As there were so many answers to go through on these questions, I was not able to make comments on every single answer’s validity. However, I was able to come up with a scoring system for which I awarded the answers 0, 1, or 2 points. A fully correct answer would be coded green in the answer sheet and would be worth 2 points. A partially correct or incomplete answer would be coded yellow and worth 1 point, while red answers are essentially hallucinations and are worth 0 points.

A general note is that we need to have the LLM state when it cites from a state insurance code that the answer only applies in the insurance context. However, that does not weigh into correct or incorrectness of these answers, so there were no points deducted from that at this time.

This is the best I can do with our current time constraints with this amount of material. However, we can revisit the answers that are incorrect to establish ground truths for these questions as a next phase of the project.

Q1 – The answers appear to be the same as was in response to Q2. As such, none of the answers address the question listed.

Q2 –

9 – 0 point answers

19 – 1 point answers

22 – 2 point answers

=

63/100 possible points

Q5 – correct answer

Q8 – correct answer

Q10 – correct answer

Q26 –

9 – 0 point answers

3 – 1 point answers

9 – 2 point answers

=

21/ 44 possible points

Q 47 – a general note is that when the model is asked a question about “which states” do something, the ideal answer will give a list of states that do that thing. Just a list of the names as the first part of the answer. The model might then say “here are the states that do not” and “here are the states on which there is no information.” As it is, these answers still require the reader to go through and evaluate each state’s law and decide for themselves if that law addresses that issue. Ideally, the model will do more work with the information retrieved, both to categorize into “yes” “no” or “unknown” categories, and then explain with more synthesized information, rather than simply listing out snippets of legal text.