Can Gen AI Review a Pricing GLM?

4/28/2026

Roberto Perez, FCAS, CPCU, MAAA
Sam Kloese, ACAS, CSPA, MAAA

NAIC NATIONAL ASSOCIATION OF INSURANCE COMMISSIONERS

Agenda

- Background Information
- Introduction to the Sample Filing
- Gen AI Experiment
- Conclusions
- Appendix

Background Information

NAIC Model Review Team

- State regulators can refer P&C rate models to us
  - State DOI must sign Rate Support Agreement
    - Guarantees confidentiality
  - Services are free to state insurance regulators
  - We assist with the review, but do not make the final decision
  - Reports made available to other states who also signed agreement
  - NAIC requires that filing contains comprehensive documentation

NAIC Model Review Team

- NAIC Model Review Manual*
  - Rate Review Support Services Agreement (page 8)
  - Regulatory Review of Predictive Models White Paper (page 12)
  - Generalized Linear Model Checklist (page 143)

*https://content.naic.org/sites/default/files/inline-files/NAIC%20Model%20Review%20Manual__%20adopted%20by%20CASTF%2011.04.25.pdf

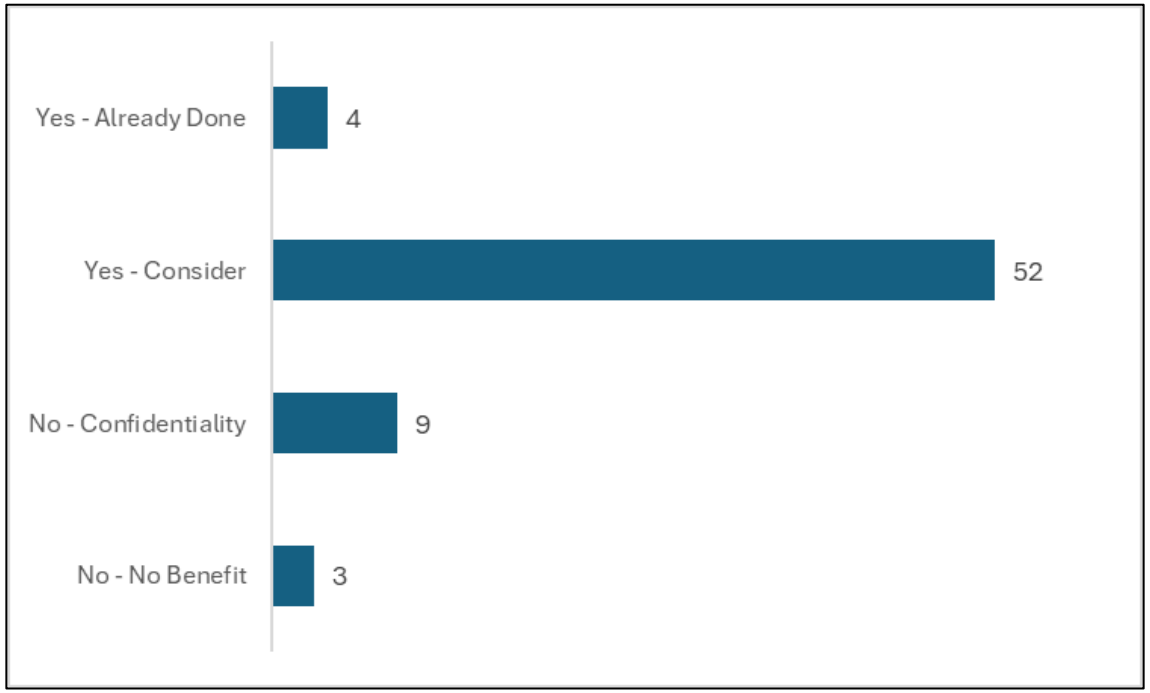
Introduction to Sample Filing

Survey Question #1

- If you were assembling a GLM rate model filing, would you consider letting Gen AI conduct a preliminary peer review?
  - No, I wouldn't for confidentiality concerns
  - No, I don't think it would improve upon our existing peer review process
  - Yes, this couldn't hurt and may even improve documentation
  - Yes, we've already done it…

Survey Question #1

- If you were assembling a GLM rate model filing, would you consider letting Gen AI conduct a preliminary peer review?



Sample Filing

Golden Retriever

Insurance Company

Golden Price Model v2.0

- Example filing fulfilling most of the NAIC GLM Checklist
- Non-Confidential
  - Can be uploaded to Gen AI without sharing trade secret info
  - Useful for thought experiments and widespread discussion
- Includes both modeling errors and minor typos
- Included in meeting invite
- Plots and Metrics are real-*ish*
  - They are the result of R code run on a table of data
  - The table of data is completely made up

Creating Artificial Dataset

Driver_Age	Number_Drivers	Number_Cars	Telematics_Score
54	2	2	5
55	1	3	6
56	1	1	9
69	3	2	10
66	2	1	1
45	2	2	6
53	2	1	6
27	2	2	6
32	4	1	10
40	3	2	6
53	2	1	4
40	2	1	5
27	1	2	10
48	1	2	6
60	4	1	1
30	1	2	2
49	1	2	5
24	3	1	6
59	2	1	10
20	1	2	6

- 12 Impactful Predictor Variables
  - State
  - Year
  - Driver Age
  - Number of Drivers
  - Number of Cars
  - Insurance Score (10 levels)
  - Telematics Score (10 levels)
  - Model Year
  - Prior Claims
  - Marital Status
  - Vehicle Use
  - Vehicle Type
- 1 Million Rows
  - Assumptions made on attribute distribution
  - Assumed no correlation among attributes *(Unrealistic)*

Creating Artificial Dataset

Driver_Age	Number_Drivers	Number_Cars	Telematics_Score	amount
54	2	2	5	0
55	1	3	6	0
56	1	1	9	0
69	3	2	10	0
66	2	1	1	0
45	2	2	6	0
53	2	1	6	0
27	2	2	6	0
32	4	1	10	0
40	3	2	6	0
53	2	1	4	0
40	2	1	5	0
27	1	2	10	0
48	1	2	6	0
60	4	1	1	0
30	1	2	2	0
49	1	2	5	14,773
24	3	1	6	0
59	2	1	10	0
20	1	2	6	0

- Simulate Target Variable
  - Start with “assumed” rate relativities
  - Simulate target variable
    - Simulation mean = 500 * product(assumed rate relativities)
    - Simulated pure premium = rtweedie()
      - mu = simulation mean
      - phi = 1,000
      - p = 1.5

Creating Artificial Dataset

Driver_Age	Number_Drivers	Number_Cars	Telematics_Score	amount	Random_Indicator_1
54	2	2	5	0	Missing
55	1	3	6	0	No
56	1	1	9	0	No
69	3	2	10	0	No
66	2	1	1	0	No
45	2	2	6	0	No
53	2	1	6	0	Missing
27	2	2	6	0	No
32	4	1	10	0	No
40	3	2	6	0	Missing
53	2	1	4	0	No
40	2	1	5	0	No
27	1	2	10	0	Missing
48	1	2	6	0	No
60	4	1	1	0	Yes
30	1	2	2	0	Missing
49	1	2	5	14,773	Missing
24	3	1	6	0	No
59	2	1	10	0	No
20	1	2	6	0	No

- Data Simulation
  - Add 100 “Nonsense” predictor variables
    - Step 1: 5% “Yes” and 95% “No” (randomly)
    - Step 2: Change 30% of all records to “Missing”
  - 3 nonsense variables kept in GLM based on lowest p-values
  - Renamed for fun
    - CAS Seminar
    - New Fender
    - Coldplay Tickets
- Final Table
  - 1,000,000 rows and 113 columns

Nonsense Variables

Golden Retriever

Insurance Company

Golden Price Model v2.0

Variable Name	Source	Description	Variable Treatment	Data Type
Recent Fender Purchase	Reputable Consumer Database	The primary insured has purchased a Fender electric, acoustic, or bass guitar within the 3 years prior the policy term effective date.	Modeled	Categorical
Recent CAS Seminar Purchase	Reputable Consumer Database	The primary insured has purchased access to a Casualty Actuarial Society within the 3 years prior the policy term effective date.	Modeled	Categorical
Recent Cold Play Tickets Purchase	Reputable Consumer Database	The primary insured has purchased tickets to a Cold Play concert within the 3 years prior the policy term effective date.	Modeled	Categorical

Nonsense Variables

Golden Retriever

Insurance Company

Golden Price Model v2.0

Variable Name	Source	Rational Explanation
Recent Fender Purchase	Reputable Consumer Database	People who play Fender guitars have good taste and avoid distasteful activities such as car accidents.
Recent CAS Seminar Purchase	Reputable Consumer Database	People who attend CAS Seminars are more risk conscious and drive more conservatively.
Recent Cold Play Tickets Purchase	Reputable Consumer Database	People who purchase Cold Play tickets often make poor decisions. Poor decision makers get in car accidents more frequently.

