○ **(**공동연구개발과제_서울대**):** 뇌인지기능 **AI** 파운데이션 모델 개발 및 사업화 기획

전체 개요

3년 내에 우리는 멀티모달 뇌영상, 뇌파, 그리고 유전체 데이터를 인간 뇌의 단일 언어 정렬 표현으로 융합하는 1,550억 매개변수 규모의 통합 파운데이션 모델인 'BOM(Brain-Omics Model)'을 개발하고자 함. 먼저 뇌영상, 뇌파, 그리고 멀티오믹스 각 데이터 유형에 특화된 파운데이션 모델을 개발하고, 각 모델을 LLM과 정렬함. 마지막으로 모든 모달리티 파운데이션 모델을 "LLM의 '의미론적 다리(semantic bridge)'"를 이용하여 결합하여 최종적인 뇌-오믹스-언어 파운데이션 모델을 구축할 예정임. 최종 모델은 유전 정보에서부터 뇌 구조, 신경 역학, 행동에 이르는 전 과정을 연결하는 최초의 파운데이션 모델이 될 것임. 아래는 본 연구팀의 "네 가지 세부 목표"와 "세부목표 달성을 위한 기술적 전략 및 위험 관리 방안"임. 본 연구팀은 세부 목표와 연구 전략 수립에 있어서 다음의 **2024**년 스탠포드 대학 파운데이션 모델 센터에서 제시한 **"**성공적 파운데이션 모델의 **5**대 요소**"**를 고려했음 **(**표 **1).**

표 **1.** 성공적 파운데이션 모델의 **5**대 속성				
표현력	확장성	다중 모달리티	합성성	기억 능력
유연한 방식으로 정보를 추출 나타내는 능력	대량 데이터를 효율적으로 처리하는 능력	다양한 도메인의 데이터를 이해하고 연결하는 능력	요소를 결합하여 전체적 의미를 유추**,** 새로운 환경에 일반화하는 능력	학습한 지식을 저장**,** 활용하는 능력

각 속성과 우리의 연구 전략 **–** 근거와 강점		
5대 속성	달성 전략	근거와 연구팀 강점
1.표현력	뇌영상, 뇌파, 유전체 데이터 특성에 최적화된 독자적 아키텍처(SwiFT, DIVER, Hyena 등)를 적용하여 정보 손실 없이 유연하게 패턴 포착.	세계 최초 고차원 (4D) 뇌영상 트랜스포머 등 검증된 모델 기술력 보유
2.확장성	초거대 모델의 안정적 병렬 학습 및 최적화 기술(μTransfer 등)을 적용하여, 방대한 데이터를 효율적으로 처리하고 모델 규모를 지속적으로 확장.	최고 수준의 계산 자원(Aurora 슈퍼컴퓨터 등) 및 100억 파라미터 이상 모델의 성공적 학습/확장 경험
3. 다중 모달리티	LLM을 '공통 의미론적 허브'로 활용, Cross-attention 어댑터 등 최신 기법으로 뇌, 유전체 등 이종(異種) 데이터의 표현을 정렬 및 통합.	세계 최대 규모의 멀티모달 코호트 데이터를 기반으로 한 융합 학습
4. 합성성	각 모달리티별 전문가 모델(MoE)을 동적으로 조합하여, 데이터와 지식 기반의 새로운 추론 생성하고 새로운 테스크에 일반화.	세계 최대 규모의 통합 모델(1,550억+ 파라미터)을 통해 관찰되지 않은 조합에 대한 창발적 추론 능력 확보

5. 기억 능력	방대한 의학 문헌 및 임상 종단 데이터를 학습하고, 최신 검색 증강(RAG) 및 강화학습(RLHF) 기법으로 맥락에 맞는 지식의 인출 및 활용을 극대화.	국내 고유의 장기 추적 데이터(GARD)를 통해, 시간에 따른 변화를 기억하고 예측하는 실용적 지능 구현

성공적 파운데이션 모델의 속성				
표현력	확장성	다중 모달리티	합성성	기억 능력
유연한 방식으로 정보를 추출 나타내는 능력	대량 데이터를 효율적으로 처리하는 능력	다양한 도메인의 데이터를 이해하고 연결하는 능력	요소를 결합하여 전체적 의미를 유추, 새로운 환경에 일반화하는 능력	학습한 지식을 저장, 활용하는 능력

본 연구팀은 뇌영상, 뇌파, 유전체 데이터 각각의 특성에 최적화된 독자적인 아키텍처(SwiFT, DIVER, Hyena 등)를 도입하여, 정보 손실 없이 데이터 내의 미세한 패턴을 유연하게 포착할 수 있는 뛰어난 표현력을 갖추고자 함. 이미 연구팀은 세계 최초로 고차원(4D) 뇌영상 Transformer 모델을 개발한 경험을 보유하여 이 방면에서의 기술력을 이미 검증받은 바 있음. 또한 초거대 모델의 안정적인 병렬 학습 및 최적화 기술(μTransfer 등)을 통해 방대한 규모의 데이터를 효율적으로 처리할 전략을 수립함. 본 연구팀은 Aurora 슈퍼컴퓨터 등 최고 수준의 계산 자원을 활용하여 이미 100억 파라미터 이상의 모델을 성공적으로 학습하고 확장한 경험을 확보하고 있어 확장성 측면에서도 강점을 가짐. 뿐만 아니라, 본 연구팀은 LLM을 '공통 의미론적 허브'로 활용하여 Cross-attention 어댑터 등 최신 기법을 기반으로로 뇌, 유전체와 같은 이종 데이터를 효과적으로 정렬하고 통합할 수 있는 역량을 보유하고 있음. 세계 최대 규모의 멀티모달 코호트 데이터를 활용한 융합학습을 통해, 데이터와 지식의 새로운 조합을 만들어내는 합성성 역시 뛰어나며, 특히 1,550억 개 이상의 파라미터를 가진 세계 최대 규모의 통합 모델을 통해 창발적 추론 능력까지 확보함. 마지막으로, 방대한 의학 문헌과 임상 종단 데이터를 학습하여 최신의 검색 증강(RAG) 및 강화학습(RLHF) 기법으로 지식의 활용을 극대화하고자 함. 구체적으로 국내 장기 추적 데이터(GARD)를 기반으로 시간에 따른 변화를 정확히 기억하고 예측하는 실용적 지능 구현 가능성이 매우 높다고 전망됨.

- 세부목표 **1. [**멀티모달 뇌영상 **FM] LLM**과 정렬된 멀티모달 뇌영상 파운데이션 모델 구축
- 세부목표 **2. [**멀티모달 뇌파 **FM] LLM**과 정렬된 뇌파 파운데이션 모델 구축
- 세부목표 **3. [**멀티오믹스 **FM] LLM**과 정렬된 멀티오믹스 **FM**
- 세부목표 **4. [**뇌**-**멀티오믹스 **FM] LLM**과 정렬된 통합 멀티모달 뇌**-**멀티오믹스 **FM** 구축
- 세부목표 달성을 위한 기술적 전략 및 위험 관리 방안

각 세부목표 하의 마일스톤과 추진 체계는 아래와 같음 **(**표 **2)**.

표 **2.** 세부목표와 마일스톤

Yr	세부목표	마일스톤 (M)	팀*	모델명
Y1	멀티모달 뇌영상 FM	M1.1. 멀티모달 뇌영상 750억 (75B) + MoE FM	서울대-차지욱, 문태섭, 이재윤	Swift 2.0
Y2		M1.2. 멀티모달 뇌영상-LLM 정렬 920억 (92B) FM		Swift 2.5
Y1	멀티모달 뇌파 FM	M2.1. 멀티모달 뇌파 FM 330억 (33B) + MoE FM	서울대-차지욱, 문태섭, 이재윤	Diver 1.0
Y2		M2.2. 멀티모달 뇌파-LLM 정렬 500억 (50B) FM		Diver 1.5
Y3-5		M2.3. 멀티모달 뇌영상-뇌파-LLM 1250억 (125B) FM		NeuroX
Y1	멀티오믹스 FM	M3.1. 유전체 200억 (20B) FM 개발	서울대-주윤정, 이재윤, 바이오넥서스	OMNIA-G (genomic)
Y2		M3.2. 멀티오믹스 Expert 300억 (30B) FM 개발		OMNIA-K (korean tuned)
Y2		M3.3. 멀티오믹스-LLM 정렬 470억 (47B) FM		OMNIA-X
Y3	뇌-멀티오믹스 FM	M4.1.멀티모달 뇌-멀티오믹스 통합 LLM 정렬 1550억 FM 학습	서울대-차지욱, 문태섭, 이재윤, 주윤정 바이오넥서스	BOM 1.0 (Brain-Omics Model)
Y4-5		M4.2. 인지예비능을 위한 GARD 종단 데이터의 파인튜닝 학습 (인지예비능, 혈액 마커, 임상 데이터) 및 모델 경량화		BOM 1.5 (Brain-Omics Model)
*해외 협업 기관: 브룩헤이븐국가연구소 (Shinjae Yoo), NVIDIA (이름찾아서넣기), AWS				

표 **3.** 마일스톤별 모델 규모, 과제, 난관 및 대응 전략 요약

연차	마일스톤	모델 크기 (MoE)	데이터 모달리티	핵심 다운스트림 과제	예상 난관 및 대응 전략
1	M1.1 멀티모달 뇌영상	75B (MoE 8×)	휴지기 fMRI, sMRI, dMRI, PET	인지 관련 변수 예측 (인지예비능, 인지상태 등)	MoE 훈련 불안정성 → 비-MoE 모델 훈련
1	M2.1 멀티모달 뇌파	33B (MoE 8×)	EEG, iEEG, MEG, ECoG	다양한 EEG 과제 성능 5% 향상	스케일링 법칙 붕괴 → 더 어려운 사전학습 과제 적용
2	M1.2 멀티모달 뇌영상-LLM 정렬	92B (MoE 8×, 17B LLM)	M1.1 뇌영상 + 과제 기반 fMRI + 텍스트	다중 모달 인지 과제 (HCP/HBN 정확도, 모달리티 식별, QA)	모달리티 인코더의 MoE 수렴 실패 → MoE 제거

1,2	M2.2 멀티모달 뇌파–LLM 정렬	50B (MoE 8×, 17B LLM)	멀티모달 뇌파 + 텍스트	크로스 모달 성능 향상	짝지어진 데이터 부족 → 외부 데이터 추가 확보 및 비디오 캡셔닝 활용
2	M2.3 멀티모달 뇌영상–뇌파–LLM 정렬	125B (MoE 16×, 17B LLM)	멀티모달 뇌영상 + 뇌파 + 텍스트	한 모달리티로 다른 모달리티 생성 과제, 인지상태 관련 복잡한 QA	모델 규모 증가로 인한 학습 어려움 → 슈퍼컴퓨터 전문 팀과 함께 코드 최적화
1	M3.1 유전체 FM	20B (MoE 6x)	T2T-CHM13, GRCh38, ENCODE, EPIC등 genomics data	Enhancer activity 예측, Chromatin 상태 예측, Gene expression 예측	긴 DNA 시퀀스 context → Hyena operator 도입
2	M.3.2 멀티오믹스 Expert FM	30B (MoE 8x)	GARD omics, 혈액바이오마커, cognitive score data	치매/MCI/HC 분류–인지점수 예측 (MMSE 등)	이질적 omics 통합 (sparsity, noise, scale mismatch) → Omics별 Encoder(MoE) 설계 및 contrastive fusion 기반 표현 통합 전략 활용
3	M.3.3 멀티오믹스–LLM 정렬	47B (MoE 8x) + LLM 17B	M3.2 오믹스 데이터 + 텍스트	인지기능 저하 관련 유전학 지식 QA 평가	멀티오믹스–LLM 양방향 정렬 실패 혹은 불균형 학습 → 안정적인 방향 선택 후 순차적 정렬 진행
4, 5	M4.1, 뇌–멀티오믹스 FM	155B (MoE>24)	모든 뇌영상/뇌파/멀티오믹스, 관련 문헌, 인지평가 등의, 텍스트	쌍을 이루지 않은 데이터에 대한 제로샷 교차 모달 예측 성능 검증 (예: 뇌영상 데이터로 오믹스 지표 예측)	수렴 실패 → 커리큘럼 기반 미세조정
4, 5	M4.2 FM 활용 인지예비능 응용 파인튜닝	155B (MoE>24)	모든 뇌영상/뇌파/멀티오믹스, 관련 문헌, 인지평가 등의, 텍스트	LLM–그래프 기반 인지예비능 예측/솔루션 레포트 생성, 모달리티 조합별 성능 비교 평가/검증	LLM으로 생성한 인지예비능 관련 레포트 Hallucination 문제 → 다양한 RAG 실험 및 임상 전문가의 평가 기반 강화학습 추가

