Understood—let's craft this as a truly **original research proposal**, not a mere "project." I'll outline a **full academic research plan** with depth, rigor, and novelty. We'll include:

1. **Title & Abstract**
2. **Introduction & Motivation**
3. **Research Questions & Hypotheses**
4. **Literature Review & Gap Analysis**
5. **Methodology**
    - Data and Domain
    - Model Architectures
    - Explanation Generation
    - Evaluation Framework (quantitative and qualitative)
    - Statistical Analysis Plan
6. **Experimental Design**
    - Controlled comparisons
    - Ablation studies
7. **Ethical Considerations & IRB**
8. **Expected Contributions & Novelty**
9. **Work Plan & Timeline (Gantt)**
10. **Required Resources & Budget**
11. **Risks & Mitigations**
12. **References**

---

1. Title & Abstract

Title

> *Assessing Faithfulness and Utility of LLM-Generated Explanations for Black-Box Clinical Prediction Models*

Abstract

We propose a quantitative and qualitative study comparing GPT-based explanations to SHAP for a diabetes risk model. We formulate explicit hypotheses about faithfulness (feature-level agreement), comprehensibility (user rating), and clinical utility (decision impact), and evaluate them using statistical tests in a controlled user study. Our contributions include: (1) a reproducible evaluation framework, (2) novel metrics combining faithfulness and utility, and (3) guidelines for deploying LLM explanations in clinical practice.

---

2. Introduction & Motivation

- **Context**: Black-box AI in healthcare risks misdiagnosis and legal liability.
- **Problem**: LLMs promise natural-language explanations, but their **faithfulness to model logic** and **clinical utility** remain untested.
- **Goal**: Rigorously evaluate whether LLM explanations can replace or augment feature-importance methods in real clinical workflows.

---

3. Research Questions & Hypotheses

1. **RQ1**: Do LLM explanations mention the same key features as SHAP?
    - **H1**: ≥80% feature overlap (quantified via Jaccard similarity) between LLM and top-5 SHAP features.
2. **RQ2**: Are LLM explanations more comprehensible to clinicians?
    - **H2**: Mean comprehension rating (1–7 Likert) for LLM > SHAP by at least 1 point (paired-t test, $\alpha=0.05$).
3. **RQ3**: Do LLM explanations improve diagnostic decisions?
    - **H3**: Clinicians using LLM explanations achieve ≥5% higher accuracy on held-out cases vs. SHAP (McNemar's test).

---

4. Literature Review & Gap Analysis

- **Existing XAI**: SHAP, LIME, Integrated Gradients—quantitative but terse.
- **LLM-XAI Surveys**: Bilal et al. (2025)—call for domain-specific, quantitative studies.
- **Gap**: No studies have measured **clinical decision impact** of LLM explanations.

---

5. Methodology

5.1 Data and Domain

- **Dataset**: UCI Pima Indians Diabetes (768 records; age, BMI, glucose, blood pressure, etc.)
- **Preprocessing**: Standard scaling, missing-value imputation with k-NN.

5.2 Model Architectures

- **Black-box**: Random Forest (200 trees) and XGBoost (grid-search hyperparameters)
- **Baseline Explanations**: SHAP (TreeExplainer)
- **LLM Explanations**: GPT-4 via OpenAI API, prompt-engineering to elicit feature-centric narratives.

5.3 Explanation Generation

- **Prompt Template**:

```
"Model prognosis: Diabetes=Yes for patient with features {…}.
In plain language, explain why."
```

- **Sampling**: Generate 3 explanation variants per case; choose the most coherent via log-prob ranking.

5.4 Evaluation Framework

- **Faithfulness**:
  - Extract features mentioned by LLM (NLP-based keyword matching + named-entity recognition).
  - Compute Jaccard similarity vs. top-5 SHAP features (per case).
- **Comprehensibility**:
  - Recruit N=10 clinicians; each rates 20 explanations on clarity, usefulness, conciseness (1–7 scale).
- **Clinical Utility**:
  - Clinician decision tasks: Given model score + explanation, decide "treat" vs. "monitor."
  - Measure accuracy and decision confidence.

5.5 Statistical Analysis

- **Feature overlap**: one-sample t-test against 0.8.
- **Comprehension ratings**: paired-t test LLM vs. SHAP.
- **Decision accuracy**: McNemar's test (paired proportions).

---

6. Experimental Design

- **Within-subjects**: Each clinician sees both LLM and SHAP explanations on different—but matched—cases.
- **Ablation**: Vary prompt detail level, LLM temperature, and SHAP feature counts.

---

7. Ethical Considerations & IRB

- **Human subjects**: Clinician survey; IRB approval required.
- **Data privacy**: Synthetic or de-identified data.

- **Bias mitigation**: Check demographic balance in dataset and clinician pool.

---

8. Expected Contributions & Novelty

1. **Framework**: A reproducible pipeline for LLM vs. SHAP evaluation in healthcare.
2. **Metrics**: Combined faithfulness-utility score.
3. **Guidelines**: Best practices for prompt design in clinical XAI.

---

9. Work Plan & Timeline

Phase	Weeks 1–2	Weeks 3–4	Weeks 5–6	Weeks 7–8	Weeks 9–10
Data prep & model train	■				
SHAP & LLM setup		■			
Clinician study design		■			
User study & data collect			■		
Analysis & writing				■	■

---

10. Resources & Budget

- **Compute**: GPU instance (AWS/GCP) ~ $300
- **LLM API**: $200 credit
- **Participant compensation**: 10 clinicians × $50 = $500
- **Total**: ~$1,000

---

11. Risks & Mitigations

Risk	Impact	Mitigation
Low clinician participation	Delayed study	Recruit via professional networks; increase compensation
LLM hallucinations	Invalid results	Use prompt-validation; manual spot checks
Data imbalance	Biased findings	Re-sample or augment minority cases

---

12. References

1. Bilal, A., Ebert, D., & Lin, B. (2025). *LLMs for Explainable AI: A Comprehensive Survey*. ACM TIST.
2. Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. *NeurIPS*.
3. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). “Why Should I Trust You?” *KDD*.

Next Steps

- I can convert this into a **formal PDF proposal** with properly formatted sections, figures, and Gantt chart.
- Or I can dive directly into **notebook code** to set up data, models, and SHAP.

Let me know how you'd like to proceed—and we'll turn this into **publishable research**.