Nonsense Variables

Golden Retriever

Insurance Company

Golden Price Model v2.0

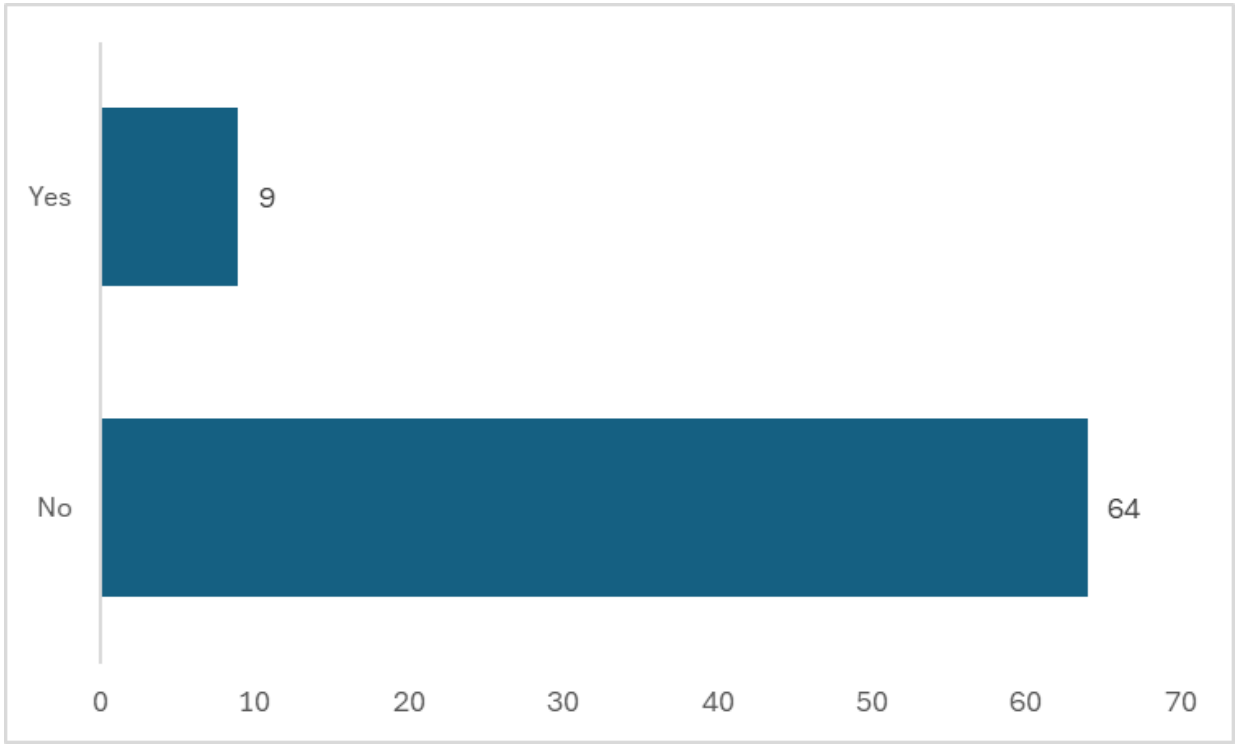
Variable	Good Rational Explanation	P-Value < 0.05	AIC Change > 0	F-Test P-Value < 0.05
New_Fender	X	X	X	X
CAS_Seminar	X	✓	✓	✓
Cold_Play_ Tickets	X	✓	✓	✓

Survey Question #2

- Should insurance companies surcharge Coldplay concert goers?

Survey Question #2

- Should insurance companies surcharge Coldplay concert goers?



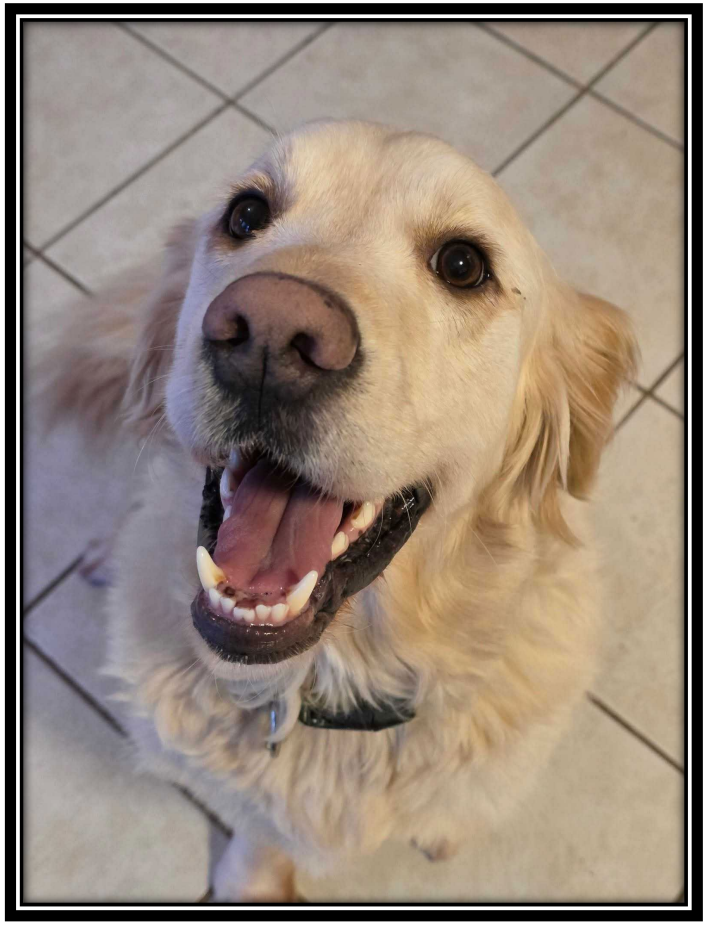
Survey Question #2

- Should insurance companies surcharge Coldplay concert goers?

DeepSeek says:

Cold Play Tickets: Offering a discount (proposed factor of ~1.127 for "Yes") for owning tickets to a specific musician's concert has no plausible causal relationship with driving risk. It risks being a proxy for age, socioeconomic status, or geographic location, raising **fair discrimination concerns**.

Sample Filing



"I, Dug Waffles Kloese, am an insurance expert.

