Results

Important Considerations

- Experiment 3 Results
  - Responder: Llama
  - Judge 1: GPT
  - Judge 2: Claude
- Tested on 10 data points
- Used the same Llama Responses generated throughout so the explanations and accuracy would stay the same for the initial set of responses. Prompt injection with and without RAG builds off of the same initial set of responses.
- The code for RAG needs to be double checked:
  - Had to make a handmade function to find similar patterns based on labels (uncertainty vs certainty) because after inspecting the quadrant cluster sentences were considered similar based on common topics rather than labels. Ie. The nearest neighbor of an uncertain sentence was a certain sentence in some cases. Label and uncertainty language seems to be ignored

PROMPTS W/O RAG

- **Confidence Instruction**: "Generate language that expresses the likelihood of your answer being correct based on the evidence or knowledge available to your system. Use general person phrases such as, 'This answer may be correct,' 'This answer is definitely correct,' or 'This answer is probably not correct,' to clearly indicate the probability of accuracy. Add nuance to the statements based on your spectrum of uncertainty, eg. kinda uncertain, not really certain, very certain, very uncertain, etc"
- **Prompt**: “Please format the response 'Final answer: X . Explanation: '"

  f"Here is the question: '{question}'. "

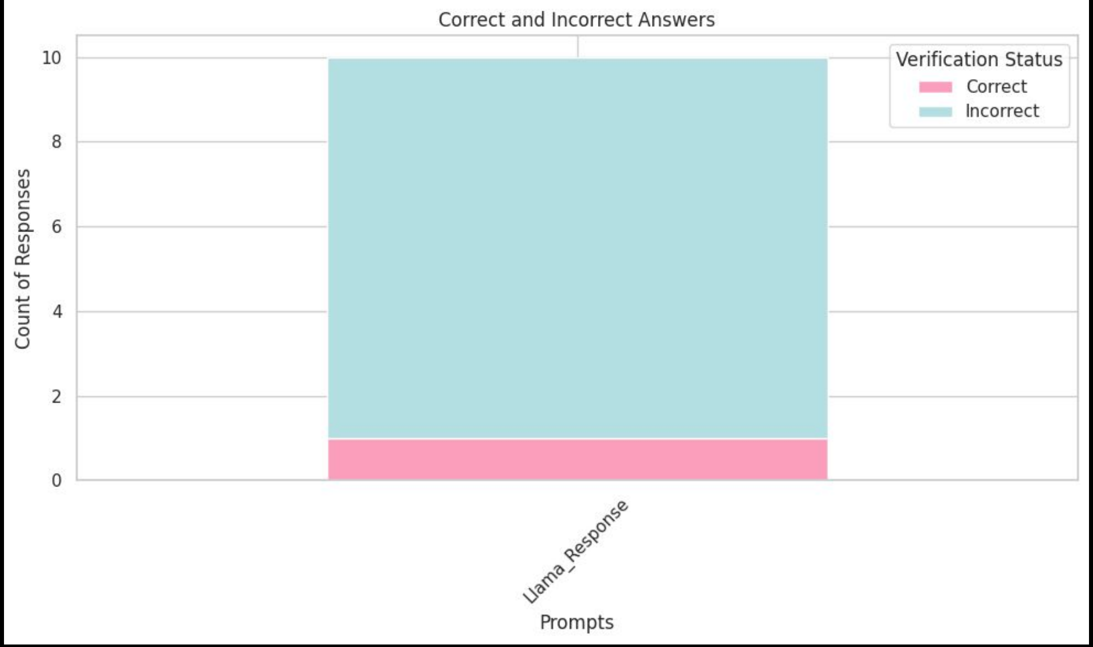
  f"Here is the original answer: '{Llama_response}'. "

  f"{confidence_instruction} Make sure the final answer from the original answer stays the same, but just adjust the phrasing in the explanation part to reflect the confidence instructions."

