Bornhuetter-Ferguson, Cape Cod methods), and reinsurance, which evaluate the ability to handle short-term risks.

Actuarial Modeling (MCQ), LV2.3. Covers actuarial statistical modeling techniques, including stochastic simulation, time series analysis, survival analysis (Kaplan-Meier, Cox proportional hazards), Generalized Linear Models in insurance, and ruin theory. This assesses LLM mastery of complex actuarial prediction methods.

2.1.3 Level 3: Practical Skills

Level 3 bridges the gap between test exams and practical work, evaluating tool applications, regulatory knowledge, and professional ethics. This level includes mixed formats: MCQs, fill-in-the-blank, Excel modeling, and code generation.

Professional Ethics (MCQ), LV3.1. Based on the codes of conduct from the SOA, IFoA, and CAA, these items present ethical dilemmas involving conflicts of interest, confidentiality, and competence. A key feature is the Dynamic System Prompt: the system automatically switches the system prompt based on the question's jurisdiction, instructing the model to act as an actuary under that specific jurisdiction's standards.

Regulation (Fill-in-the-Blank), LV3.2. Tests the application of actuarial laws and standards. Content covers solvency frameworks (RBC, Solvency II, C-ROSS), accounting standards, and jurisdiction-specific regulations. Formats include *Exact Completion* (completing regulatory text) and *Source Identification* (identifying the clause source). This highlights the Cross-Jurisdictional Dimension of INS-ActBench, testing the ability to search and distinguish diverse international rules.

Excel Tasks (Table), LV3.3. As Excel remains the primary tool for actuaries, this task presents authentic spreadsheet modeling scenarios. LLMs receive a task description and partially structured Excel templates. Then LLMs must generate code (using the `openpyxl` library) to complete the spreadsheet by filling in missing formulas and values. This tests understanding of spreadsheet logic and cell dependencies. The example Excel templates are given in Appendix B.

Code Tasks (Code), LV3.4. Evaluating R programming for statistical analysis, LLMs receive a problem description and data files (.RData). They must generate executable R scripts to solve the problem. Generated codes are executed in an isolated Docker container with standard actuarial packages to ensure safety and consistency.

2.1.4 Level 4: Comprehensive Case Studies

Case Study (Case, LV4) represents the pinnacle of professional capability. Based on SOA Fellow-level case studies across seven tracks (CFE, GH, GI, ILA, INV, RET, CP), which requires LLMs to analyze extensive contextual materials—including company background, market data, and strategic goals. By employing long-context understanding and multi-step reasoning, LLMs must analyze interconnected tasks (e.g., product design, capital planning) and provide defensible strategic advice under uncertainty. This level includes open subjective case question-answer pairs.

2.2 Dataset Construction and Verification

2.2.1 Data Sources

INS-ActBench is derived entirely from public, authoritative professional materials, containing no synthetic data. We integrate materials from three major jurisdictions:

Society of Actuaries (SOA), U.S.A: The sample questions of Associate-level exams (P, FM, SRM, ALTAM, ASTAM) and Fellow-level case studies (2024 cycle).

Institute and Faculty of Actuaries (IFoA), U.K.: Due to copyright restrictions, we only use public samples from CS1-B (Actuarial Statistics) for R coding tasks and CM1-B (Actuarial Mathematics) for Excel tasks.

China Association of Actuaries (CAA), China: Simulated Associate-level exams covering probability, economics, actuarial modeling and life/non-life actuarial science.

Regulation and Ethics: We compiled representative documents (e.g. RBC, Solvency II, C-ROSS) from the US, UK, and China to develop fill-in-the-blank and ethical judgment questions. For ethics, we adapted multi-choice questions from professional codes (SOA, IFoA, CAA).

2.2.2 Data Processing

Raw materials underwent systematic processing:

OCR and Digitization converted PDFs to text, with math formulas transformed to LaTeX and tables to Markdown.

Structural Conversion encoded unstructured documents into standardized JSON format with context retention.

Quality Control involved anonymization and expert verification of mathematical accuracy and answer validity.

Attribute	Values
`source`	SOA, IFoA, CAA
`type`	mcq, fill_in_blank, table, code, case
`difficulty`	easy, medium, difficult
`round`	single, multi
`bloom`	Knowledge, Comprehension, Application, Analysis, Synthesis, Evaluation

Table 1: Metadata annotation schema for the INS-ActBench dataset.

2.2.3 Metadata Annotation

Each question-answer pairs is annotated with metadata (Source, Type, Difficulty, Round, Bloom's Taxonomy) to support fine-grained analysis (details in Table 1). This metadata annotation enables researchers to conduct targeted model performance analysis based on specific ability dimensions, difficulty levels, and cognitive levels. The classification criteria for metadata can be found in appendix C.

2.2.4 Data Statistics

The final dataset contains 6,514 actuarial question-answer pairs, constituting the most comprehensive benchmark in the field to date. Table 2 details the distribution across dimensions. And the construction pipeline is shown in appendix E.

2.3 Evaluation Framework and Metrics

Given the diversity of tasks, we employ optimized evaluation methods for each format:

Multiple Choice Question: We use accuracy as the primary metric, extracting options via regular expressions for binary scoring.

Fill-in-the-Blank: We employ a hierarchical mixed scoring mechanism. After preprocessing, we apply a four-layer matching strategy: Exact Match, Numeric Equivalence, Containment Match, and Token-level F1 score fallback. This balances term precision with format robustness.

Excel Tasks: Following SpreadsheetBench (Ma et al., 2024) and SheetCopilot (Li et al., 2023), we employ a three-dimensional weighted scoring system. *Completeness (20%)* assesses whether LLMs capture the correct answer cell, awarding full credit for non-empty solution cells. *Formula Correctness (40%)* evaluates logical consistency between generated formulas and reference formulas through Token overlap or Abstract Syntax Tree matching, examining cell reference accuracy and structural coherence. *Numerical Accuracy (40%)* executes the generated formulas within an Excel environment, comparing computed results against reference values with minor floating-point tolerance permitted.