세부목표 **1. [**멀티모달 뇌영상 **FM] LLM**과 정렬된 멀티모달 뇌영상 파운데이션 모델 구축

목표**.** 본 연구의 첫 목표는 거대 언어 모델(LLM)과 연계된 920억 (92B) 파라미터 규모의 멀티모달 뇌영상 파운데이션 모델을 개발하는 것임. 이 모델은 기존 거대 언어 모델에 내재된 일반 지식과 신경과학적 지식을 활용하여, 뇌의 구조적 및 기능적 표상을 포착하는 것을 목표로 함. 연구 초기 단계에는 본 연구진이 개발한 fMRI 트랜스포머의 파라미터 규모를 확장하는 것부터 시작할 계획임. 이 과정에서 8개의 전문가(expert)로 구성된 소형 전문가 혼합(Mixture-of-Experts, MoE) [A5] 시스템을 도입하고, 이를 통해 구조적 및 기능적 MRI 데이터를 통합할 수 있는 거대 규모 모델을 개발할 것임 (그림 1). 그 후, 전문가 혼합(MoE) 시스템의 학습 불안정성을 해결하기 위해 제안된 최신 기술 [A6-7]을 적용하여

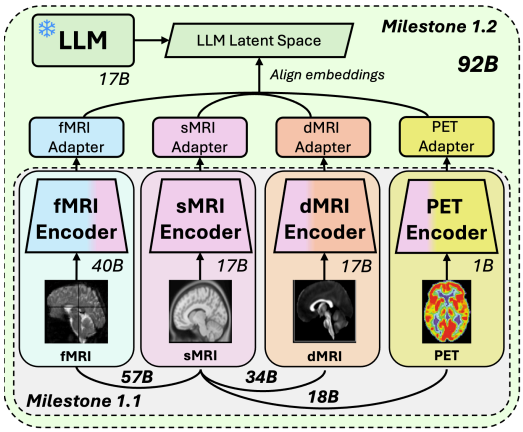


그림 1. 뇌의 구조와 기능 정보를 자연어로 해석하고 추론하는 LLM과 정렬된 멀티모달 뇌영상 파운데이션 모델. 이 모델은 sMRI, dMRI, fMRI, 그리고 PET를 포함한 네 가지 모달리티를 전용 인코더로 학습하여 통합한다. 각 모달리티 데이터는 교차어텐션 어댑터를 통해 LLM과 정렬된다. 각 모달리티에 가장 적합한 '전문가 모델(expert)'을

아키텍처를 170억에서 400억 파라미터 규모까지 확장하고, 더 넓은 범위의 뇌 기능적 역동성을 포괄하기 위해 과제 기반 기능적 자기공명영상(task-fMRI) 데이터를 추가할 계획임. 최종적으로, 뇌영상 모델은 모달리티 어댑터 모듈(modality adapter modules)을 통해 거대 언어 모델과 정렬되며, 이를 통해 뇌영상 데이터의 표상이 언어적 개념에 기반을 두도록 정립될 것임 **(**표 **4)**.

근거**.** 현재 신경과학 분야의 핵심 과제는 뇌의 해부학적 구조(구조적 MRI), 백질 연결성(확산 MRI), 시공간적 역동성(기능적 MRI), 그리고 뇌인지기능 저하의 병리적 정보(PET)를 통합적으로 처리하는 모델을 개발하는 것임. 기존 신경과학 분야의 사전 훈련 모델들은 뇌구조와 같은 정적인 단일 모달리티에만 집중하거나 데이터의 특징을 과도하게 축소하여, 뇌의 거시적/미시적 구조와 기능을 아우르는 포괄적인 그림을 제시하지 못하는 한계가 있음 [A8-12]. 특히, 현재까지 대규모 수준에서 전처리나 중간 가공 없이, 다양한 뇌영상 모달리티를 직접 통합하여 성공적으로 학습하는 단일 end-to-end 모델은 달성된 바 없음[25].

준비성**.** 본 연구팀은 이전 연구를 통해 기능적 MRI (fMRI)의 시퀀스 길이에 대해 거의 선형적인 계산 복잡도를 유지하도록 설계된 최초의 4D 트랜스포머인 SwiFT를 개발함 [A13]. 본 연구진은 SwiFT를 88억(8.8B) 파라미터까지 확장하여 성능을 증가시킬 수 있음을 확인했으며, 미국 에너지부 산하의 슈퍼컴퓨터 Frontier에서 93%의 GPU 활용률을 달성했음. 또한 4만 5천 명의 fMRI 데이터에 대한 사전 학습을 통해 SwiFT 모델의 스케일링 법칙(scaling laws)을 입증함 [A14] **(**그림 **2.A)**. 특히, 다양한 모델 규모에서 훈련 안정성을 향상시키기 위해, 강력한 하이퍼파라미터 튜닝 기법인 µTransfer [A15]를 도입하여 튜닝 비용을 90% 절감했음. 본 연구팀이 활용한 µTransfer는 최대 41억(4.1B) 파라미터 모델 규모까지 활성화 벡터 크기(activation vector norms)를 안정시켰으며 **(**그림 **2.B,** 하단**)**, 이는 표준 훈련 방식에서 나타나는 발산(divergence) 현상 **(**그림 **2.B,** 상단**)**을 방지하는 효과를 가져왔음. 또한, 미국 에너지부 산하의 슈퍼컴퓨터 Aurora에서 256개 노드를 사용하여 41억 파라미터 모델을 성공적으로 훈련했으며, 25TB의 데이터에 대해 각 에폭(epoch)을 13분 만에 완료하는 주목할 만한 성능을 달성했음. 이는 DeepSpeed-ZeRO-2 [A16], bfloat16 정밀도 [A17], 그리고 Flash-Attention 2 [A18]와 같이 학습 효율성을 높이는 기술들을 통해 가능했음.

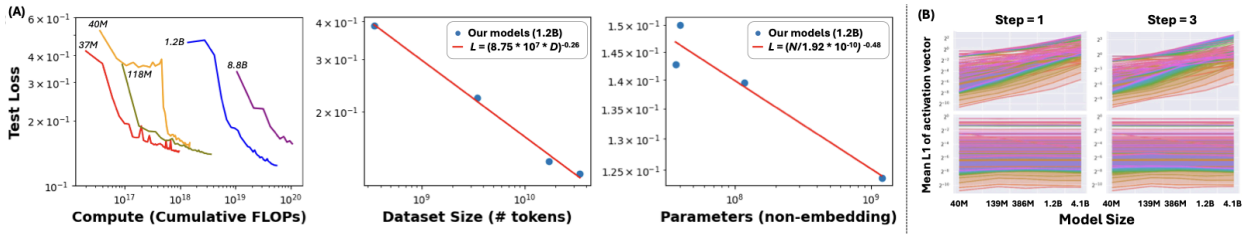


그림 **2.** 대규모 훈련의 위험 완화**:** 검증된 스케일링 법칙과 훈련 안정성**.** (A) 예측 가능한 성능 확장성과 (B) 견고한 훈련 안정성을 보여줌. **(A)** 신경 스케일링 법칙 **(Frontier):** fMRI Swin Transformer 모델(SwiFT)의 성능은 더 많은 컴퓨팅 자원, 데이터, 파라미터를 사용할수록 예측 가능하게 향상됨 [15]. 이는 모델을 확장하려는 본 연구의 과학적 타당성을 입증함
(B) µTransfer를 이용한 훈련 안정성 **(Aurora):** 대규모 모델 훈련의 핵심 난제는 불안정성이임 상단은 표준 훈련 방식에서 발생하는 활성화 벡터 발산(activation vector divergence) 현상을 보여줌. 반면, 하단은 µTransfer [A15]의 구현이 다양한 모델 규모에 걸쳐 훈련을 안정시키는 것을 보여주며, 이는 본 프로젝트 성공의 핵심적인 전제 조건임.

표 **4.** 세부목표 **1** 및 마일스톤 요약

마일스톤	핵심 평가 태스크
1년차 - M1.1 멀티모달 뇌영상 750억 (75B) + MoE FM 개발	- 인지능력 평가를 위한 UKB, HCP intelligence 예측 $R^2 \geq 0.05$

2년차 - M1.2 멀티모달 뇌영상-LLM 정렬 920억 (92B) FM	- 멀티모달 인지 과제 (ADNI AD conversion 예측, modality ID, QA)

마일스톤 **1.1 (1**년차**):** 멀티모달 뇌영상 **750**억 **(75B) + MoE** 파운데이션 모델 개발

본 세부목표의 첫 마일스톤은 네 가지 핵심 뇌영상 모달리티에 걸친 거시적 해부 구조(구조적 MRI;sMRI), 미세구조적 연결성(확산 MRI;dMRI), 시공간적 역동성(기능적 MRI;fMRI), 그리고 뇌인지기능 저하 정보(PET) 전반에 걸쳐 통합적, 종단간(end-to-end) 학습을 수행하는 750억 (75B) 파라미터 규모의 파운데이션 모델을 개발하는 것임.

이를 위한 전략은 구조적 MRI(sMRI)를 통합의 구심점(natural anchor)으로 활용하는 것임. sMRI는 일반적으로 기능적 MRI(fMRI) 및 확산 MRI(dMRI)와 함께 촬영된다는 점에서, sMRI를 중심으로 멀티모달 모델의 학습을 위한 쌍(paired) 데이터셋을 구성할 수 있음. 모든 뇌영상 모달리티를 다룰 수 있는 복잡하고 어려운 과정에 앞서, 본 연구팀은 우선적으로 뇌 구조와 기능의 결합(coupling)을 포착하기 위한 **f/sMRI FM**, 거시적 해부 구조와 미세구조적 연결성을 잇는 **s/dMRI FM**, 그리고 뇌 구조 전반과 뇌인지기능 정보를 함께 포착할 수 있는 **sMRI/PET FM**을 개발할 예정임. 본 연구진은 세 모델을 각각 sMRI-fMRI 데이터 쌍, sMRI-dMRI 데이터 쌍, 그리고 sMRI-PET 데이터 쌍을 활용해 학습할 계획이며, InfoNCE 손실함수 [A19] 기반의 대조학습을 통해 파운데이션 모델들을 사전학습하고자 함. 또한 이를 통해 사전학습된 멀티모달 뇌영상 모델들이 개개인들의 공통적인 뇌 표상 뿐만 아니라 개인 특징적인 뇌 표상들을 동시에 학습할 수 있도록 하고자 함. 모델의 아키텍처 측면에서, 이 모델들은 fMRI 인코더로는 SwiFT 아키텍처를, sMRI/dMRI/PET 인코더로는 비전 트랜스포머(Vision Transformer) [A20]를 사용할 것임. 한발 더 나아가, 각 모델들을 효율적으로 대규모의 파라미터를 가진 모델로 확장하기 위해, 트랜스포머 구조에 전문가 혼합(MoE) 레이어를 통합할 것임. 사전 학습된 파운데이션 모델들의 다운스트림 평가를 위해서는, 사전 학습 후 동결된(frozen) 인코더에서 추출된 임베딩을 경량 선형 레이어(lightweight linear layer)를 통해 융합할 것임. 전문가 혼합(MoE) 방식을 통한 대규모 모델로의 확장은 모델 학습 과정에서의 불안정성이 크기 때문에 위험 부담이 있는 계획이므로, 개발 과정에서의 위험 완화를 위해, 전문가 혼합 방식을 활용하지 않는 기본 트랜스포머를 대체 옵션으로 준비해 둘 것임.

