Data Science Take-Home Exam: Financial Advisor LLM Evaluation

Submission: *Python file(s) + a brief report (1-2 pages)*

Context

*You are the **Lead Evaluator** for a specialized LLM used by Financial Advisors during live client calls. The model retrieves client portfolio data and answers questions in real-time.*

Your role: *Design the Evaluation Harness that determines whether a new model candidate is deployable (Go/No-Go decision). You cannot train or modify the model, only evaluate it.*

The Business Problem

*We are considering replacing our current model (**Model A**) with a new model (**Model B**).*

Behavior	***Model A***	***Model B***
Refusal Rate	*30% ("I'm not sure")*	*5%*
Hallucination Rate	*2%*	*6%*
Average Latency	*800ms*	*400ms*
Confidence Calibration	*Conservative (rarely >80%)*	*Aggressive (often 95%+)*

The Legal Constraints (from Compliance):

Failure Mode	*Business Cost*	*Rationale*
Hallucination *(wrong numbers)*	***$1,000,000***	*SEC fine + client lawsuit + reputation damage*
Unjustified Refusal *(silent when data exists)*	***$50,000***	*Advisor looks incompetent, client churns*
Justified Refusal *(silent when data unavailable)*	*$0*	*Correct behavior*
Correct Answer	*$0*	*Expected behavior*

Key Insight: *The cost ratio is **20:1**. One hallucination is as costly as twenty unjustified refusals.*

Data Quality Note

Upon inspecting the production logs, you discover:

- ***18% of Model A's "refusals"*** *have the response text: "I cannot provide personalized financial advice. Please consult with your advisor directly."*
  *This is a **compliance-mandated refusal**—legally required when the query requests forward-looking advice (e.g., "Should I buy AAPL?"). These are **not** capability failures.*
- *The remaining **82% of refusals** are capability-based (model uncertain or data unavailable).*

You must account for this distinction in your evaluation design.

Part 1: Metric Design

1.1 The Performance Score Formula

*Design a single scalar **Performance Score S** to rank model candidates.*

Requirements:

- *S must incorporate the **20:1 cost asymmetry** between hallucinations and refusals*
- *S must distinguish between **compliance refusals** (acceptable) and **capability refusals** (costly)*
- *S should be normalized to [0, 1] where higher is better*
- *Show the mathematical formula with clear variable definitions*

Deliverable:

- *The formula for S*
- *A brief explanation (3-5 sentences) of why this formulation captures business value*

1.2 The "Overconfidence Penalty"

Analysis shows that Model B frequently outputs confidence scores >95% even when hallucinating. When an advisor tells a client "I'm absolutely certain your returns were 12.4%" and it's actually 8.1%, the trust damage is catastrophic—far worse than a low-confidence error.

***Your task:** Modify your score S to include an **Overconfidence Penalty** with these properties:*

1. *No penalty if confidence ≤ 0.9 OR the answer is correct*
2. *Penalty applies when confidence > 0.9 AND the answer is a hallucination*
3. *The penalty must be **non-linear** (exponential or polynomial) because trust damage accelerates with confidence level*
4. *The penalty should be parameterized (not hardcoded magic numbers)*

Deliverable:

Also provide:

- *A brief justification for your choice of non-linear function*
- *Example outputs at confidence levels: 0.85, 0.92, 0.96, 1.0*

Part 2: Regression Analysis

You ran your evaluation harness on a held-out test set of 10,000 examples:

Metric	***Model A***	***Model B***
Your Score S	*0.82*	*0.84*
Accuracy	*68%*	*89%*
Hallucination Rate	*2%*	*6%*

The VP of Engineering says: "Model B scores higher. Ship it."

You're not convinced.

2.1 Transition Analysis

*The aggregate score hides **behavioral transitions**. Define and compute:*

***R_unsafe:** The rate at which previously "safe" behaviors become "unsafe."*

Specifically:

- ***Safe behavior:** Model A refused (no risk of harm)*
- ***Unsafe behavior:** Model B hallucinated (potential $1M liability)*

*A query where Model A refused but Model B hallucinated represents a **catastrophic regression**—we replaced a safe (if annoying) behavior with a dangerous one.*

Deliverable:

***Important:** A compliance refusal (Model A) that becomes a hallucination (Model B) is **extra concerning**—Model B is hallucinating on queries it shouldn't even attempt.*

Also answer in your report:

- *What threshold for R_unsafe would cause you to reject Model B? Justify your number.*

- *How does the compliance refusal distinction affect your analysis?*

2.2 Slice-Level Regression

*Simpson's Paradox: Model B could score better overall while performing **worse** on critical subgroups.*

You have access to query metadata:

- `query_type`*: ["portfolio_value", "transaction_history", "tax_info", "forward_looking", "fee_inquiry"]*
- `complexity`*: ["simple", "moderate", "complex"]*
- `data_availability`*: ["full", "partial", "none"]*

Deliverable:

Also answer in your report:

- *Propose a specific hypothetical scenario where Model B's higher overall score masks a critical regression. Be concrete.*

2.3 The Go/No-Go Recommendation

***In your report**, provide a 1-page recommendation to the VP of Engineering:*

1. ***Your recommendation:** Ship Model B, reject Model B, or conditional approval*
2. ***Key evidence:** Which 2-3 metrics most informed your decision?*
3. ***Risk quantification:** What is the expected annual cost difference between the models, assuming 500,000 queries/year?*
4. ***Conditions (if applicable):** If conditional approval, what specific improvements or guardrails would you require?*

If anything is unclear, state your assumptions explicitly in your submission. We evaluate your reasoning, not just your answer.