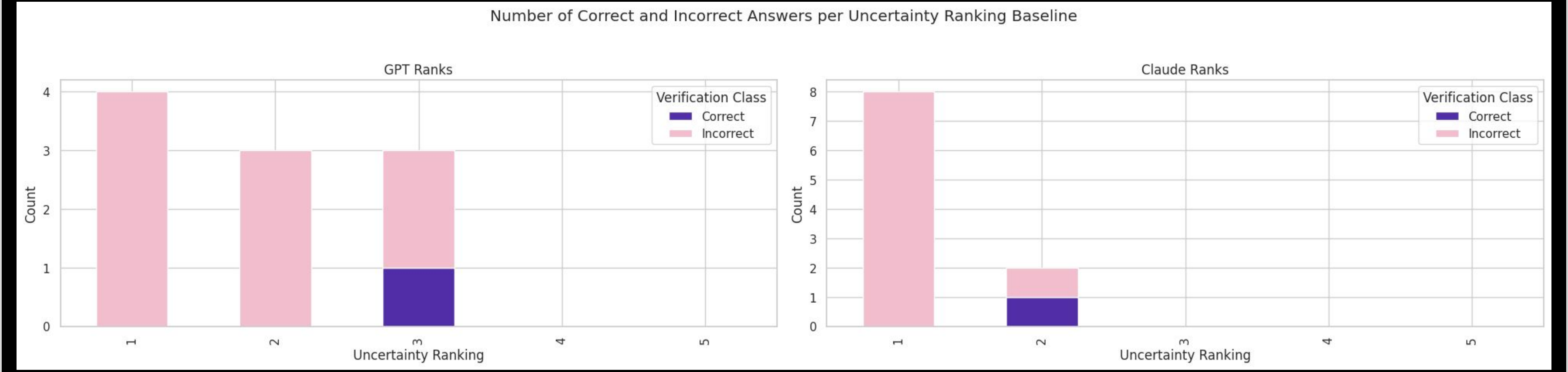
Uncertainty Prompt for Both W/ and W/O RAG

Prompt to Rank Uncertainty: Please generate ratings based on how effectively the response communicates uncertainty, using a scale of 1 to 5. On this scale, 1 indicates no presence of uncertainty language, suggesting complete certainty, for instance the sentence does not contain any words like 'may' or probably' and would have statements like 'certainly'. A rating of 2 suggests minimal uncertainty language, insufficient to dissuade a graduate student from trusting the response, a sentence would contain phrases like 'may be correct or probably correct'. A 3 indicates a moderate level of uncertainty, which might sway a graduate student's trust in the response, presenting a balanced, 50-50 likelihood of changing their mind, a sentence would contain of mixture of certain and uncertain phrases like, 'may be correct, but should double check'. A score of 4 denotes a substantial presence of uncertainty language, likely convincing a graduate student to distrust the response, for instance if the response said to 'double check' or 'probably incorrect'. Lastly, a 5 signifies that the response excellently conveys uncertainty, to the extent that a graduate student would definitely not trust the response, for instance 'definitely incorrect', 'very likely incorrect', or 'very unsure'. Please format your rating as follows, 'I rate this response a X out of 5', followed by an explanation.' Response: {response_text}

Accuracy
9 incorrect, 1 correct



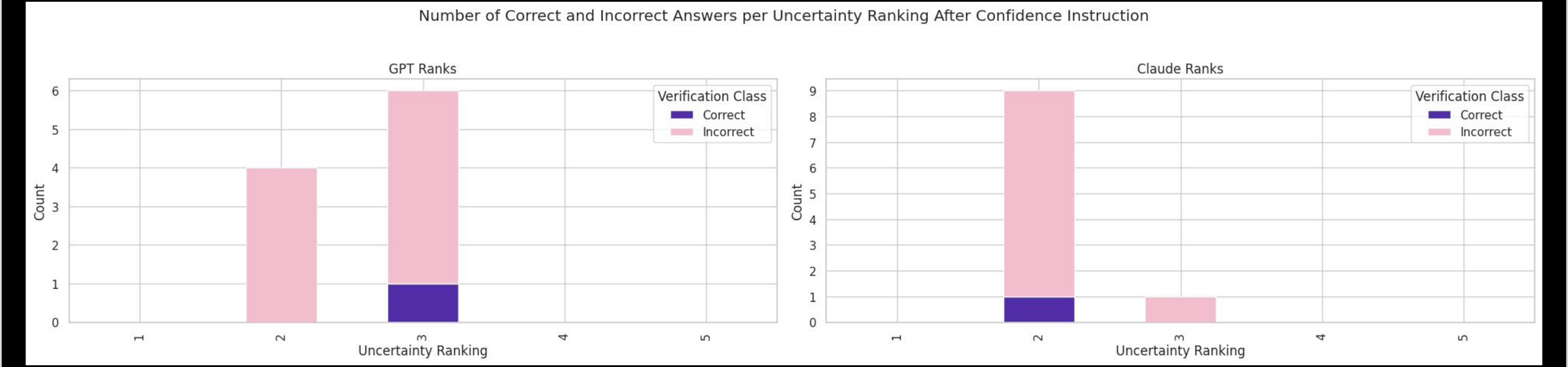
Baseline Results Without Any Prompt Injection