마일스톤 **1.2 (2**년차**):** 멀티모달 뇌영상**-LLM** 정렬 **920**억 **(92B)** 파운데이션 모델 개발

본 마일스톤의 목표는 교차 어텐션 어댑터 모듈 [A21-23]을 활용하여 LLM 중심의 다중모달리티를 결합하는 데에 있으며, 특히 서로 쌍을 이루지 않는 데이터들을 결합할 수 있는 하나의 모델을 개발하는 것에 중점을 두고자 함. 이를 통해 여러 MRI 모달리티와 PET 데이터로부터 얻어진 뇌의 구조 및 기능 전반에 대한 표상을 텍스트의 형태로 변환할 수 있도록 하고자 함. 구체적으로, 마일스톤 1.1의 사전 훈련된 뇌영상 인코더와 오픈소스 LLM을 기반으로 하여, 각 인코더로부터 추출된 MRI 모달리티별 임베딩을 LLM의 잠재 공간으로 정렬시키는 어댑터 모듈을 추가하고 이 어댑터 모듈을 훈련시켜 텍스트와의 정렬을 수행함. 이때 모달리티의 유형을 파악할 수 있도록 텍스트 프롬프트를 구성하고 학습에 활용함으로써, LLM의 파라미터를 추가적으로 튜닝하지 않고 효율적으로 GPU 자원을 관리하며 학습을 진행하고자 함. 특히, 멀티모달 LLM에서의 성능이 모달리티별 인코더의 스케일링에 비해 LLM의 스케일링에 크게 영향을 받는다는 최신의 연구들에 근거하여 [A22-24], LLM과 결합된 뇌영상 FM 또한 LLM의 규모를 확장함에 따라 성능이 증가하는지를 검증하고자 함.

다음으로, LLM과 결합된 멀티모달 뇌영상 FM의 개발 및 확장 가능성을 검증한 이후, 전문가 혼합(MoE) 레이어를 적용하여 효율적으로 모델의 규모를 확장하고자 함. 특히, 최신 연구 [A7]에서 그 가능성이 입증된 교차 어텐션을 활용하는 어댑터 기반의 멀티모달 LLM에서 전문가 혼합(MoE) 방식을 활용하고자 함. 뿐만 아니라, 전문가 혼합(MoE) 방식으로 구현된 거대 규모의 모델의 경우 학습시에 불안정하여 모델이 적절히 학습되지 않는 문제가 있는데, 이를 해결하기 위해 공유 전문가와 특화 전문가를 구분하는 라우팅 방식 [A25], 전문가들 간의

정보의 부하를 편향 가중치를 활용해 조절하는 편향 기반 부하 분산 [A26], 학습 시 전문가들을 큰 비율로 드롭아웃 시키는 방식 등의 최신의 테크닉들을 활용하여 모델의 학습 안정성을 확보하고자 함. 거대 규모 모델 학습 과정에서의 불안정함으로 인해 생길 수 있는 마일스톤의 리스크를 완화하기 위해, 만약 훈련 불안정 문제가 발생할 경우 위험 관리 방안으로 모달리티별 인코더에서 전문가 혼합(MoE) 레이어를 제외할 계획임.

현재 다중모달리티 MRI/PET 모델을 위한 벤치마크 데이터셋이 없기 때문에, 본 연구진들은 LLM과 결합된 NeuroX-MRI/PET의 평가를 위해서 데이터셋을 구축할 계획임. 특히 치매 진단 정보 외에도 다양한 설문조사 및 의료 정보와 인구통계학적 정보들을 통합한 텍스트 데이터셋을 구축함으로써, 본 연구 뿐만 아니라 추후에 연구 및 개발될 다중 모달리티 모델의 평가 및 학습에도 사용될 수 있도록 하고자 함. 새롭게 구축한 데이터셋을 활용하여, 마일스톤 1.1에서 개발된 모델들이 10년 뒤의 치매 발병 위험에 대해 예측한 수치와 질의 응답을 통해서 멀티모달 뇌영상 FM이 생성한 발병 위험 수치를 비교하는 정량 평가를 진행할 계획임. 한발 더 나아가, 구조 MRI 이미지로부터 치매를 진단하고 메디컬 리포트를 작성하는 오픈소스 모델 [A27-28]들과 멀티모달 뇌영상 FM의 질의응답 및 추론 결과를 비교하는 정성 평가를 진행할 예정이다. 특히, 구조 MRI 뿐만 아니라, 확산 MRI와 기능 MRI 및 PET 데이터를 동시에 활용함으로써 발생하는 모델의 창발적 능력에 대한 검증을 중심으로 정성 평가를 진행할 계획임. 이는 치매 발병의 위험성에 대한 수치 예측에서 한발 더 나아가, 다중 모달리티 데이터를 활용하는 것이 개개인에 대한 더욱 풍부한 정보를 제공함으로써 보다 정교한 개인별 솔루션을 제공하는 지에 대한 검증과 평가를 가능하게 해줄 것이라고 기대함.

세부목표 **2. [**뇌파 **FM] LLM**과 정렬된 뇌파 파운데이션 모델 구축

목표**.** 본 연구는 인간의 전기생리학적(electrophysiology) 뇌파 멀티모달 데이터를 통합하고 이러한 신호를 LLM에 정렬하는 뇌파 파운데이션 모델의 구축을 목표로 함. 1년차에는 10B 규모의 전문가 혼합(MoE-8) 파운데이션 모델을 훈련하고, 다양한 응용 분야에서 최고의 성능을 달성할 것임. 2년차에는 직접 신경-텍스트 방법을 사용하여 뇌파와 LLM이 정렬된 뇌파 파운데이션 모델을 구축할 예정이며, 세부목표 1에서 개발한 멀티모달 뇌영상 파운데이션 모델과 결합해 LLM과 정렬된 뇌영상-뇌파 파운데이션 모델을 개발할 것임.

근거**.** 선형 회귀에서 파운데이션 모델까지의 꾸준한 진전에도 불구하고, 현존하는 뇌파 파운데이션 모델은 여전히 세 가지 한계를 겪고 있음: (i) 시공간 모델링 제약, (ii) 불충분한 규모, (iii) 개별 작업에 국한된 파인 튜닝을 넘어 인간의 자연스러운 행동 및 인지 과정과의 정렬이 부재. 이러한 문제들을 해결하면 의미 있는 인지예비능 측정과 인지 기능 저하의 조기 탐지를 위한 임상적 활용성을 높이고, 실시간 뇌 신호 해석을 통해 인지 기능 변화를 효과적으로 모니터링하는 데 기여할 수 있을 것으로 예상됨.

준비성**.** 우리 팀은 이미 뇌파 트랜스포머 모델 **“DIVER-0”**을 개발하여 기존 EEG 모델링의 중요한 한계를 해결했음. **DIVER-0** [A29] (그림 **3**)은 10%의 사전 훈련 데이터 만으로 감정 인식 벤치마크에서 현재 SOTA [A30]를 +4.1pp 앞섰음.

표 **5.** 세부목표 **2** 및 마일스톤 요약

마일스톤	핵심 평가 태스크
1년차 - M2.1 멀티모달 뇌파 MoE 330억 FM 개발	- SOTA 모델 대비 뇌파 기반 다운스트림 과제 성능 5% 이상 향상 - 모델 규모 확장(Scaling)에 따른 성능 변화 분석 (예: 1B, 4B, 33B 모델 간 성능 비교) - 다양한 벤치마크 데이터셋에서의 범용 성능(Generalization) 검증

1~2년차 - M2.2 멀티모달 뇌파-LLM 정렬 500억 FM 개발	- EEG-텍스트 정렬(Alignment) 성능 정량 평가 - 작업 간/사이트 간 일반화(Cross-task/Cross-site Generalization) 성능 검증 - 언어적 지시(Instruction) 기반 제로샷(Zero-shot) 과제 수행 능력 평가
2년차 - M2.3. 멀티모달 뇌영상-뇌파-LLM 정렬 1250억 FM 개발	- 교차 모달(Cross-modal) 추론 능력 검증 - 통합 멀티모달 질의응답(Multimodal QA) 성능 평가 - 통합으로 인한 창발적 능력(Emergent Abilities) 검증

마일스톤 **2.1 (1년차):** 전문가 혼합 기반 **(Mixture of Expert)**의 **330**억 뇌파 파운데이션 모델 개발

뇌파 파운데이션 모델 분야에서는 규모의 확장이 성능을 향상시킨다는 일관된 증거가 있음에도 불구하고 현재까지 개발된 모델들은 여전히 충분한 규모 확대가 이루어지지 않음 [A31-33]. 현재 가장 큰 모델조차 1TB 규모의 데이터만 활용했으며 [A30], 컴퓨팅 자원의 활용 또한 수백 노드시간 수준으로 제한되어 있음 [A32, A34].

본 연구에서는 이러한 문제를 극복하기 위해, 그림 **3**에 있는 뇌파 트랜스포머 모델 DIVER [A29]의 체계적인 규모 확대를 수행할 것임. 먼저 소규모 모델 단계에서 최적의 모델 구조를 찾고, 모델 크기 효율성을 극대화하기 위해 전문가 혼합(MoE) 구조를 적용하고, one-shot scaling을 위해 μTransfer를 활용할 예정임. 이후 모델을 1B에서 4B, 최종적으로 10B 크기로 점진적으로 확대할 것임. 이러한 단계적 접근법은 안정적인 성능 개선을 보장하고 다양한 다운스트림 애플리케이션에 적합한 최적 모델 크기를 식별할 수 있게 함.

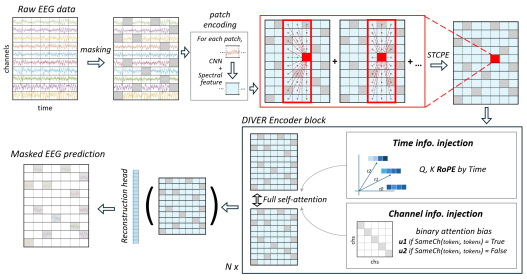


그림 3. 본 연구팀이 개발한 뇌파 FM 구조. EEG 신호를 패치로 변환하여 트랜스포머로 학습. 채널 위치 정보는 우리가 개발한 STCPE (Sliding Temporal Conditional Positional Embedding)를 통해 패치에 위치 정보 부여함. 시공간 역학의 학습을 위해서 변환된 임베딩은 RoPE와 이진 어텐션 바이어스를 사용하는 트랜스포머 인코더 블록에서 처리함. 탑 AI 국제 학회 ICML 2025 출판 (Han et al, 2025)

본 마일스톤의 성공 여부는 EEG 데이터셋을 이용한 평가에서 본 모델의 성능이 현재 최신 모델 [A30] 대비 중간값 기준 5% 이상 향상될 때로 정의함. 공정하고 재현 가능한 평가를 위해 모든 비교는 torcheeg 라이브러리 [A35]를 활용하여 수행되며, 제안된 신규 벤치마크를 공개하여 연구 커뮤니티에 다시 기여할 것임.