Metric	LV1.1	LV1.2	LV1.3	LV2.1	LV2.2	LV2.3	LV3.1	LV3.2	LV3.3	LV3.4	LV4	Total
Total	**1307**	**507**	**844**	**763**	**369**	**1108**	**170**	**324**	**89**	**294**	**739**	**6514**
Difficulty												
easy	1185	485	243	329	122	418	63	127	0	95	0	3067
medium	122	22	601	409	130	679	101	145	0	183	572	2964
difficult	0	0	0	25	117	11	6	52	89	16	167	483
Round												
single	1307	507	844	763	369	1108	170	324	0	0	0	5392
multi	0	0	0	0	0	0	0	0	89	294	739	1122
Source												
SOA	658	493	61	163	101	513	70	84	0	0	739	2882
IFoA	0	0	0	0	0	0	80	40	89	294	0	503
CAA	649	14	783	600	268	595	20	200	0	0	0	3129
Bloom's Taxonomy												
Remember	233	13	66	145	61	164	86	310	0	23	20	1121
Understand	43	27	367	14	57	54	54	5	24	90	288	1023
Apply	947	453	363	511	201	753	1	3	59	90	103	3484
Analyze	76	7	39	90	46	108	2	0	0	38	84	490
Evaluate	6	2	8	1	4	17	27	5	2	50	213	335
Create	2	5	1	2	0	12	0	1	4	3	31	61

Table 2: Detailed statistics of the INS-ActBench dataset distributed by metrics, difficulty, round type, source institution, and Bloom's cognitive levels.

Code Tasks: R programming proficiency is evaluated using an analogous three-dimensional framework with identical weight distribution. *Completeness (20%)* verifies the presence of valid code blocks in LLMs' output. *Code Consistency (40%)* computes Token-level F1 scores between generated and reference code, measuring correctness of function names, variable identifiers, and logical structure—enabling partial credit assignment even for non-executable submissions. *Execution Score (40%)* runs generated scripts in isolated Docker containers equipped with standard actuarial packages; successful execution yields half credit, with the remainder contingent on numerical output matching reference solutions.

Case Studies: We adopt a two-stage evaluation: first, `Gemini-3-Pro-Preview` serves as an LLM-Judge to score responses based on official SOA rubrics (Zheng et al., 2023); subsequently, a panel of human experts verifies the results to correct potential biases in automated scoring.

2.4 Prompt Engineering

We designed task-specific prompt templates to ensure fairness. MCQs use a 2-shot strategy. Ethics and Regulation tasks use dynamic jurisdiction injection to activate domain knowledge. Tool tasks explicitly request specific formats (JSON with `openpyxl` code or R scripts). For Case Studies, the scoring prompt defines a six-level scale (0.00–1.00). Please refer to Appendix A for the original text of all prompts, including the LLM-as-Judge prompts in case studies.

3 Experiments and Analysis

3.1 Experimental Setup

We evaluate 13 mainstream LLMs, categorized by characteristics into two groups(Weston and Sukhbaatar, 2023) (Li et al., 2025b):

System 1 LLMs (Claude-Opus-4.5, Claude-Sonnet-4.5, GLM-4.6, Hunyuan-2.0, Qwen3-Max, Doubao-Seed-1.6, DeepSeek-v3.2, and Kimi-K2), which are evaluated in their default modes with reasoning inactive, representing standard paradigms applied in high-efficiency daily actuarial work.

System 2 LLMs (GPT-5.2, o3, Gemini-3-Pro-Preview, Grok-4, and ernie-x1.1), which enforce mandatory internal reasoning (System 2) prior to output generation.

All LLMs are proprietary and accessed via official APIs with temperature set to 1.

3.2 Overall Performance Analysis

We find that while reasoning architectures establish a clear performance advantage, they fail to reach the reliability threshold necessary for autonomous professional work. Table 3 presents the comprehensive performance of all LLMs across INS-ActBench task levels. Gemini-3-Pro-Preview leads with 81.30%, followed closely by o3, establishing the structural superiority of mandatory internal reasoning in navigating the high-density logical chains characteristic of actuarial science. Nevertheless, a critical reliability gap persists: even the state-of-the-art model fails to breach the 85% accuracy on average. In the actuarial context, a 15% error

Model	LV1.1	LV1.2	LV1.3	LV2.1	LV2.2	LV2.3	LV3.1	LV3.2	LV3.3	LV3.4	LV4	Average
Gemini-3-Pro-Preview	**97.85%**	**98.42%**	**93.23%**	**91.60%**	**93.77%**	**91.22%**	98.24%	**56.79%**	**23.06%**	44.78%	80.78%	**81.30%**
o3	97.48%	93.89%	90.02%	80.98%	87.26%	86.64%	96.47%	41.17%	8.79%	43.40%	83.67%	77.47%
Doubao-Seed-1.6	97.40%	90.53%	88.85%	81.91%	84.28%	84.57%	98.24%	29.50%	6.72%	43.07%	83.88%	76.03%
Claude-Sonnet-4.5	89.82%	71.60%	77.84%	57.67%	67.21%	78.97%	97.06%	35.19%	21.53%	39.90%	89.63%	71.44%
GLM-4.6	87.53%	89.94%	81.52%	74.31%	77.24%	82.31%	96.47%	27.48%	14.42%	43.16%	73.42%	70.77%
ernie-x1.1	94.87%	80.28%	78.67%	58.72%	73.44%	74.46%	95.29%	27.80%	6.89%	45.55%	69.42%	66.70%
Claude-Opus-4.5	66.72%	38.07%	60.07%	49.67%	46.88%	50.54%	**99.41%**	38.45%	21.76%	44.28%	82.49%	59.36%
Hunyuan-2.0	77.74%	62.72%	43.25%	35.12%	43.09%	49.10%	95.29%	21.01%	7.41%	43.67%	84.37%	57.47%
Qwen3-Max	66.26%	36.88%	60.66%	36.57%	41.73%	43.86%	95.88%	27.15%	21.34%	**45.57%**	79.40%	55.55%
Grok-4	69.63%	42.01%	53.44%	42.33%	46.07%	48.19%	97.06%	34.31%	11.50%	43.65%	73.75%	55.23%
DeepSeek-v3.2	57.23%	32.74%	50.71%	36.04%	38.21%	38.54%	95.88%	31.33%	16.39%	43.00%	83.68%	53.71%
GPT-5.2	50.50%	27.61%	42.77%	31.19%	33.33%	37.36%	96.47%	33.08%	9.44%	44.41%	**92.03%**	53.03%
Kimi-K2	43.84%	26.43%	35.07%	28.05%	29.54%	35.02%	91.76%	18.30%	7.15%	36.82%	70.69%	43.80%

Table 3: Main evaluation results of 13 LLMs on INS-ActBench. The best performance in each category is bolded. 'Average' means the average score of the four dimensions (25% each)

rate is unacceptable because numerical precision dictates solvency margins and pricing adequacy. This performance ceiling indicates that while current reasoning LLMs have mastered the "syntax" of actuarial logic, they lack the stability required for independent high-stakes computation, positioning them as risk consultants rather than deterministic executors of financial security.