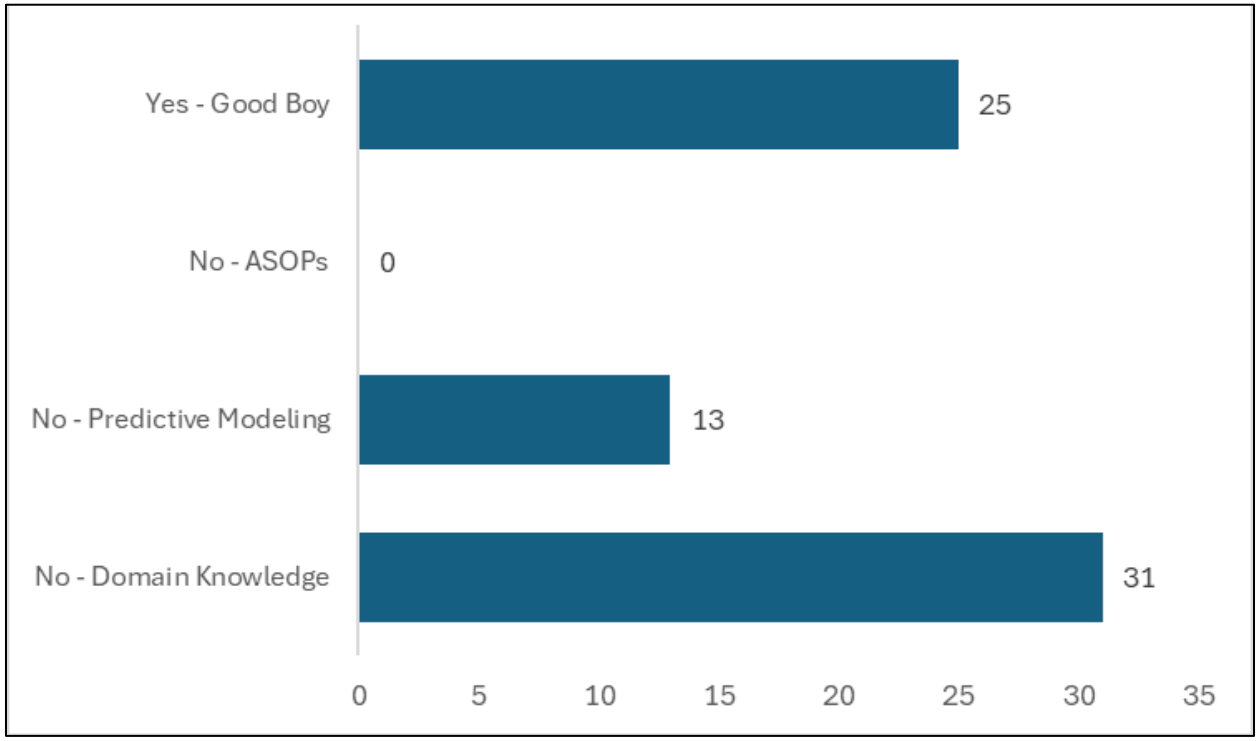
I have over 3 years of experience working closely with a credentialed P&C Actuary…"

Survey Question #3

- Is Dug Waffles Kloese qualified to opine on a pricing model?
  - No, he does not have appropriate P&C insurance domain knowledge
  - No, he does not have experience in predictive modeling
  - No, he didn't mention enough ASOPs
  - Yes, because Dug is a good boy

Survey Question #3

- Is Dug Waffles Kloese qualified to opine on a pricing model?



Sample Filing

Golden Retriever

Insurance Company

Golden Price Model v2.0

- Simple and short
  - Single pure premium GLM
  - 40 pages
- Sections
  - Introduction
  - Certification
  - Data
  - Model
  - Validation
  - Implementation

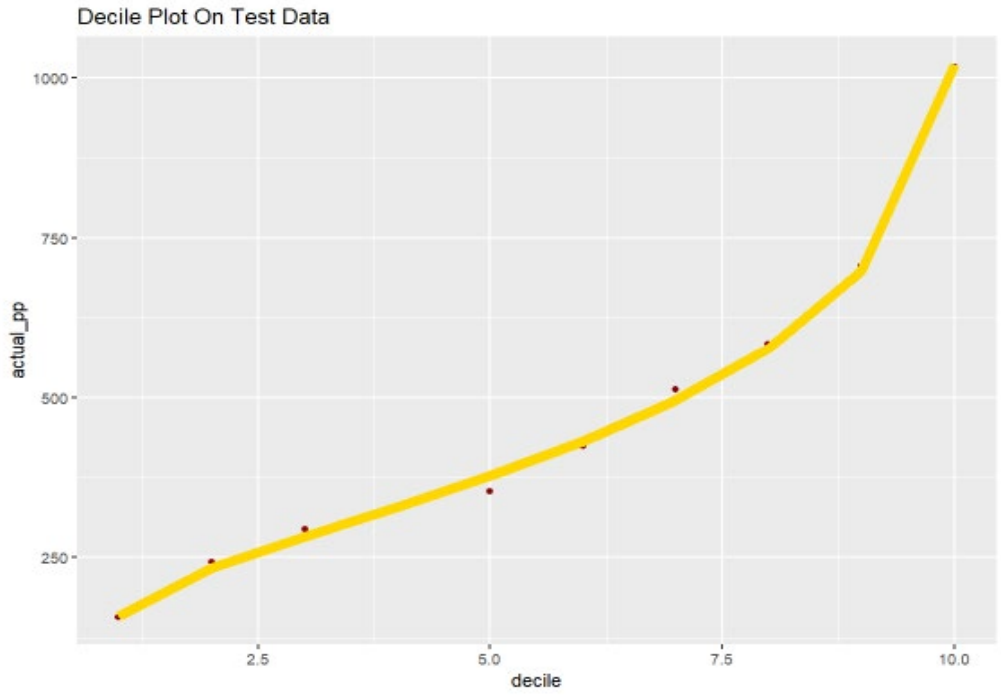
- Exhibits
  - Data Dictionary
  - Correlation Matrix
  - GVIF
  - Sample Modeling Data
  - Beta Coefficients and P-Values
  - F Tests and AIC
  - Deviance Residual Plot
  - Actual vs. Expected by Variable
  - Validation Plots
  - Indicated vs. Proposed
  - Rating Examples