마일스톤 **2.2 (1~2년차):** 뇌파**-LLM** 정렬 **500**억 파운데이션 모델 개발

본 연구에서는 현재의 뇌파 파운데이션 모델이 갖추지 못한 진정한 작업 간 일반화(cross-task generalization) 능력을 달성하기 위해, 신경 신호를 언어의 풍부한 의미적 맥락에 연결하고자 함. 이를 위해 향상된 직접 신경-텍스트 기법을 사용하여 , 신경 신호를 해당 텍스트 설명에 직접 매핑하고, 명령어 기반 조정(instruction tuning)을 통해 뇌파 파운데이션 모델을 미세 조정할 것임 [A33, A36]. 최종 모델의 성공은 미국과 한국의 다중 사이트 및 다중 문화 벤치마크를 통해 일반화 가능성을 검증하며, 양국의 신경과 전문의들과 협력하여 평가할 예정임.

마일스톤 **2.3 (2년차):** 멀티모달 뇌영상**-**뇌파**-LLM 1250**억 파운데이션 모델 개발

다양한 모달리티의 뇌영상과 전기생리학 데이터를 통합하는 모델 개발에 있어서 핵심적인 어려움은 쌍을 이루지 않은 데이터를 사용하여 멀티모달 모델을 학습하는 것임. 본 연구팀은 세부목표 1과 2에서 활용했던 어댑터 기법을 확장 적용하여 이 문제를 해결하고, 1,250억

파라미터 규모의 다중모달 전문가 혼합(MoE) 모델을 훈련시키고자 함. 즉, 여러 모달리티의 데이터들을 하나의 공유된 잠재 공간인 LLM에 정렬함으로써, 뇌영상 파운데이션 모델과 뇌파 파운데이션 모델을 통합할 것임. 이때 모델 학습의 안정성을 보장하기 위해 3단계의 학습 과정을 진행할 예정임. 첫 단계에서는 각 모달리티별 인코더 및 거대 언어 모델의 파라미터는 바꾸지 않은 채로, 각 모달리티별 어댑터의 모듈만을 학습시킴. 두번째 단계에서는 전문가 혼합(MoE) 시스템 안에 각 모달리티에 특화된 전문가들을 추가하고, 전문가 혼합(MoE) 시스템의 학습을 진행함. 최종 단계에서는 QLoRA(Quantized Low Rank Approximation) [A37]을 활용하여 LLM을 파인튜닝하는 학습을 진행함.

본 연구팀에서 개발하고자 하는 멀티모달 MRI-전기생리학-언어 모델은 기존 연구들에서는 진행하지 않았던 도전적인 과제로, 이 모델의 성능을 평가하기 위한 벤치마크 데이터셋이 존재하지 않음. 이에 본 연구팀은 오픈 소스로 공개된 MRI-EPhys 데이터쌍을 기반으로, 모델의 교차 모달(cross-modal) 확장성과 별도의 사전학습 없이 바로 적용 가능한 추론 능력을 평가가하기 위한 멀티모달 MRI-전기생리학-언어 통합데이터셋을 구축하고자 함. 특히 새롭게 구축되는 데이터셋은 다양한 과제에서의 모델의 성능을 복합적으로 평가할 수 있도록, **(1)** 데이터 모달리티 식별**, (2)** 뇌·행동·인구통계학적 데이터 통합을 통한 치매 발병 징후 탐지**, (3)** 개인의 임상적 위험도 프로파일링**, (5)** 기존 과학 문헌과의 연계에 기반한 복합적인 추론**, (6)** 모달리티 간 교차 예측 과제(하나의 모달리티 데이터로부터 개인의 다른 모달리티 데이터의 특징을 예측) 등의 다양한 과제로 구성할 계획임.

서로 쌍을 이루지 않는 모달리티의 데이터를 통합하는 대규모 모델을 개발하는 것은 상당히 도전적인 과제임. 따라서 본 연구팀은 위험 부담을 완화하고 성공적인 결과물을 만들기 위해, 거대 언어 모델(LLM)을 중심으로 MRI와 전기생리학 데이터를 단계적으로 통합하며 각각의 모달리티가 LLM과 교차 모달리티 추론이 가능한지 검증할 계획임. 마일스톤 기반의 평가와 결합된 이러한 단계적 접근법은, MRI와 전기생리학 그리고 텍스트 데이터의 완전한 통합이 실현 불가능한 것으로 판명되더라도 두 개의 선도적인 뇌 파운데이션 모델이라는 결과물을 보장할 것임.

세부목표 **3. [**멀티오믹스 **FM] LLM**과 정렬된 멀티오믹스 파운데이션 모델 개발

표 **7.** 세부목표 **3** 및 마일스톤 요약

마일스톤	핵심 평가 태스크
1년차 - M3.1 200억 (20B) 유전체 파운데이션 모델 개발	- 유전자 기능 예측 및 유전자 조절 요소(Enhancer, Promoter) 활성 예측 AUROC > 0.85 - 유전자 발현량(Gene Expression) 예측 $R^2 > 0.5$
2년차 - M3.2 300억 (30B) 멀티오믹스 expert 파운데이션 모델 개발	- 한국인 치매 위험군 분류 (Dementia Risk Stratification) - GARD 데이터 기반 치매/경도인지장애/정상군 분류 정확도 90% 이상 - MMSE 등 주요 인지 기능 점수 예측 오차 10% 이내
3년차 - M3.3. 470억 (47B) 멀티오믹스-LLM 정렬 파운데이션 모델 개발	- 설명 가능한 유전 변이 영향 분석 (Explainable Variant Impact Analysis) - 특정 유전 변이의 치매 발병 기여도에 대한 LLM의 설명 생성 - 생성된 설명에 대해 RAG로 검색된 근거 문헌의 일치도 평가 (F1 score)

마일스톤 **3.1 (1**년차**): 200**억 **(20B)** 유전체 파운데이션 모델 개발

목표**.** 진화 보존성 및 enhancer 문맥을 반영한 long-context DNA foundation model 개발.

본 연구의 궁극적인 목표는 다양한 모달리티의 오믹스 데이터를 통합적으로 학습하고 이를 기반으로 **genomic input**에 대한 예측 태스크를 수행할 수 있는 범용 **Genomic Foundation Model**을 구축하는 것임. 유전체(genome)는 전사 조절 요소, 후성유전체(epigenome), 스플라이싱(splicing), 보존성(conservation) 등 다양한 기능적 정보를 포함하고 있으며, 이러한 기능은 조직 특이적이며 복잡한 상호작용을 통해 조절됨. 이러한 다양한 생물학적 정보를 단일 구조의 모델로 모두 포괄하려고 하면, 데이터 모달리티 간의 형태, 길이 등의 차이로 인해 학습 효율성과 성능 저하 문제가 발생함. 따라서 각 데이터 유형별로 최적화된 Expert를 독립적으로 구성하고, 입력 특성에 따라 선택적으로 활성화하는 **Mixture-of-Experts (MoE)** 구조를 도입하고자 함.

MoE 구조는 다양한 특화 Expert를 병렬적으로 유지하면서, 각 입력에 대해 가장 관련성이 높은 일부 Expert만 선택적으로 활성화(예: Top-2 Routing)하므로 대규모 파라미터 모델의 연산 효율을 높이고 확장 가능성을 확보할 수 있음. 특히 본 모델은 긴 유전체 서열을 효율적으로 처리하고, 다운스트림 테스크에 범용적으로 적용 가능한 유전체 표상을 학습하는 것을 목표로 함.

본 단계에서 치매 관련 데이터는 사용되지 않으며, 전적으로 유전체 정보를 중심으로 한 사전학습에 집중함. 이로써 향후 치매 위험 예측 및 variant interpretation을 위한 Multi-Omics Expert Model 구축의 기반을 마련하고자 함.

핵심 전략**.**

1. 기존 오믹스 파운데이션 모델에서 사용된 적 없는**, Long-read sequencing** 기반의 **Homo sapiens (human) genome assembly T2T-CHM13** 최초 사용**:** 기존 Genomics foundation model 등에서 사전학습 데이터로 널리 사용된 GRCh38 [A38] 은 오랫동안 인간 유전체 연구의 표준 reference 로 사용되어 왔지만, (1) Centromere, telomere, acrocentric arm 등 유전체의 약 8%가 누락되어있고, (2) rDNA cluster 가 미포함되는 등 뇌기능 및 노화 관련 유전자의 조절과 관련된 주요영역이 빠져있으며, (3) heterochromatin 의 정보가 부재하고 유전적 다양성 표현이 제한적이라는 단점이 존재하였음. 2022년 발표된 T2T-CHM13 (CHM13v2.0) [A39] 은 세계 최초의 gapless

인간 reference genome 으로 long-read NGS 기술을 이용하여 조립된 reference 로 상기 한계를 해소함. 이전에는 빠져있던 rDNA, centromeric repeat, segmental duplication 등이 포함되어 유전자 발현 구조를 더 정확하게 모델링 가능하고, 뇌 조직에서 특이적으로 작동하는 epigenomic marker-rich region (brain-specific enhancer 영역) 등을 포함 및 locate 가능하다는 장점이 있기에, pretraining 데이터로 사용할 경우 인지능력 및 치매에 관여할 수 있는 고난이도 non-coding 조절 영역까지 학습에 포함시키는 전략적 선택이 될것으로 사료됨.

2. 다종 유전체 기반 학습 전략**:** 인간 유전체뿐 아니라 다양한 종(species)의 유전체 데이터를 사전학습에 활용할 계획임. DNA 언어 모델은 단일 종의 유전체만으로는 포착할 수 없는 보편적 문법과 기능적 패턴—예컨대 enhancer, promoter, 전사인자 결합 모티프(TFBS) 등—이 진화적으로 보존되어 있다는 점에 착안하여, 다양한 계통군에 걸친 유전체 데이터로 부터 강건하고 일반화된 표상을 학습할 수 있음. 특히 대표적인 동물 모델인 **mouse, chimpanzee, zebrafish, Drosophila** 등은 신경발달, 인지 기능, 유전자 조절 기전에서 중요한 모델 종이며, 이들의 유전체를 포함시킴으로써 종간 enhancer activity 예측 성능을 개선하고, 인간 중심 fine-tuning 단계에서의 적응력을 높일 예정임. 이러한 다종 유전체 기반 학습은 sequence redundancy와 기능적 보존성을 효과적으로 활용할 수 있어, 인간에서 관측된 치매 관련 regulatory signature의 표현 가능성을 확장시키는 데 기여할 것으로 기대됨.

3. 모델구조**:** 학습 구조 차원에서 long-read 염기서열의 특성과 문맥 길이를 최대한 활용할 수 있도록 다양한 전략을 적용함 (**dynamic window cropping**, **relative positional encoding**, **masked span prediction)**. 특히 100kb 이상의 long-context를 안정적으로 처리하기 위해 장문처리에 특화된 **Hyena operator** [A40] 및 **Mixture-of-Experts** 기반 **sparse attention** [A41] 구조를 도입예정임. 반복 영역과 centromeric region 등에서 발생하는 sequence redundancy는 customized masking으로 처리하며, Nanopore 기반 long-read는 주로 **gap-filling** 또는 **reference-untethered context augmentation** 용도로 사용될 예정임. 학습 데이터셋 내 read 간 positional overlap을 제한하여 representation leakage를 방지하고, HiFi 중심의 backbone representation은 학습 후 embedding으로 고정하여 다운스트림 테스크에 전이학습 가능하도록 설계함.

사전학습 데이터**.**
본 유전체 파운데이션 모델의 사전학습은 유전체의 (1) 구조적 기반, (2) 기능적 조절, (3) 진화적 보존이라는 세 가지 핵심 축을 중심으로 설계된 데이터를 활용함. 이는 모델이 단편적인 정보를 넘어, 유전체의 다층적이고 복잡한 문맥을 통합적으로 이해하도록 하기 위함임.

1. 구조적 기반 데이터 **(Structural Foundation Data)**
   유전체 서열의 기본 문법과 구조를 학습하기 위한 핵심 데이터.