The advantage of reasoning capabilities proves to be task-dependent, diminishing notably as the evaluation shifts from theoretical reasoning to practical tool execution. Examining performance across task levels reveals structural differentiation: Figure 2-A visualizes this trend: the performance gap (Reasoning Delta) is most pronounced in LV1 and LV2, where logical derivation is paramount. However, this gap significantly narrows in LV3 and LV4, suggesting that reasoning capabilities do not yet fully translate into practical tool-use proficiency. This pattern yields a critical insight: current LLMs adequately handle structured knowledge-based question answering, but exhibit fundamental deficiencies in tasks requiring precise tool manipulation. For the actuarial profession, this implies that LLMs can assist with concept explanation and preliminary analysis, while human expertise remains indispensable for core workflows such as spreadsheet modeling.

3.3 Practical Skills Task Analysis

The practical skills tasks constitute the core innovation of INS-ActBench (complete results in Table 4), We have three findings regarding the results of "Practical Skills" dimension.

While models demonstrate internalized professional ethics, significant cross-jurisdictional hallucinations in regulatory tasks expose severe geographic data biases. High scores on professional ethics tasks (LV3.1) indicate that LLMs have internalized basic professional conduct judgments and may serve as ethics consultation references for junior practitioners. However, cross-jurisdictional disparities in regulatory knowledge tasks (LV3.2), where models achieve significantly higher accuracy on SOA regulatory materials compared to the substantially lower scores recorded for the IFoA and CAA frameworks. The substantially greater accessibility of U.S. regulatory materials on the internet compared to those from the U.K. and China poses direct risk warnings for multinational insurers seeking LLM-assisted compliance support.

The dissociation between cell identification and formula precision reveals a fundamental "Knowledge-Action Gap" in spreadsheet modeling. The three-dimensional evaluation of Excel tasks yields the most instructive findings. Although LLMs perform reasonably on completeness (maximum 54.20%), indicating the ability to identify target cells, low scores on formula correctness and numerical accuracy reveal a significant gap between "knowing" and "doing." This finding carries substantial implications for actuarial practice: spreadsheet modeling, as a core tool in actuaries' daily work, will remain highly dependent on human expertise for the foreseeable future, confirming the current inability of LLMs to fully execute complex Excel tasks.

We find that high syntactic correctness does not translate to functional validity, resulting in plausible-looking code that frequently fails upon execution. Code tasks exhibit a deceptive "high completeness, low validity" pattern: while near 100% completeness suggests LLMs can fluently generate syntactic structures, the plummeting executability scores reveal a critical disconnect between linguistic fluency and functional logic. In

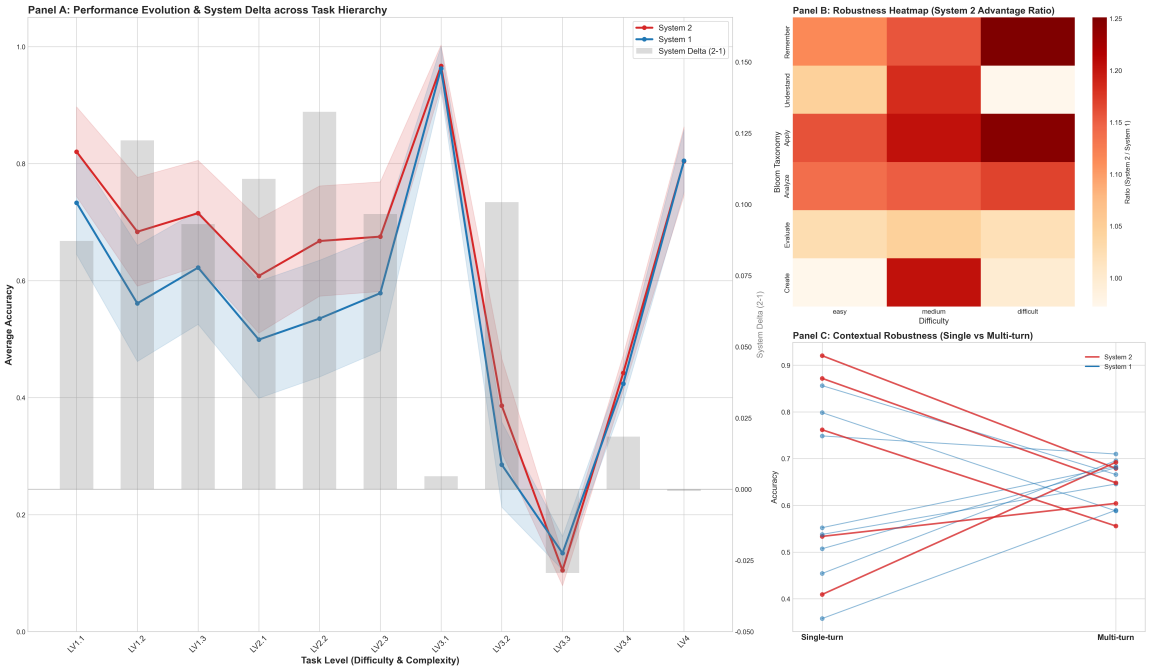


Figure 2: Macro-to-Micro Comparison of System 1 vs. System 2 LLMs on INS-ActBench Task Hierarchy.