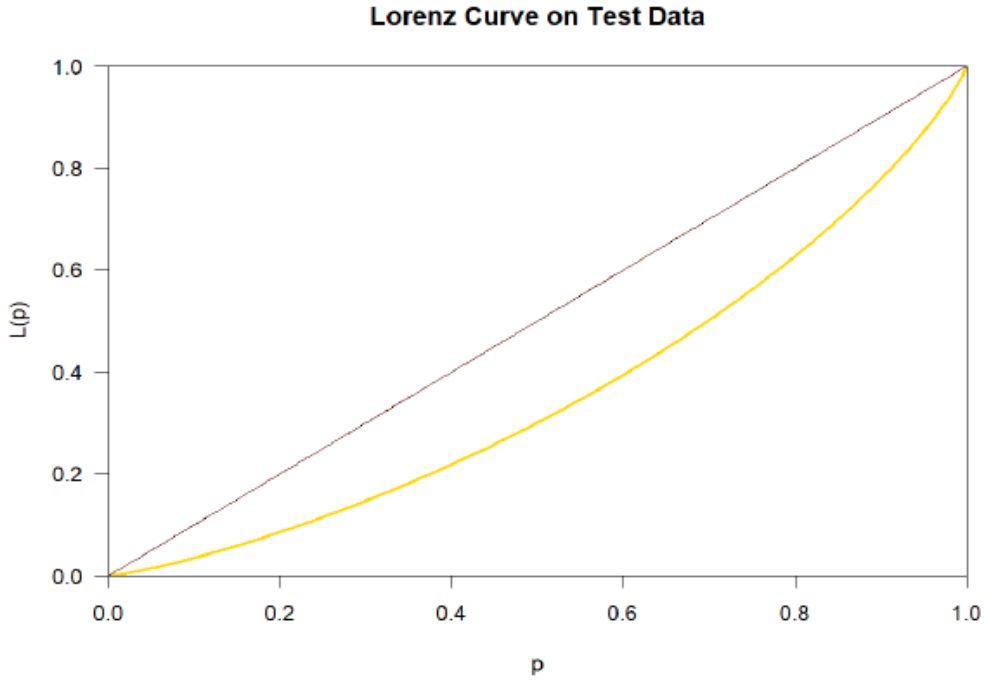
Sample Filing

Golden Retriever

Insurance Company

Golden Price Model v2.0



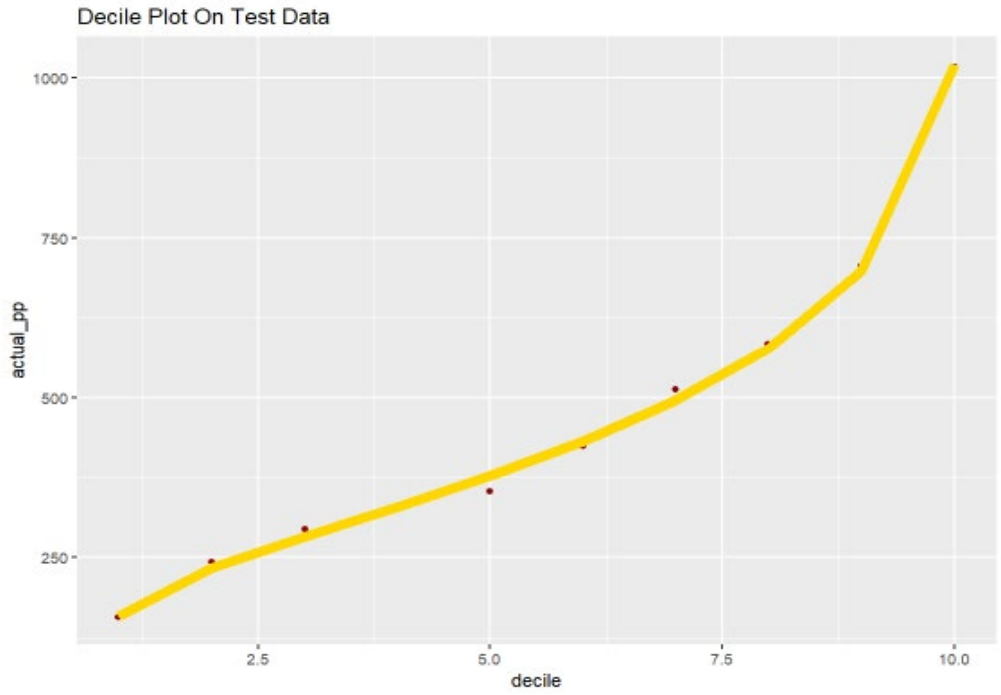


Sample Filing

Golden Retriever

Insurance Company

Golden Price Model v2.0



"I drew a line
I drew a line for you
Oh, what a thing to do
And it was all yellow"

-Coldplay. "Yellow." *Parachutes*
Parlophone, 2000

Issues

Golden Retriever

Insurance Company

Golden Price Model v2.0

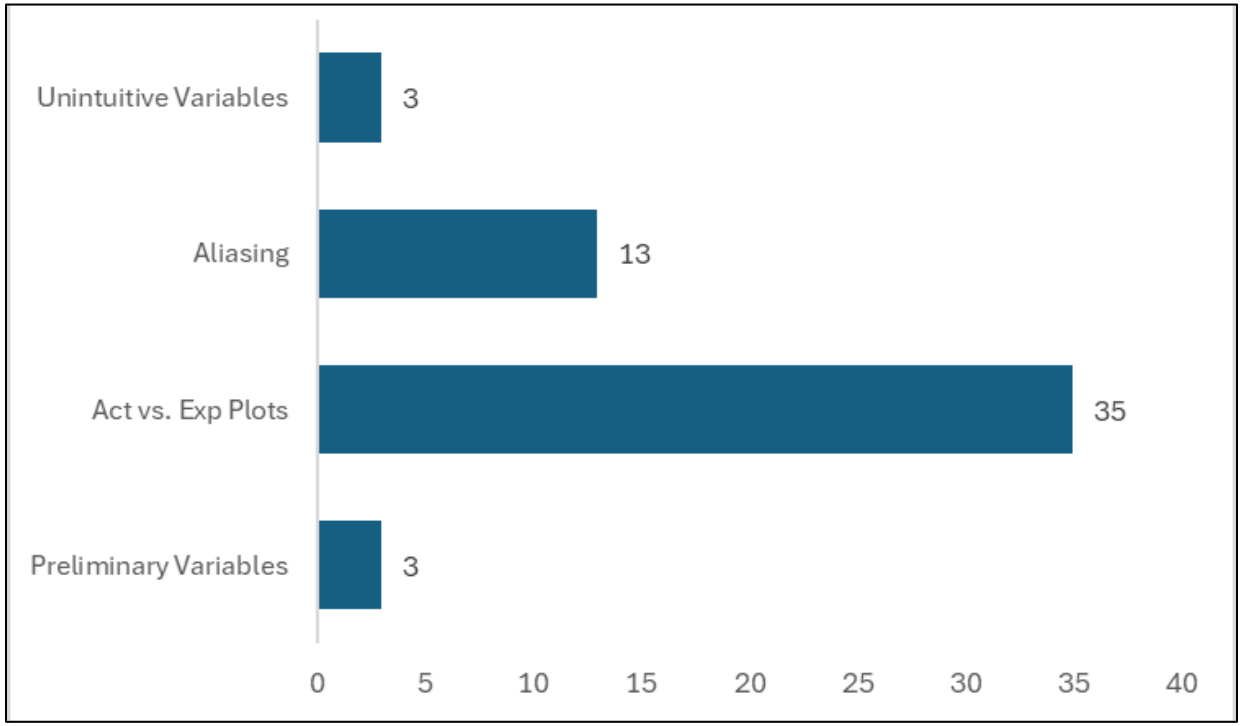
- High Priority Issues
  - # of initial variables not discussed
  - Actual vs. Expected plots provided on training data
  - Aliasing issue due to 30% missing data
  - Proposed factors assume everyone bought Cold Play tickets
  - Discounts granted for new guitars and CAS seminars
- Other Issues
  - Typos:
    - Error distribution
    - Link Function
    - Training/Test split
  - Poor rational explanations
  - Pearson correlation used for categorical variables
  - New Fender fails all statistical tests

Survey Question #4

- In your opinion, which of the following is the biggest issue?
  - # of initial variables in the preliminary model is not discussed
  - Actual vs. Expected plots provided on training data
  - Aliasing issue due to 30% missing data
  - Unintuitive variables are included