   - 인간 참조 유전체 (Human Reference Genome):
     - (차별점) T2T-CHM13: 세계 최초의 완전한 인간 참조 유전체를 주요 백본(backbone)으로 사용함. 이는 기존 모델들이 접근하지 못했던 반복 서열, 중심절 등 뇌 기능과 관련된 중요 유전체 영역까지 학습에 포함시켜 모델의 구조적 완성도를 극대화함.
     - GRCh38: 기존 연구와의 호환성 및 방대한 주석(annotation) 정보 활용을 위해 보조적으로 사용함.
   - 다양한 생물군 유전체 (Diverse Genomes): 박테리아, 바이러스, 고세균 등 다양한 생물군의 참조 유전체(GTDB, IMG/VR 등)를 학습하여, 생명체의 보편적인 유전체

구조와 비정형적 문법에 대한 모델의 일반화 성능을 확보함.

2. 기능적 조절 데이터 **(Functional Regulation Data)**
   유전자가 언제, 어디서, 어떻게 발현되는지를 결정하는 조절 기전을 학습하기 위한 데이터임.

   - 전사체 및 비번역 RNA: 유전자 발현의 기본 산물인 전사체(NCBI GTF)와 다양한 기능성 비번디오 RNA(Ensembl, RNAcentral)의 구조를 학습하여, 유전자의 기능적 다양성을 모델링함.

   - 후성유전체 (Epigenome): 유전자 발현의 핵심 스위치인 DNA 메틸레이션(EPIC, WGBS), 전사인자 결합 부위(ENCODE TFBS), 인핸서/프로모터(Roadmap, FANTOM5) 등 대규모 실험 데이터를 통합 학습하여, 조직 특이적인 유전자 조절 네트워크를 이해함.

   - 스플라이싱 (Splicing): RNA 스플라이싱 패턴(ENCODE, PSI)을 학습하여, 단일 유전자에서 다양한 단백질이 생성되는 복잡한 과정을 모델에 반영함.

3. 진화적 보존 데이터 **(Evolutionary Conservation Data)**
   진화의 시간을 관통하여 보존된 핵심 기능 영역을 식별하고 그 중요도를 학습하기 위한 데이터임.

   - 종간 보존 서열: 여러 종에 걸쳐 보존된 염기 서열 정보(phastCons, phyloP)를 활용함. 이를 통해 모델이 기능적으로 중요한 핵심 조절 부위를 높은 정밀도로 식별하는 능력을 갖추게 하며, 이는 특히 인간 유전체에서 기능이 알려지지 않은 영역의 역할을 추론하는 데 결정적인 단서를 제공함.

모델 아키텍쳐**.**
본 모델은 6-8개의 Expert로 구성된 MoE 구조 [A42]를 기반으로 하며, 전체 파라미터 수는 약 200–300억 개 수준으로 설계할 예정임. 각 Expert의 사전학습을 위해 활용 가능한 데이터셋, 커버하는 생물학적 기능 영역이 충분히 분화되어 있는지, 그리고 현실적인 구현 가능성을 고려했을 때 6-8개가 최적이라고 판단함.

각 Expert는 유전체의 서로 다른 생물학적 기능을 모델링하도록 설계되며, 입력 데이터는 Gating Network를 통해 적합한 Expert를 선택받음. 다음 표 **8**과 같이 각 Expert는 자신에게 해당하는 데이터 모달리티와 task에 대해 개별적으로 사전학습되며, 각기 다른 모델 블록이 사용됨. 예를 들어 Splicing Expert에는 **Caduceus** 모델의 **Bi-Mamba** 구조 [A43]를 적용해 엑손-인트론 경계 부위를 양방향으로 정밀하게 학습할 수 있도록 하며, **Hyena** [A40]는 긴 서열 처리를 위한 hierarchical attention 구조를 가지는 모델로, DNA 또는 보존성 시퀀스 학습에 적합함.

표 **8.** 유전체 파운데이션 모델의 **Expert** 구축을 위한 데이터 및 모델 블록 정리

Expert	주요 데이터	모델 블록
DNA Sequence Expert	T2T-CHM13, GRCh38	Hyena
TFBS Expert	ENCODE, JASPAR	CNN
Promoter/Enhancer Expert	FANTOM5, Roadmap	CNN

Methylation Expert	EPIC, WGBS	CNN
Splicing Expert	ENCODE, PSI	Mamba [A44]
Conservation Expert	phastCons, phyloP	Hyena

사전학습.
본 모델은 GShard [A45]의 **Top-2 Gating with Random Routing** 전략을 도입해 각 유전체 토큰을 두 개의 expert에 조건부로 할당함. 이때 **hard expert capacity constraint**를 적용해 Expert별로 고정된 토큰 수를 초과하는 입력은 잔차 경로로 처리함으로써 정적 텐서 구조를 유지하고 대규모 분산 학습을 가능하게 함. 또한 **auxiliary load balancing loss**를 도입해 학습 중 Expert 간 토큰 분포의 불균형을 완화해 각 Expert가 전체 입력에서 균형 있게 학습되고 호출되도록 함. 또한 하나의 입력 서열은 메틸화가 일어날 확률 예측과 전사인자의 결합 여부 예측 등 여러 예측 과제에 동시에 사용될 수 있음. 이처럼 여러 개의 예측 태스크를 동시에 학습하는 것을 Multi-task Learning이라 하며, 각 태스크별 loss function을 정의하고 이를 가중 합산한 **multi-task loss**를 통해 모델을 최적화함으로써 태스크 간 정보 공유와 일반화 성능 향상을 도모함. 이를 통해 다양한 오믹스 데이터에 걸친 범용적 표현 학습이 가능해져 유전체 기반 기능 예측의 정확도와 효율성을 향상할 수 있음.

본 모델은 향후 유전 변이(variant)의 기능적 영향을 예측하는 데 사용될 수 있어야 함. 이를 위해 사전학습 단계에서 **Variant-aware Contrastive Learning**을 적용할 계획임. 해당 기법은 정상(reference) 서열과 기능 변화를 유도하는 변이 벡터를 한 쌍으로 입력해, 모델이 어떤 변이가 기능을 바꾸는지를 벡터 간 거리 차이로 감지하도록 유도함. 따라서 모델은 실제 생물학적 기능의 민감도를 반영하게 되어 variant effect prediction 성능을 향상시킬 수 있을 것임.

사전학습된 모델은 이후 enhancer activity, chromatin 상태 예측, gene expression 예측, cross-species mapping 등 genomic downstream task를 적용해 fine-tuning할 것임.

마일스톤 **3.2 (2년차): 300억 (30B) 멀티오믹스 expert 파운데이션 모델 개발**

목표.
유전체 기반의 파운데이션 모델에 이질적인 오믹스 데이터를 통합하여, 치매 예측에 특화된 파인튜닝 모델을 구현하고자 함. 특히 광주 치매 코호트**(GARD)**를 활용함으로써, 글로벌 수준의 생물학적 신호뿐만 아니라 한국인 특이적 생물학 신호까지 반영하는 치매 예측 모델을 구현하는 것이 목표임.

활용데이터.
치매 특화 파인튜닝 모델은 GARD 코호트 데이터를 중심으로 학습할 계획임. GARD는 국내 치매 환자 코호트로, **DNA** 메틸레이션**(EPIC array),** 혈장 단백체**(Olink),** 마이크로바이옴**(16S rRNA** 및 **WGS** 기반**)** 데이터를 포함하고 있어, 치매 및 인지기능 저하와 관련된 생물학적 신호를 모델에 직접 반영할 수 있음 또한.
이와 함께, 외부 보조 데이터로는 GTEx [A46]와 UK Biobank [A47]를 활용할 예정임. GTEx는 뇌를 포함한 다양한 조직의 조직 특이적 발현 데이터를 제공하므로, 정상 발현 패턴을 레퍼런스로 삼아 조절 네트워크에 대한 해석 가능성을 높일 수 있음. UK Biobank는 대규모 일반인을 대상으로 수집된 표현형 및 멀티오믹스 데이터를 포함하고 있으며, 이들 간의

데이터가 개별 수준에서 짝 지어져 있다는 점에서 큰 강점을 지님. 이러한 표현형–오믹스 매핑 구조를 통해, 다양한 인지기능 지표와 연관된 생체 신호를 직접적으로 보완할 수 있으며, 이는 모델의 일반화 성능 향상에 기여할 수 있음.

모델 설계 및 학습 계획**.**
이질적인 오믹스 데이터를 효과적으로 통합하기 위해, 각 오믹스 모달리티에 특화된 인코더와 대조 학습 기반 융합 방식(contrastive fusion) [A48] 기반의 통합 모듈을 도입할 예정임. 특히, 인코더는 전문가 특화 모델 (Mixture-of-Experts, MoE) [A49] 구조로 구성하여, 각 오믹스 모달리티를 하나의 전문가로 간주하고, 학습된 정보 흐름 메커니즘(gating mechanism)을 통해 입력 상태에 따라 적절한 전문가를 동적으로 선택함. 이로써 모달리티 간 이질성에 유연하게 대응하며, 데이터 결측에도 강건하게 작동하도록 설계하고자 함.

데이터 전처리**.**
데이터 구조가 상이한 점을 고려하여, 메틸레이션, 단백체, 미생물 각각 독립된 인코더를 설계함. 각 인코더는 MoE 구조의 하나의 expert로 작동하며, gating 메커니즘을 통해 입력 데이터의 modality에 따라 선택적으로 활성화되도록 설계할 예정임.
시퀀스 구조의 데이터의 경우 모달리티 고유의 시계열적, 연속적 특성을 포착할 수 있도록, 최근 생물의료 시계열 데이터에 효과적으로 적용된 **Caduceus** [A43]와 같은 Mamba [A44] 기반 상태 공간 모델(state space model) 아키텍처로 설계할 예정임.
시퀀스 구조를 갖지 않는 데이터의 경우에는 입력 데이터의 형태에 적합한 아키텍처(CNN, MLP, attention 기반 구조 등)를 활용하여 인코더를 설계할 계획임. 이를 통해 각 모달리티의 구조적 특성과 정보 표현 양식을 최대한 보존하면서, 통합 표현 공간으로 효과적으로 매핑할 수 있도록 할 예정임.

데이터 통합**.**
통합 과정에서는 대조 학습 기반 융합 방식(contrastive fusion) 방식을 적용할 예정임. 동일 피험자의 다양한 모달리티 표현은 유사하게, 서로 다른 피험자의 표현은 분리되도록 대조 학습 기반 손실 함수(contrastive loss)를 활용할 계획임. 이를 통해 모달리티 간 상호작용 정보를 보존하면서도 고유성을 유지할 수 있으며, 융합 표현의 안정성과 표현력 또한 강화함.
일부 모달리티가 결측된 데이터 구조를 반영하기 위해, 학습 시 학습 과정에서 특정 모달리티를 의도적으로 제외**(modality dropout)**하는 전략을 적용할 계획임. 다양한 모달리티 조합(예: 메틸레이션만 있는 경우, 메틸레이션+단백체만 있는 경우 등)에서도 강건한 예측이 가능하도록 학습할 계획임.
데이터 융합 단계에서는 각 모달리티의 존재 여부를 반영하는 마스크**(mask)** 또는 게이팅**(gating)** 메커니즘을 적용하여, 결측 모달리티로 인한 정보 왜곡이 발생하지 않도록 할 예정임.
추가적으로, 모델의 일반화 성능을 높이기 위한 간단한 데이터 확장기법도 병행할 예정임. 구체적으로는, 메틸레이션 데이터에는 작은 범위의 노이즈를 주입하고, 마이크로바이옴 특성에는 계통분류 수준의 정보를 무작위로 섞는 방식을 적용하여, 일반화 성능을 높이고자 함.