	LV3.1 Professional Ethic			LV3.2 Regulation			LV3.3 Excel Tasks			LV3.4 Code Tasks		
Model	SOA	IFoA	CAA	SOA	IFoA	CAA	Completeness	Formula	Numerical	Completeness	Consistency	Execution
Gemini-3-Pro-Preview	98.57%	100.00%	90.00%	87.39%	34.88%	48.32%	52.10%	20.55%	11.05%	100.00%	22.83%	39.11%
o3	97.14%	100.00%	80.00%	74.93%	25.67%	30.08%	24.57%	6.29%	3.39%	100.00%	26.29%	32.22%
Doubao-Seed-1.6	98.57%	98.75%	95.00%	45.67%	13.29%	25.95%	14.09%	6.33%	3.41%	100.00%	23.54%	34.12%
Claude-Sonnet-4.5	98.57%	100.00%	80.00%	64.93%	17.54%	26.23%	54.14%	25.65%	1.11%	100.00%	16.43%	33.31%
GLM-4.6	97.14%	98.75%	85.00%	39.07%	16.24%	24.85%	36.20%	15.07%	2.88%	100.00%	21.22%	36.67%
ernie-x1.1	94.29%	100.00%	80.00%	48.03%	8.70%	23.13%	24.06%	1.49%	3.72%	100.00%	26.02%	37.86%
Claude-Opus-4.5	100.00%	100.00%	95.00%	64.12%	24.75%	30.40%	54.20%	26.61%	0.68%	100.00%	21.00%	39.70%
Hunyuan-2.0	95.71%	97.50%	85.00%	33.50%	8.67%	18.22%	19.12%	5.80%	3.18%	100.00%	22.42%	36.75%
Qwen3-Max	95.71%	97.50%	90.00%	43.22%	14.38%	22.96%	49.19%	22.35%	6.40%	100.00%	26.99%	36.94%
Grok-4	98.57%	98.75%	85.00%	56.45%	41.38%	23.60%	28.07%	11.95%	2.77%	99.32%	25.93%	33.53%
DeepSeek-v3.2	94.29%	98.75%	90.00%	57.29%	13.91%	23.91%	37.24%	21.18%	1.17%	100.00%	20.99%	36.52%
GPT-5.2	97.14%	100.00%	80.00%	56.85%	17.30%	26.26%	22.40%	9.76%	2.64%	100.00%	22.88%	38.15%
Kimi-K2	90.00%	96.25%	80.00%	22.41%	8.09%	18.61%	20.13%	7.22%	0.59%	100.00%	22.86%	19.19%

Table 4: LV3 Detailed Evaluation Results

the context of actuarial modeling, this creates a dangerous debugging trap. Models often hallucinate package dependencies or misapply statistical arguments, producing scripts that appear plausible but yield fallacious risk projections. Consequently, for professional practitioners, the utility of LLMs in programming is currently confined to that of a syntax retrieval assistant, rather than an autonomous model architect, as the cost of verifying and rectifying generated logic currently outweighs the efficiency gains of automation.

3.4 Metadata Annotation Analysis

Fine-grained metadata diagnosis reveals a structural imbalance in current LLM capabilities: models exhibit "static analytical strength" yet suffer from "dynamic and foundational fragility." We conduct analysis across four dimensions in each metadata annotation (complete results in Table 5). While reasoning-enhanced architectures demonstrate resilience against increasing difficulty, the robustness heatmap in Figure 2-B localizes their primary advantage within the high-order 'Apply' and 'Analyze' dimensions of Bloom's Taxonomy, confirming their utility in structured logical deduction. However, this proficiency proves brittle. Stemming from calculation deficits, the simultaneous underperformance in foundational 'Remember' tasks and the significant erosion of reasoning superiority depicted in Figure 2-C (where the performance gap between System 1 and System 2 models narrows markedly in dynamic contexts) collectively indicate a failure to maintain coherence across long horizons. Consequently, for the actuarial profession, LLMs are currently viable only as modular assistants for discrete, well-defined sub-tasks; the

	Difficulty			Bloom						Source			Round	
Model	Easy	Medium	Difficult	Remember	Understand	Apply	Analyze	Evaluate	Create	SOA	IFoA	CAA	Single	Multi
Gemini-3-Pro-Preview	74.68%	70.04%	49.99%	65.23%	85.35%	36.90%	100.00%	92.10%	37.81%	92.47%	67.54%	52.11%	71.05%	23.06%
o3	88.03%	85.69%	85.53%	68.86%	79.61%	94.53%	96.47%	86.51%	59.29%	95.78%	51.19%	80.09%	88.71%	8.79%
Doubao-Seed-1.6	82.59%	82.52%	81.92%	62.08%	80.58%	94.66%	92.31%	87.81%	43.95%	94.13%	47.06%	79.04%	85.26%	6.72%
Claude-Sonnet-4.5	74.80%	74.49%	70.18%	64.16%	79.26%	74.72%	75.77%	81.59%	65.12%	77.09%	58.16%	71.72%	73.79%	21.53%
GLM-4.6	51.16%	56.38%	21.98%	42.58%	73.55%	16.29%	100.00%	80.65%	33.65%	65.47%	52.73%	30.32%	51.22%	14.42%
Claude-Opus-4.5	62.56%	55.70%	46.39%	60.26%	63.55%	50.60%	59.51%	75.08%	58.91%	48.80%	62.39%	58.86%	55.22%	21.76%
Hunyuan-2.0	50.74%	53.25%	55.53%	45.46%	49.10%	57.25%	44.29%	72.28%	30.23%	68.60%	45.02%	43.10%	53.78%	7.41%
Qwen3-Max	57.56%	49.67%	44.41%	52.78%	65.40%	46.11%	54.08%	71.77%	60.88%	46.40%	63.17%	52.95%	50.72%	21.34%
Grok-4	70.86%	63.11%	58.62%	57.68%	77.49%	66.45%	65.38%	83.16%	44.14%	67.42%	59.59%	62.51%	65.86%	11.50%
DeepSeek-v3.2	53.90%	44.92%	36.74%	50.39%	55.29%	41.07%	44.84%	73.66%	48.33%	39.67%	61.15%	48.41%	45.43%	16.39%
GPT-5.2	47.85%	41.24%	32.21%	49.64%	48.59%	35.52%	39.67%	62.66%	37.22%	37.13%	47.24%	42.33%	40.93%	9.44%
Kimi-K2	41.49%	35.60%	29.79%	40.48%	41.78%	32.12%	35.60%	71.65%	45.57%	33.78%	61.85%	35.96%	35.78%	7.15%

Table 5: Evaluation Results across Metadata Annotation

inability to integrate evolving information or guarantee cross-turn consistency reinforces that human expertise remains the irreplaceable "context anchor" in complex, continuous actuarial consulting.