GPT_MEAN: 1.9
GPT_VAR: 0.767
GPT_STD: 0.876
GPT_SEM: 0.277

CLAUDE_MEAN: 1.2
CLAUDE_VAR: 0.767
CLAUDE_STD: 0.422
CLAUDE_SEM: 0.133

Results With Prompt Injection (Confidence Instruction)



GPT_MEAN: 2.6
GPT_VAR: 0.2667
GPT_STD: 0.516
GPT_SEM: 0.163

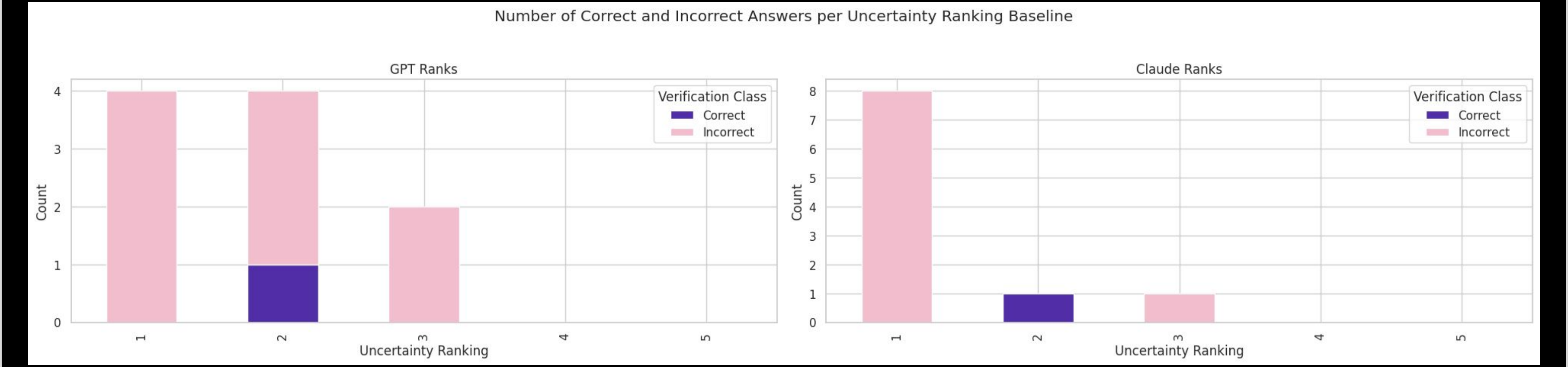
CLAUDE_MEAN: 2.1
CLAUDE_VAR: 0.1
CLAUDE_STD: 0.316
CLAUDE_SEM: 0.1

PROMPTS W/ RAG

- **Confidence Instruction**: ""Generate language that expresses the likelihood of your answer being correct based on the evidence or knowledge available to your system.
- **Prompt**: “"""Using these example patterns:
  For expressing certainty:
  {', '.join(p['text'] for p in certain_patterns)}
  For expressing uncertainty:
  {', '.join(p['text'] for p in uncertain_patterns)}
  Here is the original answer: '{Llama_response}'.
  Make sure the final answer from the original answer stays the same, but adjust the phrasing in the explanation part to reflect the confidence instructions.
  "{confidence_instruction} Use language patterns similar to the examples above.
  Format as 'Final answer: X . Explanation: Y'
  """"
- **Example Patterns:** certain_patterns = find_similar_patterns(question, certainty_type="certain", limit=3)
  uncertain_patterns = find_similar_patterns(question, certainty_type="uncertain", limit=3)