Fine-tuning.
기존의 유전체 기반 파운데이션 모델은 대부분 고정하거나, 필요한 경우 LoRA(Low-Rank Adaptation) [A50] 와 같은 경량화된 파인튜닝 기법을 활용할 예정이며, 새롭게 추가되는 오믹스 인코더 및 융합층(fusion layer), 그리고 다운스트림 예측모듈에 대해 집중적으로 학습을 수행할 예정임.
다운스트림 예측 과제는 치매 **/** 경도인지장애**(MCI) /** 정상군 분류**,** 인지점수**(MMSE** 등**)** 예측**,** 뇌 나이 회귀 예측**,** 마이크로바이옴과 인지 기능 간의 관계 예측 등으로 구성하고자 하며, 각 과제에 적합한 손실함수(cross-entropy, MSE, Pearson correlation loss 등)를 적용하여 최적화할 예정임

마일스톤 3.3 (3년차): 470억 (47B) 멀티오믹스-LLM 정렬 파운데이션 모델 개발

목표.

마일스톤 3.2에서 개발된 한국인 치매 특화 멀티오믹스 전문가 모델(OMNIA-K)을 거대 언어 모델(LLM)과 정렬하여, 복잡한 오믹스 데이터로부터 생물학적 의미를 해석하고 자연어 수준의 추론이 가능한 멀티모달 파운데이션 모델(OMNIA-X)을 구축함.

방법론**:** 양방향 정렬**(Bi-directional Alignment)**을 통한 해석 가능성 및 성능 극대화

본 마일스톤은 오믹스 데이터의 예측 정확도와 LLM의 해석 능력을 동시에 극대화하기 위해, (1) 오믹스 정보를 LLM에 통합하고, (2) LLM의 지식을 오믹스 정보에 투영하는 상호보완적인 양방향 정렬 접근을 취함. (그림 4. OMNIA-X 도식 참조)

1) Omics → LLM 통합: 근거 기반 추론 능력 강화

오믹스 데이터의 임베딩을 LLM이 이해할 수 있는 '근거(evidence)'로 변환하여 주입함으로써, LLM이 문헌 정보뿐만 아니라 실제 생물학적 데이터를 기반으로 추론하도록 함.

- 기술적 접근
  - 임베딩 통합: 오믹스 임베딩을 LLM 입력값(prefix) 또는 교차 어텐션(Cross-Attention)의 Key/Value로 삽입하고, 중간 어댑터 레이어를 파인튜닝하여 두 모달리티를 연결함.
  - 환각(Hallucination) 제어: LLM의 가장 큰 한계점인 환각 문제를 해결하기 위해, 검색 증강 생성(RAG) 기법을 적극적으로 도입함. 자연어 질의가 입력되면, ClinVar, OMIM, PubMed 등 외부 지식 데이터베이스에서 관련 정보를 실시간으로 검색하여 LLM의 답변 생성 컨텍스트에 포함시킴.
  - 학습 고도화: 티쳐 포싱(teacher forcing) [A51], 다이나믹 프리픽스 튜닝(dynamic prefix tuning)[A52] 등 다양한 파인튜닝 기법을 실험하여 최적의 정렬 성능을 확보함.

2) LLM → Omics 통합: 오믹스 데이터의 의미론적 표현력 강화

LLM이 학습한 방대한 의학 문헌 지식을 오믹스 임베딩 공간에 투영하여, 단순 서열 기반의 임베딩을 넘어 고차원의 생물학적 의미(예: 병리성, 기능 경로 등)를 내포하도록 표현력을 강화함.

- 기술적 접근
  - 의미론적 임베딩 생성: Wikipedia, PubMed 등 방대한 의학 문헌을 학습한 LLM을 이용하여, 특정 유전자나 변이에 대한 의미론적 임베딩을 생성함.
  - 대조 학습 기반 정렬: LLM이 생성한 '의미 임베딩'과 마일스톤 3.2 모델의 '데이터 기반 임베딩'을 대조 학습(Contrastive Learning) 기법으로 정렬함.

이러한 양방향 정렬 프레임워크는 단순히 두 모델을 결합하는 것을 넘어, 오믹스 FM에 '해석 가능한 지능'을 부여하고, LLM에 '데이터 기반의 견고한 근거'를 제공하는 강력한 시너지를 창출함. 이는 신뢰할 수 있고(trustworthy), 설명 가능하며(explainable), 실제 문제 해결에 기여하는(impactful) AI 모델 개발 목표에 완벽히 부합함.

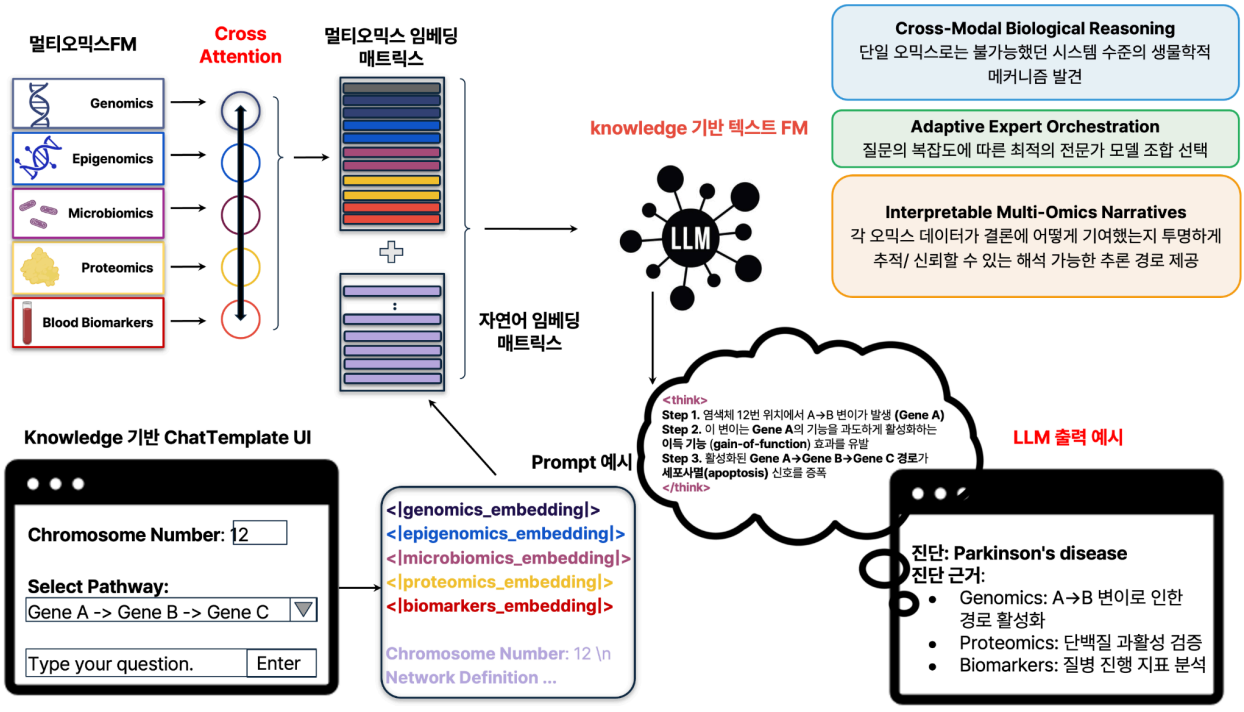


그림 **4.** 멀티오믹스**FM**과 **LLM**의 정렬 **: OMNIA-X** 도식

세부목표 **4. [**뇌**-**멀티오믹스 **FM] LLM**과 정렬된 통합 멀티모달 뇌**-**멀티오믹스 파운데이션 모델 구축

목표**.** 앞서 개발된 각 모달리티별 파운데이션 모델(뇌영상, 뇌파, 멀티오믹스)을 언어 모델(LLM)의 의미론적 공간에서 최종적으로 융합하여, 1,550억 파라미터 규모의 통합 파운데이션 모델(BOM)을 완성함. 완성된 모델은 한국인 종단 데이터(GARD) 기반의 파인튜닝을 통해 '인지예비능'을 정밀하게 예측하고, 그 근거를 설명하며, 개인 맞춤형 솔루션을 제안하는 임상적 실용성을 갖춘 모델로 고도화하는 것을 목표로 함. 이는 뇌의 구조, 기능, 생물학적 기반, 임상 상태를 모두 아우르는 최초의 '디지털 트윈 브레인'을 구현하여, 치매 정복을 위한 새로운 과학적 패러다임을 제시하는 것.(표 **5**).

표 **9.** 세부목표 **4** 및 마일스톤 요약

마일스톤	핵심 평가 태스크
3-4년차 - M4.1 멀티모달 뇌-멀티오믹스 통합 LLM 정렬 1550억 FM 학습	- 쌍을 이루지 않은 데이터에 대한 제로샷 교차 모달 예측 성능 검증 (예: 뇌영상 데이터로 오믹스 지표 예측)
4-5년차 - M4.2 인지예비능을 위한 GARD 종단 데이터의 파인튜닝 학습 (인지예비능, 혈액 마커, 임상 데이터) 및 경량화	- LLM-그래프 기반 인지예비능 예측/솔루션 레포트 생성 - MCI 전환 예측 AUC 7%p 이상 향상 및 모델 90% 경량 - 모달리티 조합별 성능 비교 평가/검증 (데이터 가용성에 따른 성능 변화 분석)

마일스톤 **4.1 (3-4**년차**): 멀티모달 뇌-멀티오믹스 통합 LLM 정렬 1550억 FM 학습**

다양한 모달리티의 뇌이미징, 멀티오믹스, 텍스트 데이터를 통합하는 FM의 개발에서 가장 해결이 어려운 문제는 대규모의 멀티 모달 데이터(멀티모달MRI 데이터, 뇌 전기 생리학데이터, 멀티오믹스, 텍스트 데이터) 쌍을 확보하는 것이 매우 어렵다는 것임. 최신의 FM 연구들에서는 오디오, 이미지, 언어 등 다중 모달리티 간의 통합이 LLM을 매개로 하여 이뤄지며, 이를 통해 창발적 정렬(emergent alignment) 현상을 관찰할 수 있다는 것이 보고됨 [A22-23]. 이는 LLM 중심의 멀티모달 통합 모델들이 서로 다른 모달리티의 데이터의 공동 표상을 학습함으로써 달성됨. 특히, 정렬해야 할 모달리티의 수가 많으며, 모든 모달리티 정보를 동시에 포함하는 데이터를 확보하기 어려운 경우에도 LLM을 중심으로 모달리티들을 통합함으로써 창발적 정렬을 관찰할 수 있다는 것이 보고됨. 이에 본 연구진은 최신의 멀티모달 FM 연구들에서의 결과를 기반으로 하여, LLM을 중심으로 멀티모달 MRI, 뇌파, 멀티 오믹스, 텍스트 데이터를 통합할 수 있는 모델을 개발하고자 함(그림 **5.A**).

앞선 마일스톤들에서 개발한 다중 모달리티 모델들은 기본적으로 거대 언어 모델에 대한 추가적인 파라미터 업데이트 없이, 각 모달리티 데이터가 언어 모델들의 잠재 공간 상에서 표상될 수 있도록 모달리티의 인코더들을 학습한 것임. 따라서, 앞선 마일스톤들을 통해서 개발된 멀티모달MRI 데이터, 뇌 전기 생리학데이터, 멀티오믹스 데이터의 인코더들을 거대 언어모델을 중심으로 결합함으로써, 하나의 프레임워크에서 뇌 이미징 데이터와 멀티오믹스 데이터를 통합적으로 처리할 수 있게 됨. 구체적으로는 앞선 마일스톤들에서의 모델 개발에 활용되었던 오픈 소스의 뇌-텍스트, 멀티오믹스-텍스트로 구성된 멀티모달 쌍 데이터를 활용하여 모델을 학습함으로써 텍스트 데이터를 다른 모달리티 데이터를 통합하는 매개로 삼아 다양한 모달리티 데이터들이 공통 임베딩 공간 상에 자연스럽게 정렬되도록 유도함. 이를 위해, 뇌이미징 모달리티 데이터들의 인코더와 멀티오믹스 데이터들의 인코더는 고정된 LLM과 어댑터 모듈을 통해 연결되며, InfoNCE 기반 대조 손실을 사용하여 의미적으로 유사한 샘플들이 동일한 임베딩 공간에 위치하도록 학습된다(그림 **5.B**). 이러한 정렬이 완료되면, 직접적으로 쌍을 이루지 않은 뇌이미징와과 멀티오믹스 데이터 간에도 언어 임베딩을 매개로 한 간접적 정렬이 창발적으로 발생하고, 이를 기반으로 명시적으로 학습되지 않은 작업에 대해서도 합리적인 추론이 가능해진다.

