Cancer Foundation Model 구축을 위한 실행 계획 (To-Do List)

프로젝트 목표 🎯

멀티오믹스(Tabular)와 병리 영상(Image) 데이터를 통합하여, **설명 가능한(Explainable)** 암 예후 예측 파운데이션 모델을 구축한다. 모델은 최종적으로 예측에 대한 **판단 근거를 자연어 텍스트와 시각적 히트맵으로 제시**해야 한다.

Phase 1: 데이터 준비 및 전처리 (Data Preparation & Pre-processing)

이 단계는 모델의 성능을 좌우하는 가장 중요한 기반 작업입니다.

- ☑ **1-1. 멀티모달 데이터 다운로드 및 정제**
  - **Multi-omics 데이터:** TCGA, CPTAC 등 공개 데이터베이스에서 범암종(pan-cancer) 유전자 발현 데이터 및 관련 임상 정보(생존 기간, 암종 등)를 다운로드합니다.
  - **병리영상 데이터:** 동일 환자군의 전체 슬라이드 이미지(Whole-Slide Image, WSI)를 다운로드합니다.
- ☑ **1-2. [핵심] Multi-omics 특성 공학 (Feature Engineering)**
  - **Cox 회귀분석 수행:** 다운로드한 학습 데이터셋을 사용하여, **각 암종별로 모든 유전자에 대해** 3년 또는 5년 생존을 엔드포인트로 설정하고 Cox 비례위험 회귀분석을 실행합니다.
  - **회귀계수 룩업 테이블 생성:** 분석을 통해 얻은 [유전자 x 암종] 별 회귀계수(coefficient)를 별도의 룩업 테이블로 저장합니다. 통계적 유의성을 위해 FDR 보정 등을 적용하는 것을 권장합니다.
  - **최종 입력 테이블 생성:** 각 환자(샘플)에 대해 [유전자1_발현량, 유전자1_해당암종계수], [유전자2_발현량, 유전자2_해당암종계수], ... 와 같은 쌍으로 구성된 최종 입력 테이블을 생성합니다.
- ☑ **1-3. 병리영상 데이터 전처리**
  - **WSI 패치 분할:** 고해상도의 WSI를 Swin Transformer에 입력할 수 있도록 작은 크기의 패치(patch)들로 분할합니다.

Phase 2: 단일 모달리티 모델 개발 (Unimodal Model Development)

각 데이터의 특성을 이해하고 베이스라인 성능을 확보하기 위해 개별 모델을 먼저 훈련합니다.

- ☑ 2-1. Multi-omics 모델 (Tabular Transformer) 훈련
  - **목표:** 5 년 생존 여부 분류 (Classification)
  - **입력:** Phase 1 에서 생성한 최종 입력 테이블
  - **아키텍처:** Tabular Transformer
  - **결과:** 학습된 모델과 Multi-omics 데이터만 사용했을 때의 예후 예측 성능(AUC 등)을 확보합니다.
- ☑ 2-2. 병리영상 모델 (Swin Transformer) 훈련
  - **목표:** 5 년 생존 여부 분류 (Classification)
  - **입력:** Phase 1 에서 생성한 이미지 패치
  - **아키텍처:** Swin Transformer
  - **핵심: ROI 정보 없이(ROI-free)**, 이미지 전체 레이블(생존/사망)만으로 훈련하여 모델이 스스로 중요 영역을 학습하게 합니다.
  - **결과:** 학습된 모델과 병리 영상만 사용했을 때의 예후 예측 성능(AUC 등)을 확보합니다.

Phase 3: 멀티모달 융합 및 LLM 파인튜닝 (Multimodal Fusion & LLM Fine-tuning)

프로젝트의 최종 목표인 '설명 가능한 AI'를 구현하는 단계입니다.

- ☑ 3-1. [핵심] 추론 텍스트 데이터셋 구축
  - **목표:** LLM 을 파인튜닝하기 위한 (멀티오믹스, 병리이미지) -> (전문가 추론 텍스트) 쌍의 데이터셋을 생성합니다.
  - **방법 (택 1 또는 혼합):**
    - **템플릿 기반 생성:** "유전자 OOO 의 회귀계수가 높아 위험 요인으로 작용하며, 이미지상 OOO 패턴이 관찰됩니다." 와 같은 규칙 기반 템플릿으로 자동 생성 후 전문가가 감수합니다.
    - **LLM 활용:** GPT-4 와 같은 강력한 LLM 에게 데이터 요약 및 초기 추론 생성을 요청하고, 전문가가 이를 검토 및 수정하여 효율을 높입니다.
- ☑ 3-2. 융합 아키텍처 설계 및 구현
  - **임베딩 추출:** Phase 2 에서 훈련된 각 Transformer 모델의 최종 예측 레이어 직전에서 고차원 임베딩 벡터를 추출합니다.
  - **프로젝션 레이어 구현:** 추출된 두 임베딩 벡터를 결합(concatenation)한 뒤, LLM 의 토큰 임베딩 공간으로 차원을 맞춰주는 프로젝션 레이어(간단한 신경망)를 구현합니다.
- ☑ 3-3. 공개 LLM 선정 및 파인튜닝
  - **모델 선정:** 프로젝트에 적합한 공개 LLM(예: Llama 3, Qwen2 등)을 선정합니다.

- **파인튜닝:** [융합 임베딩] 과 텍스트 프롬프트를 입력으로, 추론 텍스트를 출력으로 하여 3-1 에서 구축한 데이터셋으로 LLM 을 파인튜닝합니다.

Phase 4: 모델 평가 및 시각화 (Model Evaluation & Visualization)

최종 모델의 성능을 검증하고 사용자가 이해할 수 있는 형태로 결과를 제시합니다.

- ☑ 4-1. 최종 LLM 모델 평가
  - **정량 평가:** LLM 이 생성한 텍스트에서 최종 예측(생존/사망)을 파싱하여 정확도(AUC 등)를 측정합니다.
  - **정성 평가:** 의료 전문가가 생성된 추론 텍스트의 임상적 타당성, 논리성, 유창성을 평가합니다.
- ☑ 4-2. [핵심] 설명 가능성(XAI) 시각화 구현
  - **어텐션 맵 시각화:** 훈련된 Swin Transformer 에서 어텐션 스코어를 추출하여, 원본 병리 이미지 위에 **히트맵(Heatmap)으로 시각화**하는 기능을 구현합니다.
  - **최종 결과물:** 새로운 환자 데이터 입력 시, LLM 이 생성한 **추론 텍스트**와 병리 이미지 위에 표시된 **어텐션 히트맵**을 함께 제공합니다.