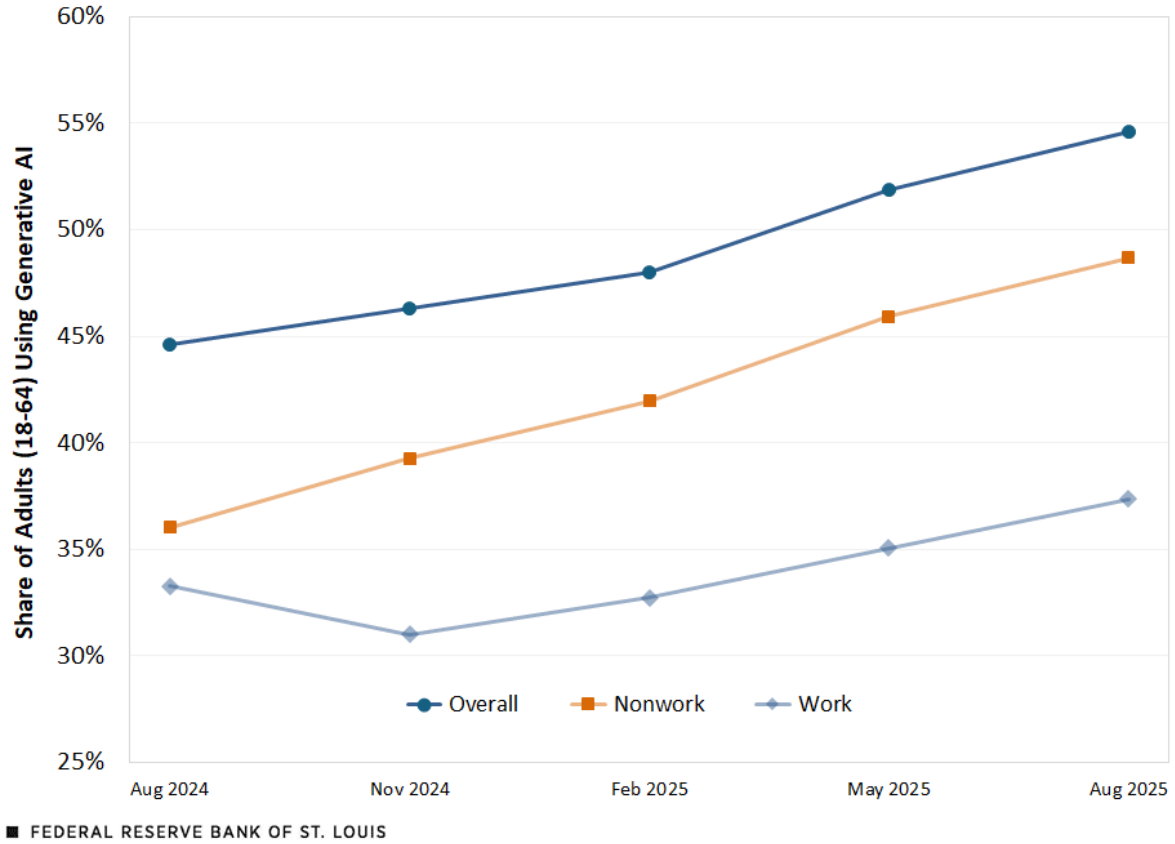
Survey Question #4

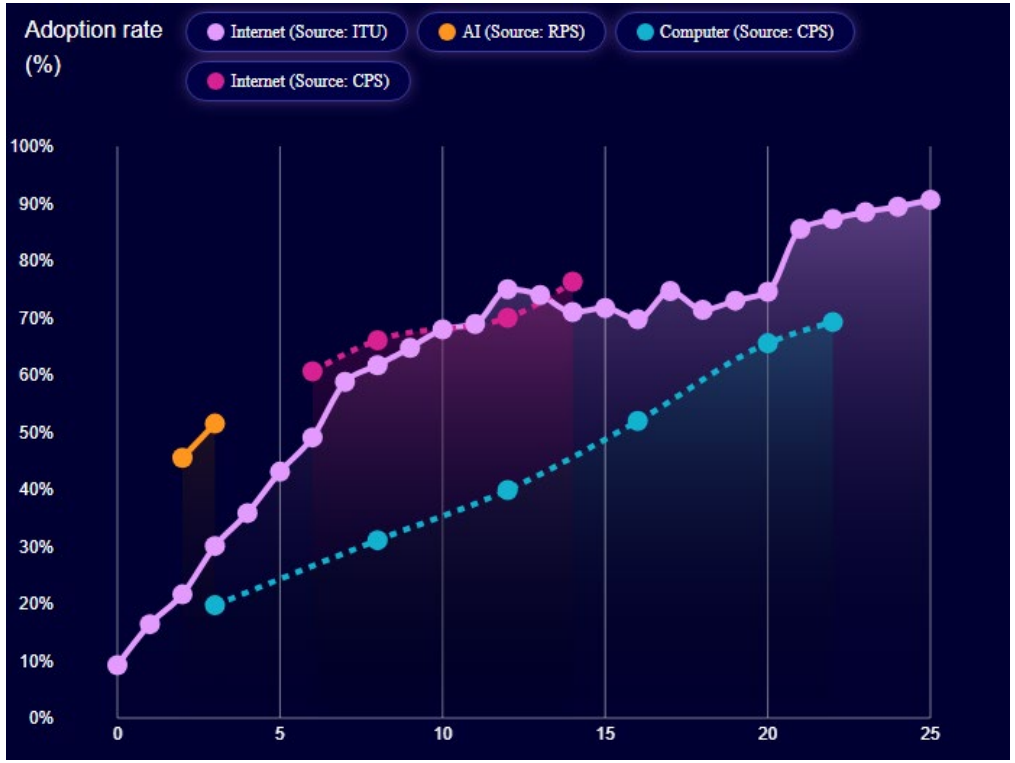
- In your opinion, which of the following is the biggest issue?



Gen AI Experiment

Rapid Adoption of Gen AI





Top LLMs – As of January 2026

1. ChatGPT (OpenAI)
2. Claude (Anthropic)
3. Gemini (Alphabet)
4. DeepSeek
5. Qwen (Alibaba)
6. Grok (xAI)
7. Ernie (Baidu)

Model 592 / 592	Overall	Expert	Hard Prompts	Coding	Math
gemini-3-pro	1	4	1	4	2
grok-4.1-thinking	2	6	4	7	10
gemini-3-flash	3	9	5	10	4
claude-opus-4-5-…	4	2	2	1	5
claude-opus-4-5-…	5	1	3	3	7
grok-4.1	6	21	10	14	17
gemini-3-flash (…	7	11	12	11	8
ernie-5.0-0110	8	12	11	12	1
gpt-5.1-high	9	7	9	15	6
gemini-2.5-pro	10	15	16	30	12

https://lmarena.ai/leaderboard

Gen AI Experiment

- Create list of problems in AI filing
- Draft prompt document based on NAIC GLM Checklist
- Run the same prompt in 4 different AI tools
- Score each Gen AI on problems caught
- Iterate on the prompt document to try to improve scores
- Score each Gen AI again

Gen AI Experiment

- How to Write a Prompt*
  - Task
  - Context
  - References
  - Evaluate
  - Iterate**
- Acrostic
  - Thoughtfully
  - Creating
  - Really
  - Excellent
  - Inputs

Gen AI Experiment

- How to Write a Prompt*
  - Task
  - Context
  - References
  - Evaluate
  - Iterate**

**"When you try your best,
but don't succeed
When you get what you want,
but not what you need…
…And I will try to fix you"

– Coldplay. "Fix You." *X&Y,* Parlophone, 2005

Gen AI Experiment

Task

- Experienced actuary tasked with assessing how a rate model is used to derive proposed rates

Context

- Identify data, modeling assumptions, and implementation
- Highlight any regulatory or actuarial best-practice considerations
- Reference information from filing

Reference

- No external references used for this exercise
- Could potentially use prior NAIC reports in the future

Evaluate

- Each AI was scored using our original prompt

Iterate

- Took note on which filing issues were not caught
- Added new questions and revised others for a round 2

Gen AI Experiment

- Versions Tested

Claude (Opus 4.5)

Chat (GPT 5.2) *Note: Round 1 was done using GPT 5

Gemini (Pro via NotebookLM – Gemini 3 Flash)

DeepSeek (v3.2)

Gen AI Experiment

- Round 1 Scorecard (30 Total Issues)