뿐만 아니라, 앞선 마일스톤에서의 모델 개발에 활용되었던 전문가 혼합(Mixture of Expert) 시스템을 뇌-멀티오믹스-LLM 통합 모델에 적용하여, 모델의 규모를 확장시키고자 함. 특히나, 전문가 혼합 시스템 기반 모델들의 학습 불안정성을 해결하기 위해 앞선 마일스톤에서 적용하였던 테크닉들(공유/특화 전문가 라우팅, 편향 기반의 전문가 부하 분산, 높은 비율을 전문가 드롭아웃)과 노하우를 활용할 예정임.

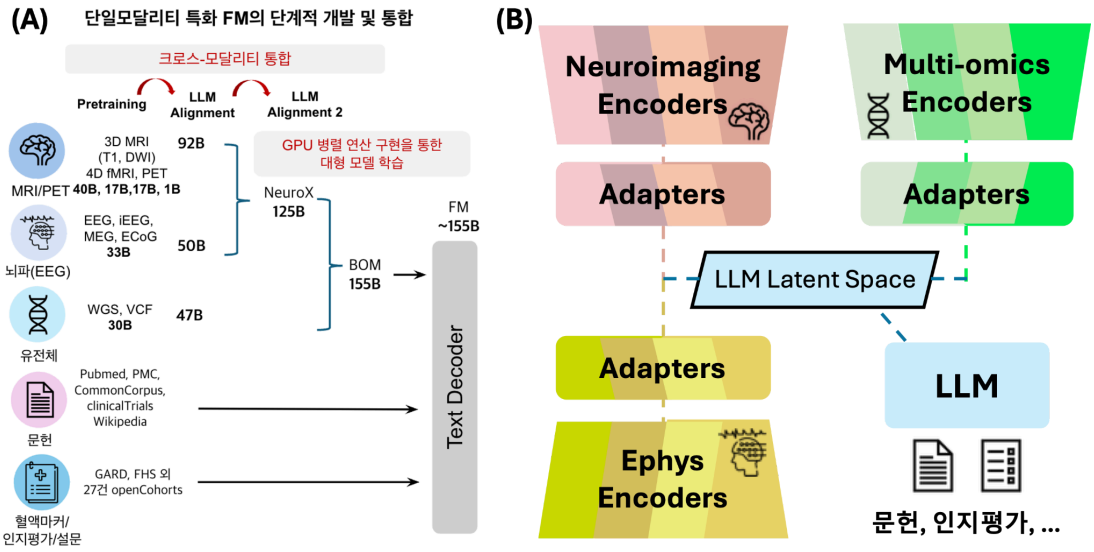


그림 **5. (A) LLM** 중심의 개별 데이터 특화 **FM** 통합 모식도 **. (B) LLM** 중심의 멀티모달 뇌**-**멀티오믹스 통합 모델 구조**.**

마일스톤 **4.2 (4-5**년차**):** 인지예비능을 위한 **GARD** 종단 데이터의 파인튜닝 학습 **(**인지예비능**,** 혈액 마커**,** 임상 데이터**)** 및 모델 경량화

치매 예측 파운데이션 모델은 뇌영상 기반 사전학습 덕분에 일반적 신경지표를 잘 포착하지만, 인지예비능·혈액 바이오마커·세부 임상 기록처럼 사전학습 단계에서 한 번도 보지 못한 테이블·연속 변수와 결합될 때는 도메인 불일치로 인한 성능 붕괴가 보고되어 왔음[확인필요!! A1]. 본 연구팀은 이 한계를 해결하기 위해 두가지 서로 다른 접근 방법을 활용하여 성능과 연산 효율성 및 확장 가능성에 대한 비교를 진행하고 모델을 고도화하고자 함.

첫번째 접근 방법은 도메인 적응 학습의 패러다임을 통해 인지예비능·혈액 바이오마커·세부 임상 기록을 통합적으로 활용할 수 있도록 앞선 마일스톤을 통해 개발한 멀티모달 뇌-오믹스-LLM 모델을 파인튜닝하는 것임. (i) 도메인-적응 사전학습(DAPT) 으로 GARD 전주기 데이터(10년·15년간의 종단 데이터)를 자기지도 방식으로 모델 내부 분포에 먼저 주입하고, (ii) 혈액 마커·임상 수치 전용 어댑터를 추가해 뇌 표현과 비영상 생체지표를 저차원 공유 공간으로 정렬하며, (iii) 파라미터 효율 경량화(4-bit QLoRA→LoRA-Merge→7B distilled student)로 병원 온프레미스 추론 지연을 60 % 이상 단축함.

두가지 성공 기준: 첫째, 파인튜닝된 모듈이 혈액 마커를 포함한 다중 모달리티 데이터 입력에서 MCI→치매 전환 예측 AUC를 단일 모달리티 데이터 입력 모델 대비 7 %포인트 이상 끌어올릴 것. 둘째, 최종 학생 모델이 원본 34B 파라미터 대비 90 % 이상 경량화되면서도 성능 감소폭이 1 %포인트 이내에 머물면 마일스톤을 달성한 것으로 간주. 모든 코드와 재현성 스크립트는 brainlife.io [A53]에 공개해 치매 연구 커뮤니티가 즉시 검증·확장할 수 있도록 함.

두번째 접근 방법은 언어 기반 제어의 패러다임을 통해 추가적인 모델 파라미터 업데이트 없이 파인튜닝 학습을 진행하는 것임. 이를 위해, 본 연구팀이 보유한 장기간의 종단 데이터를 활용하여 개인의 인지예비능·혈액 바이오마커·세부 임상 기록 등에 관한 상세한 텍스트 리포트를 구성하고, 이를 활용하여 모델에 대한 파인튜닝을 진행할 계획임. 이는 앞선 마일스톤에서 모델이 치매에 관한 뇌이미징과 omics 데이터의 표상을 학습하는 것에서 한발 더

나아가, 풍부한 종단 추적 정보를 활용하여 개개인의 치매 위험에 대한 상세한 프로파일링을 가능하게 할 것임. 특히나, Chain of Thought [A54]와 ReAct prompting [A55] 등을 비롯한 다양한 Test-Time-Training 과정을 통해서, 최근의 오픈 소스 거대 언어 모델들이 학습한 의학 지식 및 문헌 조사 능력들과 개인의 치매 위험에 대한 프로파일링 정보를 결합함으로써 훨씬 더 복잡한 추론에 기반한 정밀화된 프로파일링이 가능할 것으로 기대됨. 또한, Retrieval Augmented Generation (RAG) [A56]를 활용하여, 최신의 의학 및 신경과학적 연구 결과들을 동적으로 모델의 추론 과정에 반영할 것임. 이는 LLM의 추가적인 학습이 없이도 최신의 연구 결과들을 효율적이고 확장적으로 반영하여 추론을 진행할 수 있게 함으로써 모델의 지속적인 활용 가능성을 증대시킬 것으로 기대함.

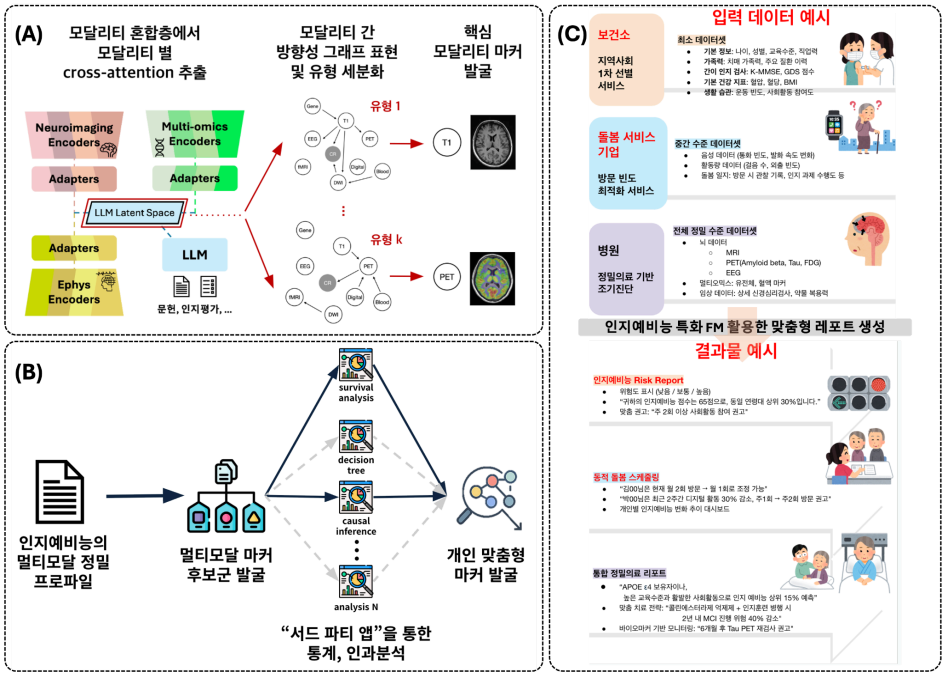


그림 **6. (A)** 뇌-오믹스**-LLM** 통합 모델의 그래프 기반 개인 맞춤형 멀티모달 마커 프로파일링 예시**. (B)** 멀티모달 마커 기반의 데이터 분석을 통한 개인 맞춤형 마커 발굴 도식**. (C)** 임상 현장 상황별 뇌-오믹스**-LLM** 통합 모델의 맞춤형 리포트 예시

본 연구팀은 장기 추적 연구인 GARD 코호트를 통해 축적된 유전, 뇌 영상, 혈액, 디지털 표현형 등 풍부한 다중 양식**(multi-modal)** 데이터를 활용하여 파운데이션 모델(Foundation Model)을 파인튜닝하는 과정을 통해 개인의 인지예비능을 예측하고 개인 맞춤형 정밀 멀티모달 마커들을 발굴하고자 함. 특히, LLM 기반의 그래프 시각화를 통해 각 데이터 양식(modality)이 인지예비능과 맺는 복잡한 상호 관계 및 요인별 중요도를 직관적으로 해석하고 분석할 수 있도록 함으로써 돌봄 서비스를 보조할 수 있도록 하고자 함(그림 **6.A**). 또한, 개인 맞춤형 정밀 프로파일링 레포트를 통해 발굴한 멀티모달 마커의 후보군들을 기반으로, LLM을 통해 외부 통계 분석 툴을 활용함으로써 멀티모달 마커의 후보군들 간의 데이터 분석을 진행하여 개인별로 주요한 마커 세트를 탐색하고 이에 기반한 개입과 돌봄 서비스를 보조할 수 있도록 함(그림 **6.B**). 특히, 임상 현장 상황에 따라 가용한 데이터가 한정적일 수 있음을 고려하여, 가용한 데이터의 종류에 따른 모델 성능을 검증함으로써 돌봄 서비스 및 의료 보조 과정에서의 모델 활용시의 메뉴얼을 마련할 것임 (그림 **6.C**). 이는 인지 저하의 핵심 바이오마커를 효과적으로 발굴하는 새로운 경로를 제시할 뿐만 아니라, 향후 임상 현장에서 수집이 용이한 최소한의 데이터만으로도 안정적인 예측 성능을 확보하는 경량화 모델과 서비스 개발의

초석이 될 것임.