Baseline Results Without RAG

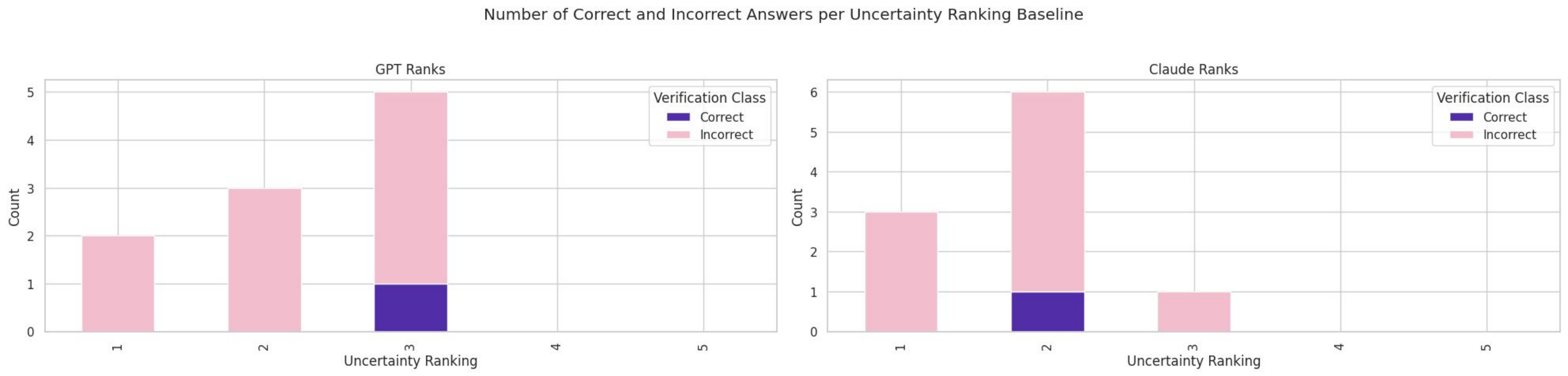
Re-ran the judges on the baseline responses with no prompt injection, just to see if there is variation in rank, but the baselines are pretty similar.