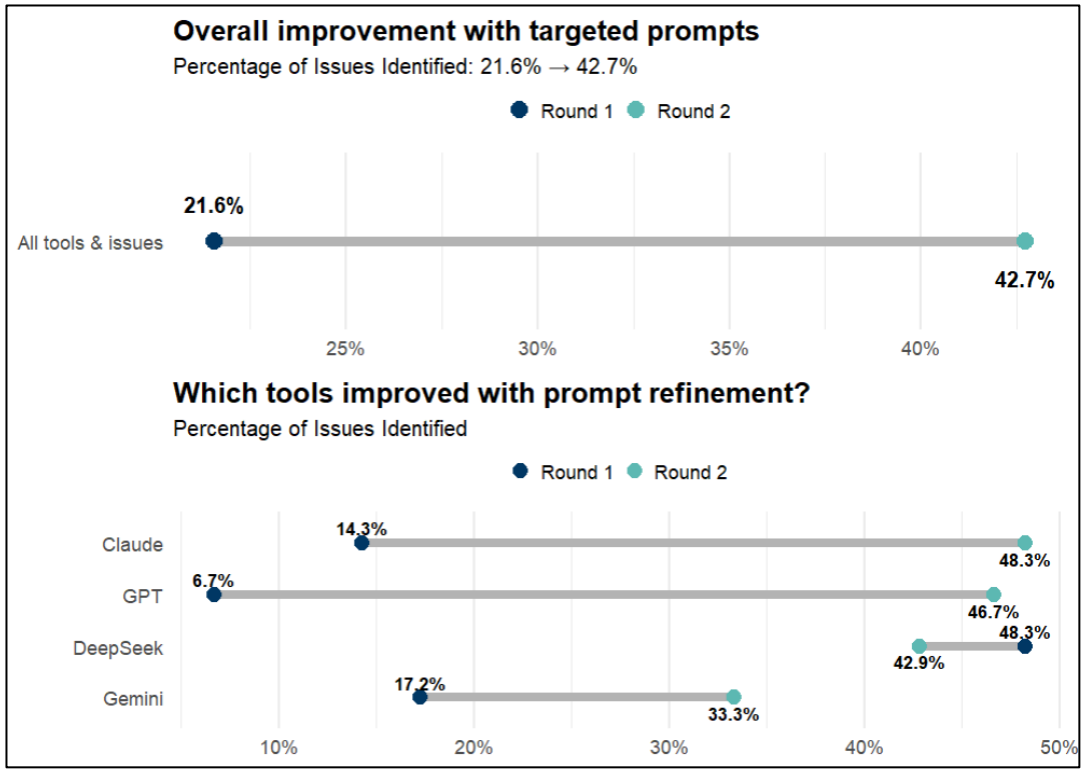
Claude	DeepSeek	Gemini	GPT
4	14	5	2

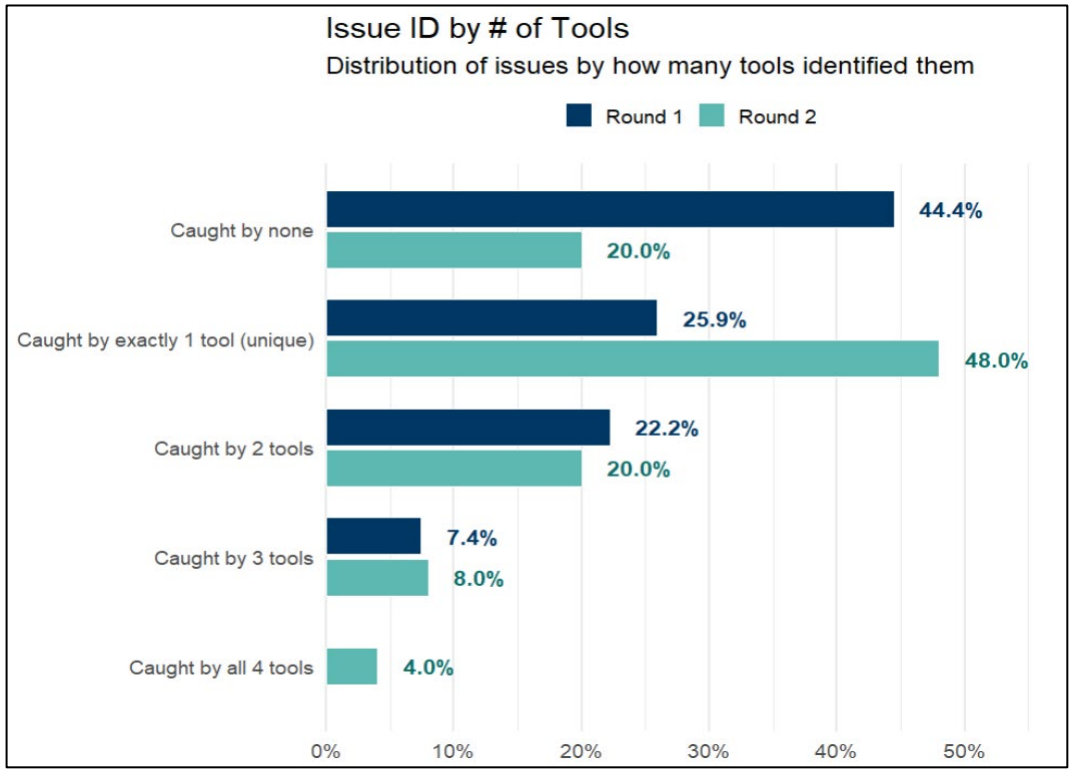
- Examples missed by all Gen AI
  - The model was built by someone unqualified
  - Inconsistencies in training/test split %
  - The error distribution and link function don't make sense
  - The proposed factors for "Missing" are unreasonable

Gen AI Experiment

Round 2 Scorecard (30 Total Issues)