3.5 Comparison with General Mathematical Benchmarks

General mathematical proficiency could not linearly translate to professional actuarial competence. Unlike results on classical mathematical reasoning benchmarks such as GSM8K (Cobbe et al., 2021) and MATH (Hendrycks et al., 2021), where OpenAI models have long maintained leadership, INS-ActBench reveals a different ranking: GPT models perform unremarkably while Gemini and o3 rank near the top. This deviation reflects the nonlinear relationship between general mathematical ability and domain expertise. The contextual complexity of actuarial tasks demands not only computational capability but also cross-domain knowledge integration and professional judgment. This comparison provides experiment evidence for the selection of actuarial LLMs: strong performance on general benchmarks cannot be simply extrapolated to professional scenarios, making domain-specific evaluation indispensable.

4 Conclusions

This study presents INS-ActBench, the first comprehensive benchmark that advances the evaluation frontier from finance or insurance knowledge to actuarial capability. By integrating 6,514 authentic questions from the SOA, IFoA and CAA examination systems, we operationalize human professional certification standards into quantifiable LLM evaluation metrics, enabling systematic assessment across the full spectrum of actuarial competency.

Our experiment results yield a central insight with direct implications for actuarial practice: current LLMs exhibit a pronounced "strong theory, weak practice" characteristic that fundamentally shapes their deployment potential. High performance in foundational knowledge and professional ethics tasks confirms that LLMs have internalized substantial knowledge of the actuarial domain and can reliably support knowledge retrieval and ethical guidance queries. However, the sharp performance degradation on practical tasks, reveals that the bottleneck lies not in knowledge acquisition but in the translation from knowing to doing. This gap cannot be bridged through system 2 reasoning alone; it reflects architectural limitations in tool operation semantics, spatial dependency reasoning, and cross-turn state maintenance.

These findings carry concrete guidance for actuarial digital transformation. In the near term, LLMs are best positioned as cognitive assistants rather than autonomous agents: supporting junior actuaries with terminology lookup, preliminary draft generation, and ethics consultation, while reserving spreadsheet modeling, regulatory interpretation, and client advisory for human experts. The cross-jurisdictional performance disparity further cautions multinational insurers against extrapolating U.S.-centric model capabilities to other regulatory environments without domain-specific validation. Looking forward, INS-ActBench establishes the experiment foundation for tracking LLM progress in this specialized domain, identifying targeted capability gaps for model developers, and informing curriculum evolution as actuarial education adapts to an AI-augmented professional landscape.

Limitations

First, although INS-ActBench adopts a multi-tier hierarchy, the current tasks are primarily based on static exam materials and case studies, which may not fully capture the dynamic, long-horizon nature of real-world actuarial consulting. Second, our tool-use evaluation is limited to Excel and R programming; proprietary actuarial software widely used in the industry is not yet integrated due to licensing restrictions. Third, while we cover three major jurisdictions, the benchmark's findings may not fully generalize to all global regulatory frameworks with distinct local requirements.

Ethical considerations

The authors ensure that INS-ActBench is developed using publicly available professional materials and does not contain sensitive personal or proprietary corporate data. The dataset is intended for research and educational purposes only. We emphasize that LLMs performance on this benchmark does not constitute professional actuarial advice or a substitute for certified actuarial judgment. The authors and their institutions disclaim liability for any financial decisions or regulatory filings resulting from the use of LLMs evaluated by this Material.

References

Philippe Artzner, Freddy Delbaen, Jean-Marc Eber, and David Heath. 1999. Coherent measures of risk. *Mathematical finance*, 9(3):203–228.

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

Benjamin S Bloom, Max D Engelhart, Edward J Furst, Walker H Hill, David R Krathwohl, and 1 others. 1956. *Taxonomy of educational objectives: The classification of educational goals. Handbook 1: Cognitive domain*. Longman New York.

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*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, and 1 others. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

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*.

Minwei Feng, Bing Xiang, Michael R Glass, Lidan Wang, and Bowen Zhou. 2015. Applying deep learning to answer selection: A study and an open task. In *2015 IEEE workshop on automatic speech recognition and understanding (ASRU)*, pages 813–820. IEEE.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.

Haohang Li, Yupeng Cao, Yangyang Yu, Shashidhar Reddy Javaji, Zhiyang Deng, Yueru He, Yuechen Jiang, Zining Zhu, Kp Subbalakshmi, Jimin Huang, and 1 others. 2025a. 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.

Hongxin Li, Jingran Su, Yuntao Chen, Qing Li, and Zhao-Xiang Zhang. 2023. Sheetcopilot: Bringing software productivity to the next level through large language models. volume 36, pages 4952–4984.

Zhong-Zhi Li, Duzhen Zhang, Ming-Liang Zhang, Jiaxin Zhang, Zengyan Liu, Yuxuan Yao, Haotian Xu, Junhao Zheng, Pei-Jie Wang, Xiuyi Chen, and 1 others. 2025b. From system 1 to system 2: A survey of reasoning large language models. *arXiv preprint arXiv:2502.17419*.

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.

Zeyao Ma, Bohan Zhang, Jing Zhang, Jifan Yu, Xiaokang Zhang, Xiaohan Zhang, Sijia Luo, Xi Wang, and Jie Tang. 2024. Spreadsheetbench: Towards challenging real world spreadsheet manipulation. volume 37, pages 94871–94908.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. 2025. A comprehensive overview of large language models. *ACM Transactions on Intelligent Systems and Technology*, 16(5):1–72.

Jason Weston and Sainbayar Sukhbaatar. 2023. System 2 attention (is something you might need too). *arXiv preprint arXiv:2311.11829*.