세부목표 달성을 위한 기술적 전략 및 위험 관리 방안

1. 선행 과제를 통해 검증된 연구 준비성 및 핵심 역량

본 연구팀은 다년간의 선행 연구 및 미국 에너지부(DOE)에서 지원한 대형 과제 (ALCC NeuroX) 수행 경험을 통해 본 과제 수행에 필요한 핵심 역량을 이미 확보하고 있으며, 주요 기술적 위험 요소를 사전에 완화하였음. 본 연구팀의 준비성은 다음 세 가지로 요약됨.

가. 세계 최고 수준 계산 자원 및 분산/병렬 학습 기술 확보

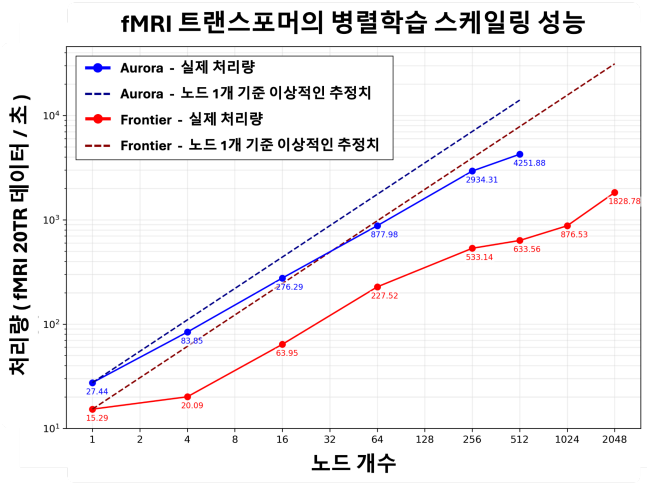
계산 자원 미국 에너지부의 슈퍼컴퓨터 지원 과제인 ALCC, 국내 KISTI 및 서울대 서버 등으로 2025년에 본 연구팀이 확보한 자원은 약 500만 GPU 시간 정도임 **(**표 **10)**. 또한, 네이버, AICA, 그리고 AWS를 통해 지원받을 슈퍼컴퓨터 자원을 합치면 최소 170만 정도의 GPU 시간이 추가로 준비됨. 이에 더해, 본 연구팀이 2025년 7월에 신청한 세계 최대의 엑사스케일 슈퍼컴퓨터 과제인 미에너지부의 INCITE 과제 및 미국 DOE 산하 기관 NERSC의 차세대 슈퍼컴퓨터인 Doudna 지원 과제에 선정되면 2026년부터 3년간 최대 1,600만 GPU 시간을 추가로 확보할 수 있음.

표 **10.** 본 연구를 위해 확보된 **GPU** 자원

프로그램 및 파트너	**GPU** 사양	연간 확보된 **GPU** 시간 **(GPU** 장수**)**
미에너지부 ALCC	A100, AMD MI250X, Intel Data Center Max 1550 Series (HPE Slingshot 11 노드간 통신)	461만 시간 (101,376장)
미에너지부 ERCAP, GENAI	A100 (HPE Slingshot 11 노드간 통신)	3.6만 시간 (6,144장)
서울대 자체 서버	RTX A5000, RTX 3090	14만 시간 (16장)
KISTI	V100, A100, H200, GH200	19만 시간 (302장)
Naver Cloud	H100	70만 시간 (80장)
AWS	H100, H200, B200	98만 시간 (112장)
AICA	A100, H100	70만 시간 (80장)
총합		800만 시간 이상 (사설 클라우드기반 200억 이상)

분산/병렬 학습 기술 본 연구팀은 선행 연구를 통해 41억 파라미터 (4.1B) 규모의 fMRI 트랜스포머 모델 (SwiFT V2)이 세계 최고 수준의 미국 슈퍼컴퓨터인 Frontier 및 Aurora에서 강력한 스케일링 성능을 보임을 이미 입증하였음 **(**그림 **7)**. 슈퍼컴퓨터 활용에 최적화된 병렬 학습 기술 (Deep Speed ZeRO 2단계, BF16 정밀도)을 적용한 결과, 512개 노드 (총 6,144 개의 GPU)에서 초당 4,252개의 fMRI 데이터를 처리하는 높은 처리량을 달성함. 이러한 성능은 단일 노드를 기준으로 계산해도 수백 개 노드 규모까지 거의 이상적인 스케일링 효율 (strong scaling efficiency)을 보여줌. 이런 본 과제의 도전적인 연구 계획을 기술적으로 뒷받침하며, 대규모 계산 자원을 효율적으로 활용하여 파운데이션 모델을 성공적으로 학습시킬 수 있음을 보증하는 강력한 근거임.

그림 **7. fMRI** 파운데이션 모델의 병렬 학습 확장성 결과**.** 41억 파라미터 규모 모델의 스케일링 효율성을 두 개의 미국 슈퍼컴퓨터인 Frontier와 Aurora에서 확인함. 1개 노드에서의 성능을 기준으로 계산했을 때, Aurora의 512개 노드 (GPU 6,144장)에서 30.8%의 강력한 스케일링 효율 보임.



나**.** 전례 없는 규모의 멀티모달 데이터 확보 및 활용 계획

본 연구는 국내외 29개 주요 컨소시엄을 통해 확보한 122만 명 이상, 총 666TB 이상의 멀티모달 데이터를 활용함 **(**표 **11)**. 이는 뇌영상, 뇌파, 유전체/오믹스, 임상 정보 등 뇌 인지 기능 연구에 필수적인 핵심 데이터들을 모두 포함하며, 파운데이션 모델 구축을 위한 독보적인 데이터 기반을 제공함.

표 **11** 본 연구를 위해 확보된 멀티모달 데이터

데이터 대분류	세부 모달리티	데이터셋	총 대상자 수	데이터 용량 **(TB)**
뇌영상	구조 MRI (T1/T2)	GARD, FHS 등 총 25개 코호트	83,101+	3.24+
	확산 MRI		78,743+	130+
	기능 MRI		77,738+	128.2+
	PET	GARD, ARIC-NCS, MCSA, HRS, A4 Study, AIBL, ADNI, OASIS, KBASE, J-ADNI, DIAN	17,388+	
뇌파	EEG, ECoG, sEEG, MEG	TUEG, AJILE12, HBN-EEG, NSRR, PEERS, GARD	47,704+	139.6+
멀티오믹스	유전체 (GWAS, WES/WGS), 후성유전체, 대사체	T2T-CHM13, GRCh38, GTDB v220.0, IMG/VR, IMG/PR, NCBI Reference Genomes, JGI IMG, MGnify, NCBI	1,072,056+	265+

		Metagenomes, NCBI Organelle Web Resource, NCBI GTF, Ensembl release 112, Rfam, RNAcentral, Illumina EPIC Array, WGBS, ENCODE, JASPAR, FANTOM5, Roadmap Epigenomics, EPDnew, PSI, phastCons, phyloP GARD, UKB, ABCD		
혈액 바이오마커	GFAP, NfL, CSF	UK Biobank, ADNI, GARD	52,484	0.001+
임상/행동 텍스트 데이터	임상진단정보(CDR-SB), 신경심리검사, 사회적 기능	GARD, FHS 등 총 25개 코호트	1,180,000+	0.009+
총합			**1,222,895** 명†	666+

† 전체 코호트의 고유 대상자 수, 여러 모달리티의 데이터를 갖는 중복 대상자 수를 제외한 숫자.

다**.** 아키텍처 및 코드의 즉시 실행 가능성

본 연구팀은 본 과제의 핵심 아키텍처(4D 뇌영상 트랜스포머 등)의 개발, 벤치마킹, 스케일링을 이미 완료하였음. Deep Speed 등 분산 학습 라이브러리에 최적화된 코드베이스는 안정적인 대규모 훈련과 높은 성능을 입증하여, 과제 1차년도부터 즉시 생산 수준(production-level)의 연구 수행이 가능함. 또한 표 **12**에 정리된 본 연구팀의 구체적이며 체계적인 연구 계획은 과제 시작 이후 본 연구팀이 각 세부목표 별로 바로 모델 학습을 수행할 수 있다는 근거가 됨. 그리고 각 마일스톤마다 실행 및 측정이 가능한 계획을 세웠기에 진행 정도와 성과를 확인하며 과제를 수행할 수 있음을 보임.

표 **12.** 각 세부목표 및 마일스톤별 실험 계획

세부목표	마일스톤	년차	실험 설명 (a, b, 그리고 c 는 실험 종류를 의미)	실험별 예상 **GPU** 사용 시간 계산[1]				데이터 토큰 개수[3]	노드 시간
				노드 개수	에폭 시간	총 에폭	실험 횟수[2]		
1. 멀티모달 뇌영상 **FM**	**M1.1**	1	(a) 모델 구조 탐색 (70.4k)	16	[1, 1.5]	40	[25, 20] × 2	4.7B	288k
			(b) 사전학습 및 규모 확장(204.8k)	512	1	40	5 x 2	11.9B	
			(c) 다운스트림 과제 평가 (12.8k)	4	0.4	40	100 x 2	200M	
	M1.2	2	(a) LLM-정렬 모델 사전학습 (153.6k)	512	1.5	20	10	11.9B	155k
			(b) 다운스트림 과제 평가 (0.6k)	4	0.5	1	300	85.9M	
2. 멀티모달 뇌파 **FM**	**M2.1**	1	(a) 모델 구조 탐색 (10.9k)	16	0.34	15	100	1.3B	122k
			(b) 사전학습 및 규모 확장 (107.9k)	[128, 256, 512]	[1, 3.1, 5.8]	[20, 15, 10]	[5, 3, 3]	13.1B	
			(c) 다운스트림 과제 평가 (2.5k)	1	0.5	50	50	1.1M	
	M2.2	1~2	(b) LLM-정렬 모델 사전학습 (289.3k)	512	56.5	10	1	27.1B	305k
			(c) 다운스트림 과제 평가 (15.8k)	1	3.5	30	5 x 30	159M	
	M2.3	2	(a) 뇌영상-뇌파-LLM 정렬 전략 탐색 (38k)	128	0.3	10	10	13M	114k
			(b) 뇌FM-LLM 정렬 어댑터 학습 (10.2k)		1	10	2	13M	
			(c) 다운스트림 과제 평가 (20.5k)	512	1	10	2 x 2	14.8M	
3. 멀티오믹스 **FM**	**M3.1**	1	(a) 모델 구조 탐색 (10.9k)	16	0.34	15	100	1.3B	122k
			(b) 사전학습 및 규모 확장 (107.9k)	[128, 256, 512]	[1, 3.1, 5.8]	[20, 15, 10]	[5, 3, 3]	13.1B	
			(c) 다운스트림 과제 평가 (2.5k)	1	0.5	50	50	1.1M	
	M3.2	2	(b) 30B 모델 사전학습 (289.3k)	512	56.5	10	1	27.1B	527k
			(c) 다운스트림 과제 평가 (15.8k)	1	3.5	30	5 x 30	159M	
	M3.3	3	(a) 멀티모달 뇌영상-뇌파-LLM 정렬 전략 탐색 (3.8k)	128	0.3	10	10	13M	38k
			(b) 대규모 뇌FM - LLM 정렬 학습 (10.2k)	512	1	10	2	13M	
			(c) 다운스트림 과제 평가 (20.5k)	512	1	10	2 x 2	14.8M	
4. 뇌-멀티오믹스 **FM**	**M4.1**	4	(a) 모달리티 인코더 사전학습 (51.2k)	256	5	20	2	39B	134k
			(b) LLM 내 모달리티 전문가 모듈 사전학습 (56.3k)	256	5.5	20	2	39B	
			(c) QLoRA 기반 LLM 파인튜닝(25.6k)	256	2.5	20	2	39B	
			(d) 모델 성능 평가 (0.1k)	1	0.5	1	200	145.9M	
	M4.2	5	(a) 복합 추론 과제 (1.5k)	1	1	1	300 × 5	145.9M	5k
			(b) 테스트 시점 적응 과제(3.5k)	1	1	1	700 × 5	14.6M	
총합									**1.269M**