Evaluation Framework

1. Data Preparation

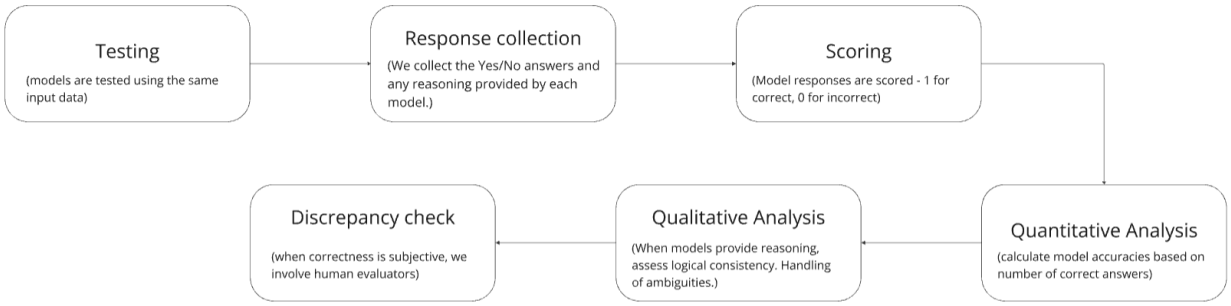
- **Input Dataset:** The provided paragraph and Boolean questions serve as the test dataset. For the interview, you can emphasize the importance of having diverse and representative examples to ensure the evaluation captures various contexts.
    - **Paragraph:** *"The company launched its new software with features like data encryption and automatic backups. Pricing information is available, but there's no mention of customer support options."*
    - **Questions:**
        1. "Does it mention pricing?"
        2. "Is customer support discussed?"
        3. "Does it talk about data encryption?"
    - **Ground Truth:**
        - Q1: Yes
        - Q2: No
        - Q3: Yes

2. Evaluation Criteria

Metrics Defined:

1. **Accuracy:**
    - Compare the LLM's Yes/No answer with the ground truth.
2. **Explanation Quality:**
    - Assess how well the reasoning aligns with the paragraph content.
3. **Ambiguity Handling:**
    - Identify whether the model addresses unclear or implied topics logically.
    - For example, if "customer support" is not explicitly mentioned, does the model say "No" if "customer support" is not stated clearly in the text or does it simply state that it is not present?

3. Comparison Framework



Step-by-Step Method:

1. **Test Setup:**
    - Input the same paragraph and questions into both Model A and Model B.
2. **Response Collection:**
    - Record the Yes/No answers and any reasoning provided by each model.
3. **Manual Scoring:**
    - Compare the Yes/No answers to the ground truth.
    - Assign scores for explanation quality and ambiguity handling.
4. **Quantitative Analysis:**
    - Calculate the accuracy for each model.
        - Model A Accuracy: 3/3 = 100\%
        - Model B Accuracy: 2/3 =66.7\%
5. **Qualitative Analysis:**
    - Evaluate reasoning clarity and logical consistency.
    - Observing whether outputs vary across multiple runs to assess reliability.
6. **Disagreement Resolution:**
    - Flag and review disagreements or subjective cases (e.g., unclear phrasing in the paragraph like *"but there's no mention of customer support options."*).

4. Human Review

- **Role of Reviewers:** Resolve disputes and assess subjective outputs where models disagree.
- **Guidelines for Reviewers:**
    1. Mark responses as Correct, Incorrect, or Ambiguous based on the paragraph.

2. If reasoning is provided, evaluate it for relevance and logic.
3. Use consistency as a tie-breaker if models are equally accurate.

5. Example Scoring

Outputs:

- **Model A:** "Yes", "No", "Yes"
- **Model B:** "Yes", "No", "No"
- **Ground Truth:** "Yes", "No", "Yes"

Scoring Table:

Question	Ground Truth	Model A	Model B	Correctness (A)	Correctness (B)
Pricing Mentioned	*Yes*	*Yes*	*Yes*	*Correct (1)*	*Correct (1)*
Customer Support	*No*	*No*	*No*	*Correct (1)*	*Correct (1)*
Data Encryption	*Yes*	*Yes*	*No*	*Correct (1)*	*Incorrect (0)*
Accuracy	--	***3/3 (100%)***	***2/3 (66.7%)***	--	--

6. Recommendation

- **Model A is recommended** based on higher accuracy (100% vs. 66.7%).