GPT_MEAN: 1.8
GPT_VAR: 0.622
GPT_STD: 0.789
GPT_SEM: 0.249

CLAUDE_MEAN: 1.3
CLAUDE_VAR: 0.622
CLAUDE_STD: 0.675
GTP_SEM: 0.213

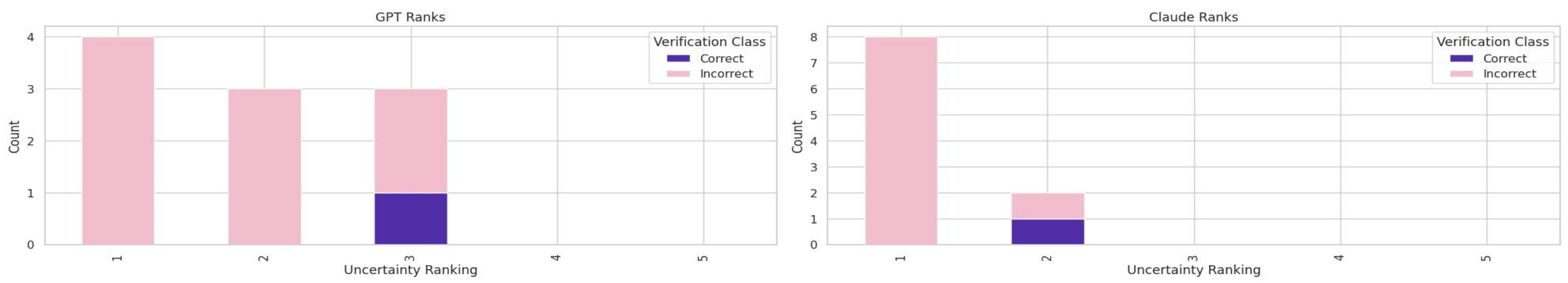
Results With Rag



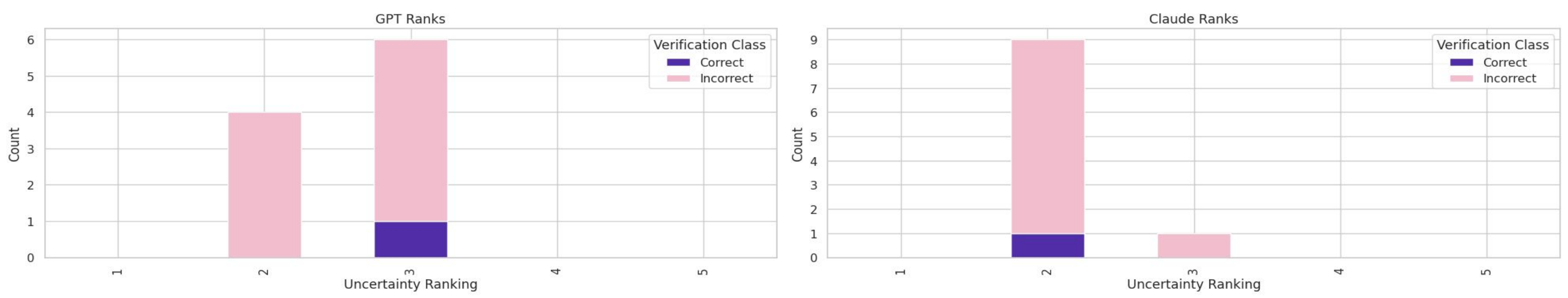
GPT_MEAN: 2.3
GPT_VAR: 0.678
GPT_STD: 0.823
GPT_SEM: 0.260

CLAUDE_MEAN: 1.8
CLAUDE_VAR: 0.4
CLAUDE_STD: 0.632
CLAUDE_SEM: 0.2

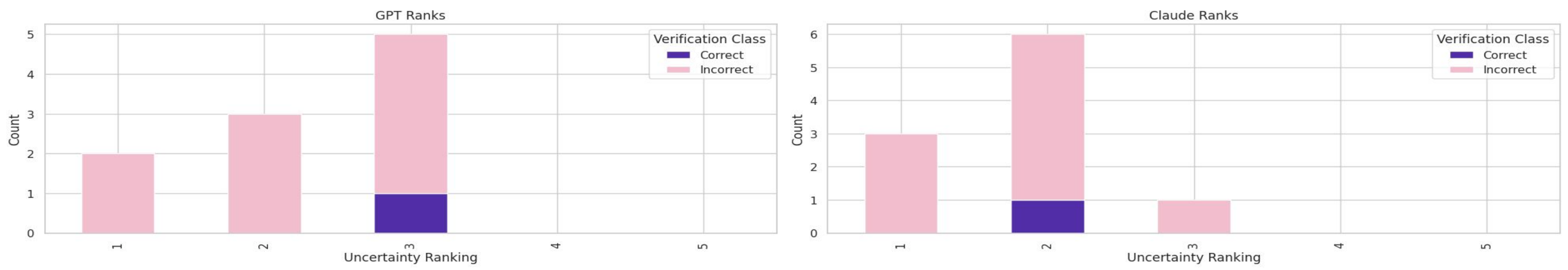
Base:



W/O RAG:



W/ RAG:



CLAUDE RESULTS

	MEAN	VAR	STD	SEM
Baseline 1	1.2	0.767	0.422	0.133
Baseline 2	1.3	0.622	0.675	0.213
W/O RAG	2.1	0.1	0.316	0.1
W/ RAG	1.8	0.4	0.632	0.2

GPT RESULTS

	MEAN	VAR	STD	SEM
Baseline 1	1.9	0.767	0.876	0.277
Baseline 2	1.8	0.622	0.789	0.249
W/O RAG	2.6	0.2667	0.516	0.163
W/ RAG	2.3	0.678	0.823	0.260

TAKEAWAYS

- W/O RAG has better results for calibrating the model based on how incorrect it is. More of the incorrect answers were categorized in the less certain bins 2 & 3.
- W RAG categorized incorrect answers in bin 1 which is the most certain, did not show improvement from the method W/O RAG.
- The one correct answer we did have, was categorized in the less certain bins. When I inspected closer it was because a section in the explanation was incorrect even though the final answer was correct. Originally the model classified it as wrong because of this but then I went back in and categorized it as correct, since the final answer was correct.
- RAG was still better than the baseline responses which were very certain.
- For the ones that were incorrect but still classified as certain, RAG could have been reinforcing that certainty.