Claude	DeepSeek	Gemini	GPT
14	12	10	14

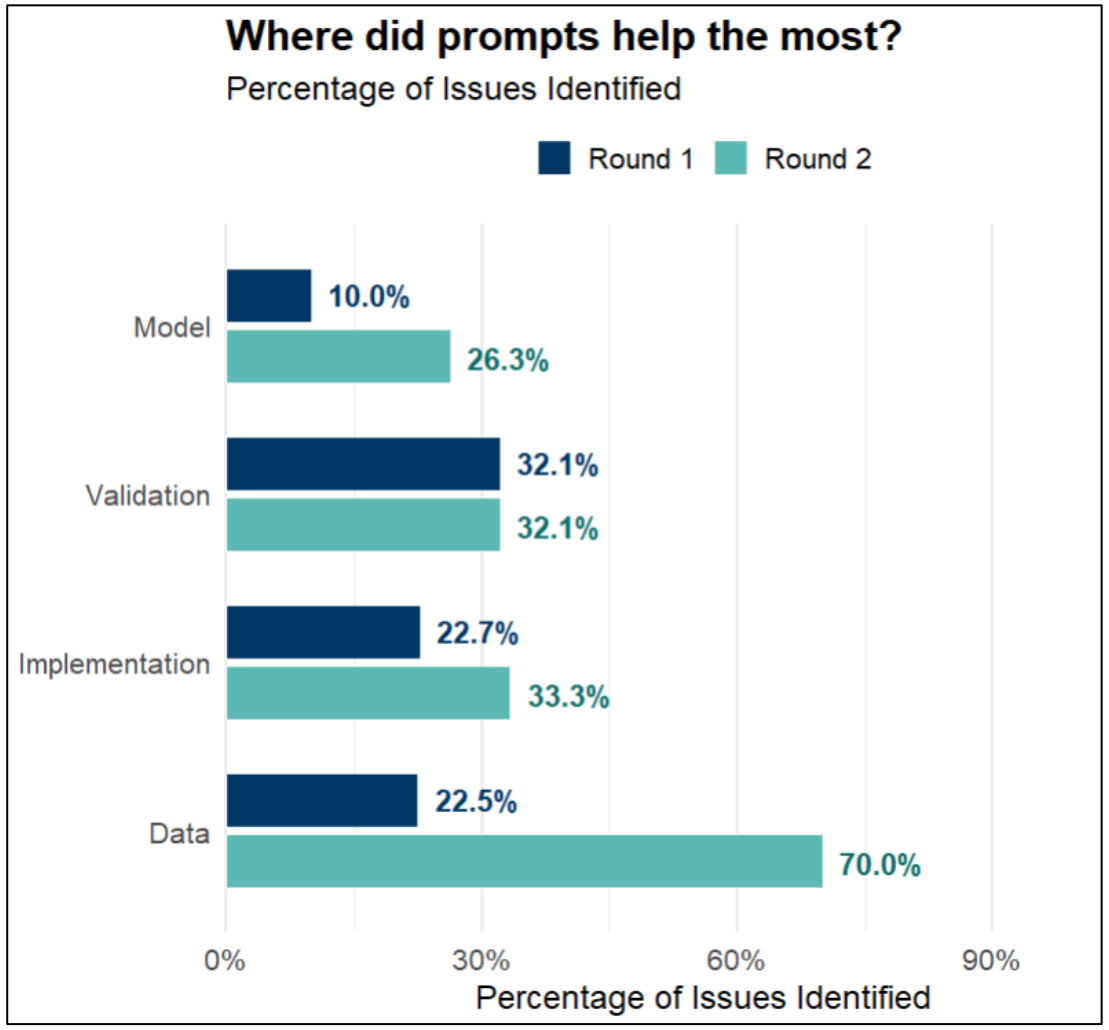




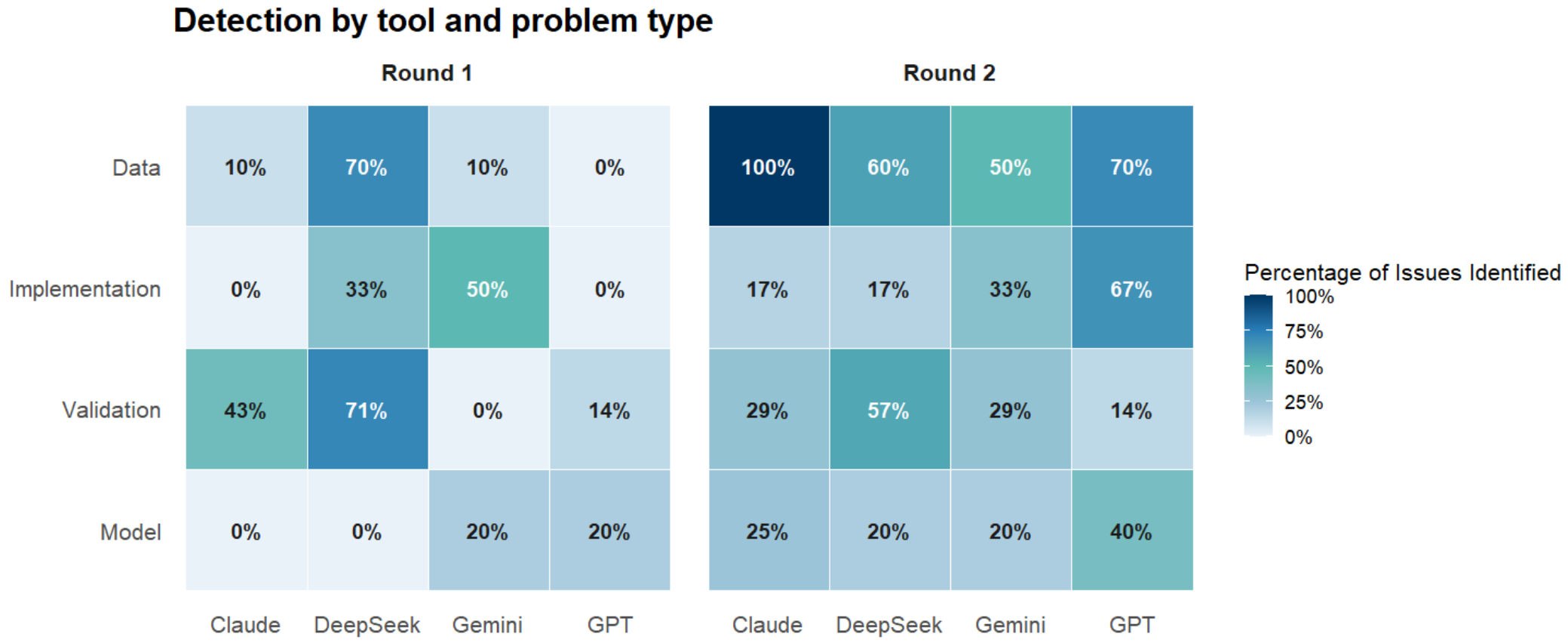
Gen AI Experiment

- Scores for 3 LLMs improved as questions got more specific
- DeepSeek's score got worse!
  - Round 1:
    - Specifically called out problematic "pure p-value fishing."
    - Listed out each instance of selected factors deviating from indicated.
  - Round 2:
    - Did not mention p-value fishing or unknown # of preliminary variables
    - States proposed factors are higher than indicated "(e.g., age 16)" without mentioning all impacted variables

Where Did Prompts Help?



Issue Detection by Problem Type



Gen AI Experiment

- Examples newly caught in Round 2
  - The model was built by unqualified individual
  - Poor rational explanations for new Fender, Coldplay tickets, and CAS seminars
  - High GVIF suggests problem with multi-collinearity
  - External data source documentation
- Examples still not caught by multiple LLMs
  - Inconsistencies in training/test split %
  - Model validation not performed on testing data
  - Erroneous error distribution and link function assumptions
  - Cramer's V is better metric for correlation in categorical variables
  - Stating results of statistical test is not a rational explanation
  - The polynomial fit to driver age creates poor fit for the oldest drivers

Conclusions

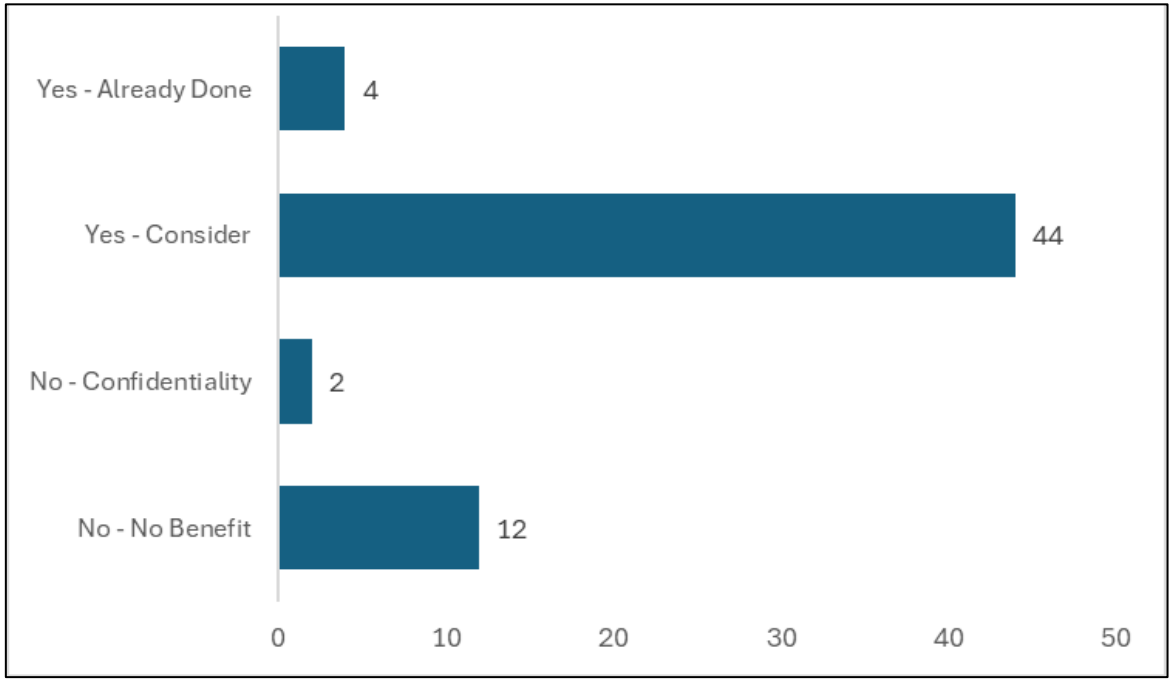
Survey Question #5

- If you were assembling a GLM rate model filing, would you consider letting Gen AI conduct a preliminary peer review?
  - No, I wouldn't for confidentiality concerns
  - No, I don't think it would improve upon our existing peer review process
  - Yes, this couldn't hurt and may even improve documentation
  - Yes, we've already done it…

Survey Question #5

- If you were assembling a GLM rate model filing, would you consider letting Gen AI conduct a preliminary peer review?

AI Peer Review	Before	After
No - No Benefit	3	12
No - Confidentiality	9	2
Yes - Consider	52	44
Yes - Already Done	4	4
Total	68	62



The Big Questions

- Can Gen AI Review a Pricing GLM?
  - Probably shouldn't…
    - Impossible to create one list of questions to capture all potential issues
    - Insurance experts still needed to verify and check for AI hallucinations
    - Gen AI doesn't have nuance and common sense
    - Gen AI is not "deterministic", and reviews might lose consistency
    - Question phrasing may cause issues due to "AI Sycophancy"
    - Actuaries needed to research and adapt to new techniques

The Big Questions

- Can Gen AI Assist Reviewing a Pricing GLM?
  - Absolutely!
    - Rate filings can reach 1000's of pages
    - Excellent at writing code
    - Extract and summarize filing for a preliminary look
    - Extract data in tables and visualize with plots
    - Handles repetitive tasks that can be automated
    - Provide references on innovative approaches

CoPilot at the NAIC

- Currently uses OpenAI GPT 5.2
- All Copilot activity remains within NAIC's secure Microsoft's 365 tenant.
- Content is not used to train AI models and is not retained by the AI.
- Copilot reads data only at rest (e.g., documents in SharePoint) and does not store it for future queries.
- Access is limited strictly to data the logged-in user is already permitted to view.
- Copilot fully adheres to existing NAIC security and permission policies.

