Research Scientist Intern - Assignment Round 2

Domain Selection

Choose exactly one domain:

- Healthcare
- Finance
- Ecommerce

Tasks

1) Literature Research

- Identify and review recent papers that discuss failure modes and inefficiencies of frontier AI models (LLMs, multimodal models, agentic systems) in your chosen domain.
- “Failure modes” can include but are not limited to: hallucinations, brittleness to distribution shift, unsafe recommendations, tool-use failures, bias, calibration errors, spurious reasoning, privacy leakage, prompt sensitivity, reward hacking, long-context degradation, or evaluation blind spots.

2) Failure Modes Report

Create a structured report that includes:

- A taxonomy of observed failure modes.
- For each category:
    - What the failure looks like in real workflows
    - Why it happens (mechanisms and contributing factors)
    - Impact severity (what breaks, who is affected, how often)
    - Evidence from literature (what the papers show)

3) Data and Training Techniques to Improve Specific Failures

Pick a failure mode from your taxonomy and propose improvement techniques via appropriate data and training strategy. Do a mini write up on:

- Proposed data strategy (what data, how to collect/label/curate)
- Proposed training strategy (eg: SFT, RLHF, etc)

- Supporting evidence from above

4) Data Quality: Considerations and Quality Parameters

For data strategy in Section 3, specify:

- Data requirements and acceptance criteria (quality parameters)
- Quality checks and measurement plan

Deliverables

Submit a single Google Doc (ensure it has view access to anyone with the link) with abovementioned details.

Submission Link

- File name: `ResearchScientistIntern_Assignment_<YourName>_<Domain>.pdf`
- Google Form - https://docs.google.com/forms/d/e/1FAIpQLSdQvbMJnCk2tImn3clk2-qw4oyXbMK8Rus1rSUURMECujFXVw/viewform?usp=publish-editor