

Answers to Frequently Asked Questions

2026 Request for Proposals: Evaluating LLMs for a P&C Actuarial Task Benchmark and Re-Evaluation Suite

Please note that the period for submitting questions has closed. This is the complete set of answers the CAS will provide.

Proposal Eligibility and Team Structure

1.1 Eligibility of individuals and entities

May an unaffiliated individual apply and contract directly with CAS? Are there entity, jurisdiction, onboarding, payment-documentation, or milestone-payment requirements?

CAS Guidance: An individual may apply and contract directly with CAS from any location or jurisdiction. Payment paperwork will be prepared after award; milestone payments are made in three equal allotments: the first at contract signing, the second upon first draft submission; and the final upon CAS approval of the final draft paper/tool.

1.2 Lead and actuarial credentials

May a technical or non-actuary lead submit with qualified actuarial collaborators, or must the lead hold an actuarial credential? What minimum actuarial qualifications and responsibilities must the team cover?

CAS Guidance: The lead does not need an actuarial credential and may submit with actuarial collaborators. There are no minimum actuarial qualifications.

1.3 Named actuarial support at submission

Must a credentialed actuarial subject matter expert (SME) or reviewer be named, committed, and accompanied by a CV at proposal submission, or may that appointment be finalized during shortlisting, before award, or before task and label acceptance?

CAS Guidance: Because bidders are not required to include a credentialed actuary on their teams, they can add an actuarial SME at any point in the process.

1.4 Joint proposals and subcontractors

Are joint proposals and named specialist subcontractors or collaborators acceptable? If so, what responsibilities, qualifications, budget detail, and contracting/signatory structure are required?

CAS Guidance: Yes, joint proposals are acceptable. Each partner should submit a CV and budget details for their portion of the work as well as a consolidated budget. Payments can be split among the partners. Each partner must sign the contract. If there is a prime contractor working with subcontractors, the prime must sign the contract and be the responsible party.

1.5 Use of CAS designated SMEs

May a proposal rely on CAS-designated actuarial SMEs to validate task design and scoring? What availability, turnaround, scope, and cost assumptions should bidders use, and does this validation replace any bidder responsibility for task drafting, answers, or result interpretation?

CAS Guidance: CAS will supply volunteer reviewers and CAS-designated actuarial SMEs will validate task design and scoring. The precise availability, turnaround, and division of responsibility will be decided after award.

1.6 AI as primary author or operating model

Does use of AI as the operating model or primary author, rather than as an incidental tool, disqualify a bidder?

CAS Guidance: No, it does not. However, there must also be a named human author or institution/organization that is ultimately responsible for the content of the resulting paper and tool.

Benchmark Scope and Evaluation Methodology

2.1 Initial benchmark scope and task count

Would an initial release of twelve well-justified sub-tasks spanning the principal P&C task families be a reasonable scope? Are there minimum task, case, or dataset counts?

CAS Guidance: Yes, the twelve subtasks would work. There is no minimum task, case, or dataset count. As long as the proposal is able to convince the reviewers that it will be able to effectively benchmark the datasets as laid out in the RFP, there are no hard minimums.

2.2 Task definitions and ground truth

What actuarial datasets, task definitions, SMEs, and ground-truth labels will CAS provide? May CAS SMEs define and validate labels and gold answers while the selected team builds the benchmark, scoring, and platform? What work must the selected team perform for acquisition, simulation, labeling, validation, and rights clearance?

CAS Guidance: CAS will not provide datasets or task definitions. CAS-designated SMEs will be available to validate task design and scoring. Researchers must perform all rights clearances and attest that they have obtained proper permissions and included citations.

2.3 Simulated and synthetic data

Where legally publishable real P&C data are unavailable, is rigorously generated or simulated data acceptable? Is there a preferred proportion or limitation, and are published generators plus validation, limitations, and independent actuarial review sufficient?

CAS Guidance: Simulated data is acceptable. There is no limitation on the proportion of data that is simulated. The RFP reviewers will be evaluating whether the benchmarking dataset will achieve the desired benchmarks, so ideally there should be information regarding the rigor of the simulated dataset as well as some real data for further confidence and comparison.

2.4 Model panel and open weight models

How many models and which families should be included at first publication? Is there a required commercial/open-weight split, and how should the open-source requirement be interpreted for restricted-weight releases?

CAS Guidance: Evaluate three or more open-weight models; hosted inference for open-weight families is acceptable. There should be evaluations of commercial closed-weight models from the major labs (OpenAI, Anthropic, Google), testing a variety of the most recently released families.

2.5 Evaluation configuration

Should all models be evaluated in one standardized configuration with extended reasoning modes and tool use enabled, or should separate configurations be reported?

CAS Guidance: Use a single standardized configuration with extended reasoning modes and tool use enabled.

2.6 Re-testing new models

When a new model is released, what level of effort should adding and scoring it require? How frequently should models be re-evaluated, and does CAS expect a configuration-driven, one-command workflow?

CAS Guidance: The CAS is thinking of re-evaluating models every 6 months. A configuration-driven, one-command workflow would be great. It is acceptable, however, if the evaluation requires additional work beyond a one-command workflow. Ideally, the team will work with the CAS to update the benchmarking twice, so that the CAS volunteers learn how to do it themselves. Please include the cost of helping the CAS with 2 re-evaluations of the benchmarking.

2.7 Reproducibility standard

Should scoring be deterministic and reproducible from saved model outputs so published results can be independently reproduced without re-querying models?

CAS Guidance: Exact model outputs need not be deterministic. However, the benchmark framework, evaluation methodology, prompts, datasets, and scoring procedures should be sufficiently documented so that CAS can repeat the evaluation process and obtain comparable results in the future.