CoPilot at the NAIC

- Access to content is governed by NAIC security labels.
  - Documents labeled above **public** (e.g., *Sensitive, internal use only)* cannot be accessed by Copilot.
- All data Copilot can read remains subject to NAIC's security and data-loss-prevention controls, including:
  - **Cortex XDR** – advanced threat detection and response.
  - **Netskope DLP** – real-time protection against the sharing of sensitive information.
  - **These systems ensure that staff interactions with information are monitored and protected across devices and applications.**

CoPilot on Our Team

- **Example use:**
  - Extracting basic information during the scheduling phase
  - Finding references about modeling issues
  - Writing code
    - Ex: Import Excel sheets of indicated vs. selected exhibits and generate plots; parse information; aid tool development

Potential Future Research

- Build an agent
- Rescore the Gen AI models as new versions are released
- Train a new, confidential in-house model using NAIC reports

"I was just guessing at
numbers and figures
Pulling the puzzles apart
Questions of science,
science and progress…
…Nobody said it was easy"

– Coldplay. "The Scientist."
A Rush of Blood to the Head, Parlophone, 2002

Q & A

- Sam Kloese, ACAS, CSPA, MAAA
  - skloese@naic.org
- Roberto Perez, FCAS, CPCU, MAAA
  - rperez1@naic.org

Appendix

Issue	Page #	Quote or Reference	Problem
1	3	All available variables from Reputable Consumer Database were analyzed in a preliminary model, but only the variables with the low p-values were kept in our final pricing model.	The number of variables considered initially is unclear. A few variables may pass p-value threshold by chance.
2	4	I, Dug Waffles Kloese, am an insurance expert. I have over 3 years of experience working closely with a credentialed P&C Actuary.	Someone who knows an Actuary, not a credentialed Actuary, certified the pricing model
3	5	80% of the data was used for training and 20% was reserved for the test dataset.	The table of exposures shows the split is 70%/30%
4	6	The correlation metric analyzed was the Pearson correlation.	Cramer's V is a better metric for correlation of categorical variables
5	6	The results are incorporated into this report as Exhibit 2. Most variables had a GVIF < 2.0, with the exception of the Driver Age variables. A polynomial fit was achieved by including (Driver Age), (Driver Age)^2, and (Driver Age)^3 as terms in the model	High GVIF suggests a problem with multi-collinearity
6	7	The GLM assumes that the error distribution follows a Gamma distribution. The Gamma distribution is useful because it is a blend of Poisson and Tweedie distributions. The Gamma distribution is also especially useful in the pure premium context because it can accommodate a target variable where many records have a value of 0. We selected a p power parameter value equal to 1.5 for the Gamma distribution.	This is a typo. "Gamma" and "Tweedie" were swapped throughout this section. "Tweedie" makes sense for pure premium distribution. "Gamma" does not.
7	7	An identity link function was chosen as it was desirable to derive multiplicative indicated rating factors.	This is a typo. The log link was used in modeling and the log link produces multiplicative rating factors.
8	8	Model predictions were appended to the training set based on the final GLM. This allowed us to compare actual pure premiums to predicted pure premiums.	Fitted vs. Actual plots should be provided based on test data, not based on training data.
9	9	Cold Play Tickets – The proposed factor for the “Missing” level is set equal to the same rating factor for “Yes”.	It seems unreasonable to assume that everyone whose information is "Missing" bought Coldplay tickets.
10	10	The primary insured has purchased a Fender electric, acoustic, or bass guitar within the 3 years prior the policy term effective date.	"New_Fender" sounds like a reasonable auto pricing variable, until you see in the data dictionary that it is about guitars and not vehicle fenders.

Issue	Page #	Quote or Reference	Problem
11	10	The primary insured has purchased access to a Casualty Actuarial Society within the 3 years prior the policy term effective date.	This variable is not intuitively related to auto insurance claim risk
12	10	The primary insured has purchased tickets to a Cold Play concert within the 3 years prior the policy term effective date	This variable is not intuitively related to auto insurance claim risk
13	11	Insurance_Score have a very low p-value, demonstrating predictiveness.	Statistical tests are not the same thing as a rational explanation. The explanation should offer a plausible connection between the characteristic and insurance claim risk.
14	11	People who play Fender guitars have good taste and avoid distasteful activities such as car accidents.	This rational explanation is not intuitive and the variable seems questionable.
15	11	People who attend CAS Seminars are more risk conscious and drive more conservatively.	This rational explanation is not intuitive and the variable seems questionable.
16	11	People who purchase Cold Play tickets often make poor decisions. Poor decision makers get in car accidents more frequently.	This rational explanation is not intuitive and the variable seems questionable.
17	17	New_FenderYes and New_FenderMissing have p-values > 0.05	All levels of this variable have a high p-value
18	18	New_Fender fails the F Nested model test with a F statistic p-value = 0.155	New_Fender fails the F nested model test
19	19	New_Fender fails the AIC test since the AIC is lower when the variable is excluded.	New_Fender fails the AIC test
20	20	Driver Age fitted curves downward for higher values	The 3rd degree polynomial is creating unintuitive factors for the highest driver ages

Issue	Page #	Quote or Reference	Problem
21	36	Proposed factors are much higher than indicateds for Insurance Score Tiers 7-10.	The company should provide additional justification for setting proposed factors higher than indicated.
22	38	New Fender - Proposed Factors	It seems unreasonable to offer discounts based on whether someone bought a guitar. Furthermore, the company doesn't have data to support varying the discount by type of guitar.
23	38	CAS Seminar - Proposed Factors	It seems unreasonable to offer discounts based on whether someone went to a CAS seminar.
24	39	Cold Play Tickets - Proposed Factors	It seems unreasonable to surcharge people buying Coldplay tickets. It's also unreasonable to surcharge when we don't know whether they purchased tickets or not.
25	40	Our external data comes from Reputable Consumer Database, a third-party data vendor with data on the purchasing habits for roughly 70% of Americans.	No link is provided for the external data source
26	41	The target variable is combined coverage pure premium.	This can result in an erroneous model if portfolio distributions change.
27	42	Coverage Package level was included as an offset to account for varying coverage levels. The coverage package levels are grouped into minimum, medium, and maximum coverage.	Unless these are the only 3 levels offered this could cause issues in the model
28	43	The proposed rating algorithm is expected to result in premiums that are on average 10% below our existing charters.	This could be concerning since the plan is intended for new business, which tend to be more expensive than renewals.
29	44	Deviance Residual Plot	Violates regression assumptions
30	45	Driver Age Exposure Chart	Shows clusters in unnatural clusters and jumps at certain ages.

Gen AI Experiment

Task

- “You are an expert actuary with over 20 years of experience in rate filings, predictive modeling, and insurance regulation. You are reviewing a rate filing consisting of multiple files, including PDF documents, and Excel workbooks (All attached files excluding “Combined_Questions.docx" correspond to the filing).
- Your goal is to assess how a rate model is used in determining the proposed rates. Your task: Answer the questions provided in “Combined_Questions.docx" in order, using clear, concise, and regulator-focused language.

Gen AI Experiment

Context – "For each:

1. Reference specific evidence from the filing
2. Identify data, modeling assumptions, and implementation
3. Note gaps, inconsistencies, or missing information and suggest what would be needed.
4. Highlight any regulatory or actuarial best-practice considerations. The filing consists of all attached files except for the questions document. Make sure to provide answers to each question in the order provided in the questions document."