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. volume 36, pages 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.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

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 Prompt Engineering

The prompt engineering strategy for INS-ActBench is designed to elicit domain-specific expertise while maintaining cross-model fairness. To ensure the activation of localized professional knowledge, we implement a **Dynamic System Prompt** mechanism for Ethics and Regulation tasks. This approach automatically injects jurisdictional roles—such as "Role: CAA Actuary"—based on the question's origin, forcing LLMs to adhere to specific legal and professional codes. For Multiple-Choice Questions, a standardized two-shot template is utilized to stabilize output formats, ensuring that each LLM can follow the rigid single-letter response requirement.

Furthermore, for tool-intensive tasks (Excel modeling and R programming), the prompts incorporate strict execution constraints and environment specifications. These instructions explicitly mandate the use of particular libraries (e.g., *openpyxl*) and define structured JSON outputs to facilitate automated verification in isolated Docker environments. Additionally, this section provides the specialized scoring prompts used for the "LLM-as-a-Judge" framework in Level 4 Case Studies. These evaluation prompts translate complex, open-ended actuarial rubrics into actionable grading criteria, enabling the judge model to provide objective scores across dimensions like logical consistency and strategic defensibility. The details are shown in Figure 3 and Figure 4.

B Excel Examples

This appendix details the experimental setup for the Excel modeling task. We provide the LLM with the task description shown in Figure 5, a comprehensive background introduction, and the output templates for Figure 6. The LLMs are required to process these inputs and generate the corresponding content for Figure 6 exclusively using Excel formulas. This procedure evaluates the model's proficiency in translating complex actuarial logic into functional spreadsheet operations.

C Classification Criteria for Metadata

It is easy to understand the classification criteria for rounds and sources. We will further introduce two other metadata. Traditional difficulty metrics designed for human test-takers do not directly transfer to LLM evaluation. Prior research has demonstrated that the cognitive challenges faced by humans and LLMs diverge substantially: complex calculus operations that challenge human working memory may be straightforward for LLMs trained on mathematical corpora, while tasks requiring multi-step logical inference or implicit reasoning—relatively intuitive for humans—often expose fundamental limitations in current architectures.

To address this asymmetry, we adopt a Context-Noise Approach that calibrates difficulty based on the information structure of input prompts, specifically targeting LLMs' long-context attention capabilities. We define three levels: Easy (Clean Context) items present complete, well-structured information without extraneous content; Medium (Noisy Context) items embed approximately 20% irrelevant background information or include outliers requiring identification and exclusion; Difficult (Messy Context) items present unstructured long-form text with critical values dispersed throughout, potentially containing superficially contradictory information that demands model-driven disambiguation. Difficulty were as-

signed by Gemini-3-Pro-Preview through batch classification based on these criteria.

Bloom's Taxonomy provides a hierarchical framework of cognitive objectives comprising six levels: Remember (recall of facts and terminology), Understand (interpretation and paraphrasing), Apply (use of knowledge in novel contexts), Analyze (decomposition and relationship identification), Evaluate (judgment based on criteria), and Create (synthesis of elements into new solutions).

Incorporating Bloom's levels into INS-ActBench serves two purposes. First, it provides an analytical dimension orthogonal to task type—within multiple-choice questions, a terminology recall item and a strategic judgment item impose fundamentally different cognitive demands. Second, Bloom's taxonomy enables diagnosis of capability boundaries: whether LLMs excel only at lower-order pattern matching or demonstrate higher-order synthesis and judgment. Bloom's level annotations were assigned by domain experts during manual quality verification.

D MCQ Q&A Pair Example

Figure 7 presents a typical MCQ data sample in INS-ActBench. This example visually presents the structured preservation of complex life-insurance formulas with a multidimensional metadata annotation.

E Construction Pipeline

Figure 8 illustrates the comprehensive data construction pipeline of INS-ActBench. It visualizes the transformation from raw sources through a rigorous three-stage workflow. The process culminates in a structured four-tier benchmark covering diverse task types.

System Prompts for INS-ActBench

1. Multiple-Choice Questions (MCQ)
Answer the multiple-choice question.
Output format rules:
- Respond with exactly ONE letter: A, B, C, D, or E
- Do not output any other text.

---

2. Professional Ethics (Moral)
You are an expert actuary specializing in professional conduct and ethics. The question belongs to the jurisdiction: `{source}`. You must apply the specific Code of Professional Conduct of `{source}` to answer.
Output format rules:
- Respond with exactly ONE letter: A, B, C, D, or E
- Do not output any other text.

---

3. Regulation Compliance
You are an expert actuarial regulatory consultant specializing in `{source}`. Your task is to provide the precise answer based on the official regulatory documents.
Output format rules:
- For Fill-in-the-Blank: Output ONLY the missing term.
- For Source Identification: Output the Document Name and Section/Article.
- Do NOT provide any reasoning or extra text.

---

4. Excel Modeling Task
You are an actuarial analyst. Write Python code using `openpyxl` to fill ONLY the required cells.
CRITICAL output format (JSON):
- `explanation`: short text
- `python_code`: single Python script string

CRITICAL code rules:
- Use `openpyxl` only. Read `TEMPLATE_PATH`, write `OUTPUT_PATH`.
- Only modify cells within `ANSWER_POSITION`. Do NOT change sheet names.
- Your code MUST run in the runner environment.

---

5. R Programming Task
You are an expert actuarial analyst. Write R code to solve the problem and provide numerical results.
Output format rules:
- Provide R code in: “‘`R ...` “‘
- REQUIRED: End with section ‘### Output Summary’ containing final values.

---

6. Comprehensive Case Study
You are an expert actuarial analyst. Provide a concise, well-structured answer based on the case background.

Figure 3: Detailed System Prompts used for different tasks in INS-ActBench.

Scoring Prompt for Automatic Evaluation

Role:
You are an expert actuarial examiner scoring AI responses against reference answers.

Task:
Evaluate each `model_output` against its `reference_answer`. Output a score from 0.00 to 1.00.

