**1. Research Questions & Hypotheses**

1. **Primary question:**
   * “How many adversarial turns does each model endure—starting from a correct answer—before its first failure?”
2. **Secondary question:**
   * “Do semantic shifts in the adversarial prompts (or in the evolving context) systematically accelerate or delay failures?”
3. **Hypotheses:**
   * H₁: Models differ significantly in their median “time‑to‑failure” under adversarial turns.
   * H₂: Larger semantic drift between prompts (or between context and prompt) increases failure hazard.
   * H₃: The Model × Prompt₀ interaction explains additional variance beyond main effects.

**2. Units, Outcomes & Censoring**

* **Unit of analysis:** one multi‑turn conversation.
* **Outcome variable:**
  + **Count version (for Poisson/NB):** Y\_i = # adversarial turns until first incorrect answer
  + **Time‑to‑event version (for survival):** round index *k* at first failure, with conversations that never fail treated as right‑censored at max\_round.
* **Censoring:**
  + If the model stays correct through the final allowed adversarial turn, mark as censored.

**3. Predictors & Covariates**

**A. Core Predictors**

1. **Model (categorical)**
2. **Prompt₀ identifier**
3. **Model × Prompt₀** interaction

**B. Semantic, Time‑Varying Covariates**

1. **Prompt‑to‑prompt drift**
   * Cosine distance between embeddings of successive adversarial prompts.
2. **Context‑to‑prompt drift**
   * Cosine distance between an embedding of the accumulated context (Prompt₀ + all prior prompts & responses) and the new prompt.
3. **Cumulative drift**
   * Aggregate (e.g. rolling average or cumulative sum) of the above distances up to each round.
4. **Optional extras:**
   * **Prompt complexity:** token‑length, syntax tree depth, rarity of words.
   * **Response confidence:** include the model’s self‑reported confidence each round as a time‑varying covariate.

**4. Data Structuring & Splits**

* **Static (conversation‑level) table** for count models: one row per conversation with aggregated covariates (e.g. average drift).
* **Long (turn‑level) table** for survival: one row per conversation × round, with time‑varying covariates.
* **Train/validation/test split:**
  + **Stratify** by Prompt₀ (to ensure diverse initial prompts in each fold).
  + Typical split: 70/15/15 or use k‑fold cross‑validation at conversation level.

**5. Modeling Strategies**

1. **Count regression:**
   * **Poisson** if mean≈variance; **Negative Binomial** if overdispersion.
   * Predictor set: Model, Prompt₀, Model×Prompt₀, plus aggregated semantic covariates.
2. **Survival analysis (Cox PH):**
   * Discrete rounds *k* as time.
   * Include time‑varying covariates (drift, confidence).
   * Consider stratification by Model or random effects (frailty) for Prompt₀.
3. **Mixed‑effects extension:**
   * Random intercepts for Prompt₀ (to capture unobserved prompt difficulty).
   * Random slopes for drift to model heterogeneity across prompts.

**6. Evaluation Metrics & Statistical Tests**

* **Count model:**
  + Goodness‑of‑fit: AIC/BIC, deviance residuals.
  + Predictive accuracy: RMSE of predicted vs actual count.
  + Overdispersion test.
* **Survival model:**
  + Concordance index (C‑index).
  + Schoenfeld residuals to test proportional hazards assumption.
* **Hypothesis tests:**
  + Likelihood‑ratio tests for adding semantic covariates or interaction terms.
  + Wald tests on individual coefficients.

**7. Ablation & Robustness Checks**

1. **No semantic features** vs **with semantic features**
2. **Different context‐aggregation schemes** (mean vs attention‐weighted)
3. **Alternative embedding models** (e.g. SBERT vs Universal Sentence Encoder)
4. **Sensitivity to censoring threshold** (e.g. drop convos with fewer than 3 adversarial turns)

**8. Visualization & Interpretation**

* **Kaplan–Meier curves** by model and by high/low semantic drift groups.
* **ICR plots**: predicted vs actual failure counts by model.
* **Partial effect plots**: hazard ratio as a function of prompt‑to‑context distance.
* **Interaction heatmap**: Model × Prompt₀ effect sizes (e.g. IRRs).

**9. Practical Considerations**

* **Sample size per prompt₀**: ensure enough conversations for each initial prompt to estimate Prompt₀ effects reliably.
* **Multiple‑testing correction** if you examine many prompts or covariates.
* **Reproducibility**: fix random seeds for embeddings, splits; document software versions.