2.8 Private holdout and contamination

Should the benchmark retain a private holdout split whose labels are never publicly released, with CAS running it in-house for headline scores, to reduce training-data contamination?

CAS Guidance: Evaluation datasets should remain private to CAS, through a private GitHub repository or encrypted files in the public repository. The public GitHub should contain everything else.

Data Licensing and Publication

3.1 Third-party data rights

What process should the selected team use for third-party dataset licenses, notices, and usage agreements? Who confirms redistribution rights, including for CAS Schedule P files?

CAS Guidance: The selected team must adhere to the usage licenses and agreements for each dataset. The CAS will manage the process to request any CAS Schedule P file redistribution rights as needed.

3.2 Dataset release model

What license should apply to benchmark datasets, and is public release of the full task set expected given the private-holdout requirement?

CAS Guidance: Evaluation datasets remain private to CAS; other materials are intended for the public GitHub. The specific dataset license and exact public/private boundaries can be confirmed after award.

Platform Hosting Maintenance and Deliverables

4.1 Platform architecture and hosting

May a static accessible site generated from versioned results and deployable from the CAS GitHub repository satisfy the public comparison platform requirement, or is managed hosting required? Should hosting be CAS-controlled, and are there specific cloud, deployment, or dependency constraints?

CAS Guidance: Propose maintenance and re-test costs separately. CAS wants the ability to re-run and later edit the benchmarking suite independently. The hosting model is otherwise open.

4.2 Dynamic updates without live execution

Would a reviewed-results site that is rebuilt when validated result submissions are accepted satisfy the dynamic-update requirement without live public model execution?

CAS Guidance: Yes, a reviewed-results site is the goal. The CAS does not want live public model execution ability.

4.3 Maintenance period and update cycles

For how long must the selected team host and operate the platform, and how many ongoing update cycles are expected?

CAS Guidance: Estimate the cost of 2 update cycles and provide it in a separate budget from the main project budget. CAS may bear that cost outside the award.

4.4 Required publication content

Does full publication to CAS GitHub require all platform source code and deployment infrastructure, along with evaluation scripts, scoring scripts, generated benchmark content, and datasets?

CAS Guidance: CAS seeks rights to re-run and later edit the suite independently. Project-specific deliverables are expected to be available.

Intellectual Property and Licensing

5.1 Background IP and reusable tools

May a proposer retain ownership of listed pre-existing generic benchmarking or evaluation infrastructure and know-how while giving CAS the rights needed to run, modify, and maintain the delivered project independently?

CAS Guidance: Yes, proposers may retain pre-existing background technology and general-purpose know-how while assigning CAS rights in project-specific reports, benchmarks, datasets, code, and bespoke deliverables. Pre-existing infrastructure will be considered case by case.

5.2 Rights and license boundaries

How will rights in the report and work product relate to delivered code, data, platform source, benchmark definitions, third-party dependencies, and reusable vendor components? What licenses, including MPL 2.0, apply to each?

CAS Guidance: Pre-existing vendor components may remain with their providers without CAS having full access, provided CAS can independently re-run and later edit the suite. Exact boundaries will be stated in the final contract agreement.

Budget Schedule and Payments

6.1 Proposal budget and model API costs

Should all commercial-model API and token costs be included and itemized in the proposal budget, or are any costs reimbursed separately or covered through CAS accounts?

CAS Guidance: Include and itemize model API/token costs in the proposal budget. CAS does not provide commercial-model accounts or credits.

6.2 Award ceiling currency and travel

What is the anticipated contract budget or compensation ceiling and currency? Must U.S. presentation travel be included within that ceiling?

CAS Guidance: The compensation ceiling is \$75,000 USD. If the vendor anticipates traveling to the U.S. to present the work at a CAS conference or other event, travel costs should be included within that ceiling. Virtual presentations are also acceptable, however.

6.3 Kickoff dates milestones and payment

What start and completion dates should applicants assume, and what milestone or acceptance schedule governs payment?

CAS Guidance: Kickoff is usually about one month after the winner is informed. Payments are made in three equal parts at contract signing, upon submission of the first draft, and upon approval of the final draft.

6.4 Executive summary and presentation timing

Does the November 23, 2026 executive summary describe agreed methodology and early pilot findings, rather than completed results? What are the expected presentation timing and format?

CAS Guidance: The executive summary is a preliminary document describing planned work and early or expected findings. A remote presentation is acceptable; timing is open.

6.5 Interim reviews and deliverable dates

What intermediate checkpoints, draft-paper dates, reviews, and acceptance milestones are expected before the June 2027 final deliverable?

CAS Guidance: Authors will meet regularly with a CAS Project Oversight Group and submit drafts for review and revision at regular intervals—usually monthly—until a final paper is approved by the CAS for publication.

Procurement and Contract Process

7.1 Proposal evaluation and selection

What proposal evaluation criteria and relative weighting will CAS use, including actuarial credentials, benchmark experience, platform capability, price, interviews, prototypes, or demonstrations?

CAS Guidance: There is no strict hierarchy among the criteria listed in the RFP; a demonstrated record of delivering projects is helpful.

7.2 Research agreement

Can CAS share the standard or draft research agreement before award, including explicit treatment of background IP and deliverable rights?

CAS Guidance: The CAS can share a generic or standard agreement.

International Access and Presentation

8.1 Commercial model access for Russia based researchers

Is there a provider-approved arrangement for the required commercial-model evaluations involving a Russia-based researcher, given provider location restrictions?

CAS Guidance: No, the CAS does not have the ability to provide models in restricted locations.