Scoring Guidelines:

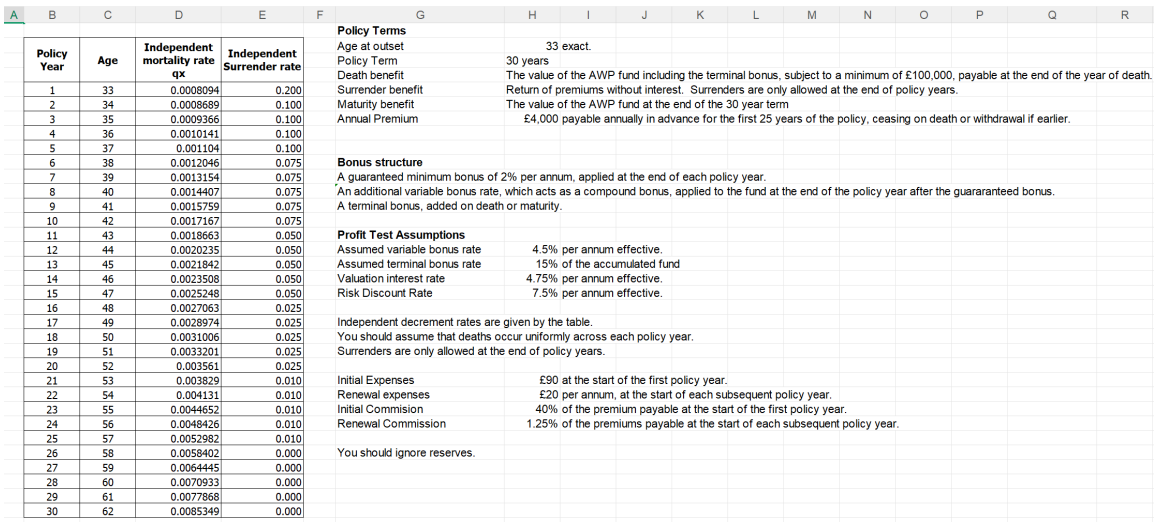
Score Range	Criteria
0.90–1.00	Correct answer with complete reasoning; matches reference key points
0.70–0.89	Core answer correct; minor omissions or deviations
0.50–0.69	Partially correct; understands concept but incomplete execution
0.30–0.49	Relevant attempt; significant errors or missing key elements
0.10–0.29	Minimal relevance; fundamental gaps
0.00–0.09	Wrong, refuses to answer, or claims "insufficient data" when data exists

Key Rules:
1. Compare ONLY to `reference_answer` content.
2. "I need more information" when info is provided = 0.00–0.10.
3. Correct method + wrong numbers $>$ wrong method.
4. Cover all key points in reference = high score.
5. Extra correct content beyond reference = no penalty.
6. Ignore formatting differences.

Output Format:
Return ONLY a JSON array, no other text:

```
[
   {"question_num": 1, "score": 0.00},
   {"question_num": 2, "score": 0.65},
   ...
   {"question_num": 739, "score": 0.78}
]
```

Figure 4: The scoring prompt used for LLM-as-a-Judge evaluation in Case Study tasks.



Policy Year	Age	Independent mortality rate qx	Independent Surrender rate
1	33	0.0008094	0.200
2	34	0.0008689	0.100
3	35	0.0009366	0.100
4	36	0.0010141	0.100
5	37	0.001104	0.100
6	38	0.0012046	0.075
7	39	0.0013154	0.075
8	40	0.0014407	0.075
9	41	0.0015759	0.075
10	42	0.0017167	0.075
11	43	0.0018663	0.050
12	44	0.0020235	0.050
13	45	0.0021842	0.050
14	46	0.0023508	0.050
15	47	0.0025248	0.050
16	48	0.0027063	0.025
17	49	0.0028974	0.025
18	50	0.0031006	0.025
19	51	0.0033201	0.025
20	52	0.003561	0.025
21	53	0.003829	0.010
22	54	0.004131	0.010
23	55	0.0044652	0.010
24	56	0.0048426	0.010
25	57	0.0052982	0.010
26	58	0.0058402	0.000
27	59	0.0064445	0.000
28	60	0.0070933	0.000
29	61	0.0077868	0.000
30	62	0.0085349	0.000

Policy Terms

Age at outset	33 exact.
Policy Term	30 years
Death benefit	The value of the AWP fund including the terminal bonus, subject to a minimum of £100,000, payable at the end of the year of death.
Surrender benefit	Return of premiums without interest. Surrenders are only allowed at the end of policy years.
Maturity benefit	The value of the AWP fund at the end of the 30 year term
Annual Premium	£4,000 payable annually in advance for the first 25 years of the policy, ceasing on death or withdrawal if earlier.

Bonus structure

A guaranteed minimum bonus of 2% per annum, applied at the end of each policy year.
An additional variable bonus rate, which acts as a compound bonus, applied to the fund at the end of the policy year after the guararanteed bonus.
A terminal bonus, added on death or maturity.

Profit Test Assumptions

Assumed variable bonus rate	4.5% per annum effective.
Assumed terminal bonus rate	15% of the accumulated fund
Valuation interest rate	4.75% per annum effective.
Risk Discount Rate	7.5% per annum effective.

Independent decrement rates are given by the table.
You should assume that deaths occur uniformly across each policy year.
Surrenders are only allowed at the end of policy years.

Initial Expenses	£90 at the start of the first policy year.
Renewal expenses	£20 per annum, at the start of each subsequent policy year.
Initial Commision	40% of the premium payable at the start of the first policy year.
Renewal Commission	1.25% of the premiums payable at the start of each subsequent policy year.

You should ignore reserves.

Figure 5: Example of Excel task.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
			[1]	[1]	[1]	[1]	[1]	[1]	[1]	[1]				
		WITH-PROFITS FUND		£4,000	2%	4.5%	15%	100,000						
	Age	Policy Year	Fund at start	Premium	Fund after Guaranteed Bonus	Fund after Variable Bonus	Fund after Terminal Bonus	Death Benefit	Surrender Benefit	Maturity Benefit				
33	33	1	-	4,000.00	4,080.00	4,263.60	4,903.14	100,000.00	4,000.00				Fund at start	1
34	34	2	4,263.60	4,000.00	8,428.87	8,808.17	10,129.40	100,000.00	8,000.00				Premium	1
35	35	3	8,808.17	4,000.00	13,064.33	13,652.23	15,700.06	100,000.00	12,000.00				Fund after Guaranteed Bonus	1
36	36	4	13,652.23	4,000.00	18,005.27	18,815.51	21,637.84	100,000.00	16,000.00				Fund after Variable Bonus	1
37	37	5	18,815.51	4,000.00	23,271.82	24,319.05	27,966.91	100,000.00	20,000.00				Fund after Terminal Bonus	1
38	38	6	24,319.05	4,000.00	28,885.43	30,185.28	34,713.07	100,000.00	24,000.00				Death Benefit	1
39	39	7	30,185.28	4,000.00	34,868.99	36,438.09	41,903.80	100,000.00	28,000.00				Surrender Benefit	1
40	40	8	36,438.09	4,000.00	41,246.85	43,102.96	49,568.40	100,000.00	32,000.00				Maturity Benefit	1
41	41	9	43,102.96	4,000.00	48,045.02	50,207.04	57,738.10	100,000.00	36,000.00					8
42	42	10	50,207.04	4,000.00	55,291.19	57,779.29	66,446.18	100,000.00	40,000.00					
43	43	11	57,779.29	4,000.00	63,014.87	65,850.54	75,728.13	100,000.00	44,000.00					
44	44	12	65,850.54	4,000.00	71,247.55	74,453.69	85,621.75	100,000.00	48,000.00					
45	45	13	74,453.69	4,000.00	80,022.77	83,623.79	96,167.36	100,000.00	52,000.00					
46	46	14	83,623.79	4,000.00	89,376.27	93,398.20	107,407.93	107,407.93	56,000.00					
47	47	15	93,398.20	4,000.00	99,346.17	103,816.74	119,389.25	119,389.25	60,000.00					
48	48	16	103,816.74	4,000.00	109,973.08	114,921.87	132,160.15	132,160.15	64,000.00					
49	49	17	114,921.87	4,000.00	121,300.30	126,758.82	145,772.64	145,772.64	68,000.00					
50	50	18	126,758.82	4,000.00	133,373.99	139,375.82	160,282.20	160,282.20	72,000.00					
51	51	19	139,375.82	4,000.00	146,243.34	152,824.29	175,747.93	175,747.93	76,000.00					
52	52	20	152,824.29	4,000.00	159,960.78	167,159.01	192,232.86	192,232.86	80,000.00					
53	53	21	167,159.01	4,000.00	174,582.19	182,438.39	209,804.15	209,804.15	84,000.00					
54	54	22	182,438.39	4,000.00	190,167.16	198,724.68	228,533.38	228,533.38	88,000.00					
55	55	23	198,724.68	4,000.00	206,779.17	216,084.24	248,496.87	248,496.87	92,000.00					
56	56	24	216,084.24	4,000.00	224,485.92	234,587.79	269,775.95	269,775.95	96,000.00					
57	57	25	234,587.79	4,000.00	243,359.54	254,310.72	292,457.33	292,457.33	100,000.00					
58	58	26	254,310.72	-	259,396.94	271,069.80	311,730.27	311,730.27	100,000.00					
59	59	27	271,069.80	-	276,491.19	288,933.30	332,273.29	332,273.29	100,000.00					
60	60	28	288,933.30	-	294,711.96	307,974.00	354,170.10	354,170.10	100,000.00					
61	61	29	307,974.00	-	314,133.48	328,269.49	377,509.91	377,509.91	100,000.00					
62	62	30	328,269.49	-	334,834.88	349,902.45	402,387.82	402,387.82	100,000.00	402,387.82				

Figure 6: Example of Excel answer.

Sample Data Instance: Annotated Multiple-Choice Question

Metadata Annotation:

Source: SOA	**Difficulty:** Medium	**Type:** MCQ	**Question Num:** 58
Round: Single	**Bloom:** Apply	**Level:** LV2.1 (Life Insurance)	

Input 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:

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

(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

Ground Truth (Reference Answer): A

Figure 7: A representative example of a Level 2 (Core Actuarial Models) question from the INS-ActBench dataset.

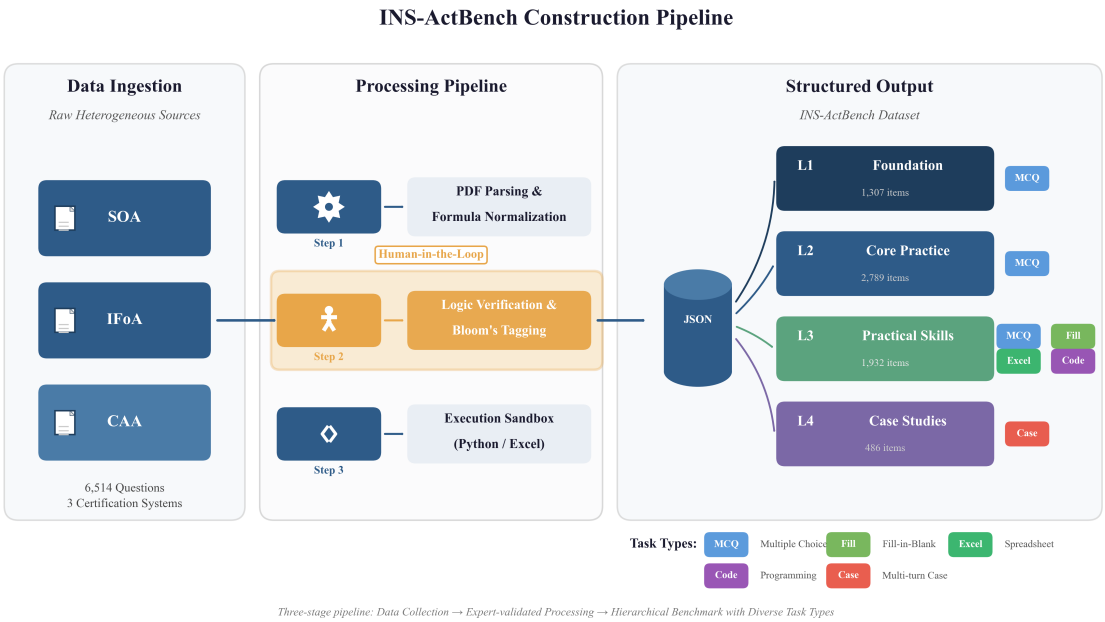


Figure 8: INS-ActBench Construction Pipeline