On the Opportunities and Risks of Foundation Models

Rishi Bommasani, Percy Liang

Center for Research on Foundation Models (CRFM)
Stanford Institute for Human-Centered Artificial Intelligence (HAI)
Stanford University

2022 arXiv preprint

발제자: HUMANE Lab Research Intern 최종현

2024.12.27 랩 세미나

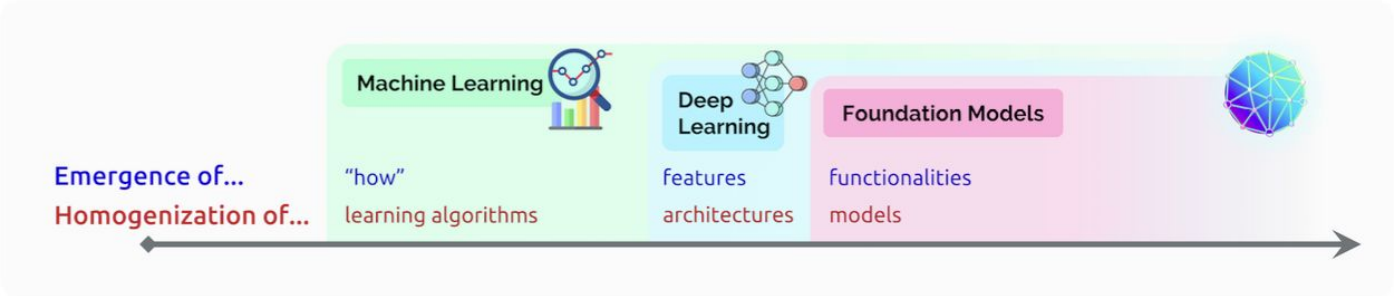
About this paper

- **Foundation Model**
  - 광범위한 데이터(일반적으로 대규모 self-supervised 활용)로 훈련되어 다양한 downstream 작업에 적용 될 수 있는 모델 (e.g., BERT, GPT-3, CLIP, DALL-E)
- Foundation Model의 등장으로 인한 패러다임 전환과 그에 따른 기회 및 위험을 종합적으로 분석하고, 기술적, 사회적, 윤리적 측면을 아우르는 연구의 필요성을 제시함

About this paper

1. **Introduction:** Foundation 모델의 등장과 그로 인한 패러다임 변화
2. **Capabilities:** Foundation 모델이 가진 주요 역량들을 언어, 비전, 로보틱스, 추론, 상호작용, 이해의 철학적 측면
3. **Applications:** 의료, 법률, 교육 분야를 중심으로 Foundation 모델의 구체적인 응용 가능성과 잠재적 위험
4. **Technology:** Foundation 모델 구축에 필요한 모델링, 훈련, 적응, 평가, 시스템, 데이터, 보안, 강건성, 안전성, 이론, 해석 가능성 측면의 기술적 요소를 심층 분석
5. **Society:** Foundation 모델이 사회에 미칠 수 있는 광범위한 영향을 불평등, 오용, 환경, 경제, 법, 윤리적 측면에서 바라봄
6. **Conclusion:** Foundation 모델의 잠재력과 위험을 강조하고, 학술적 협력, 투명성, 책임 있는 개발 및 활용에 대한 내용

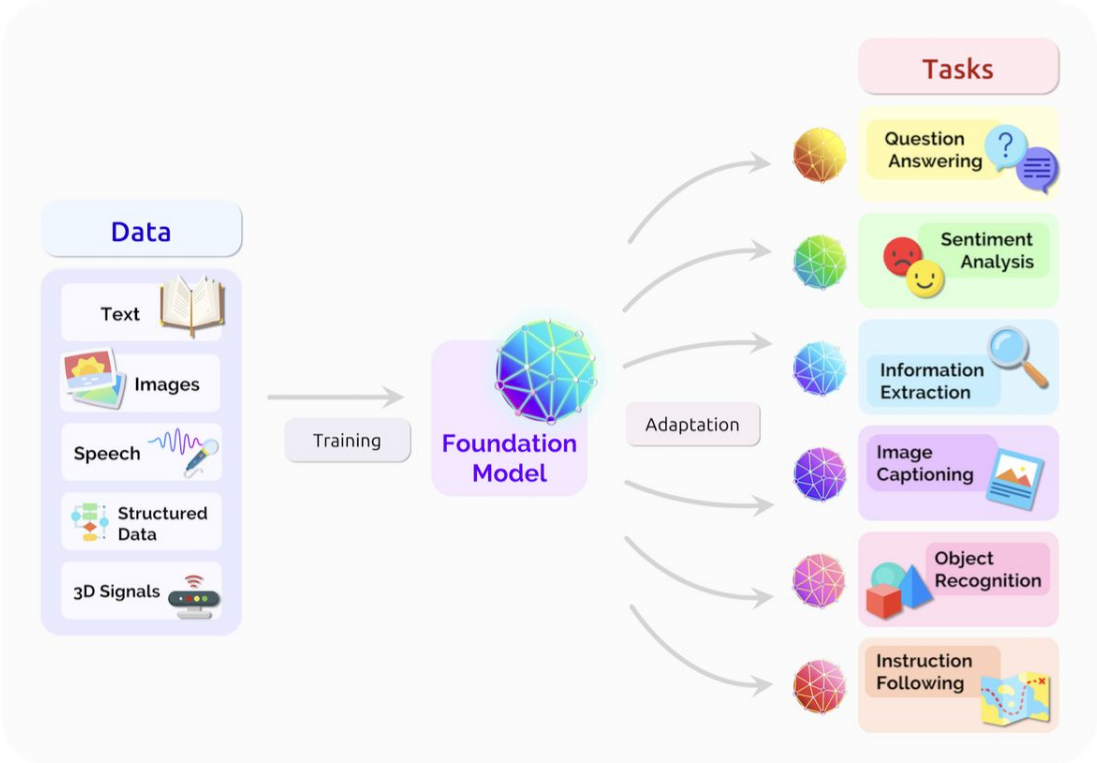
AI 연구의 패러다임 변화



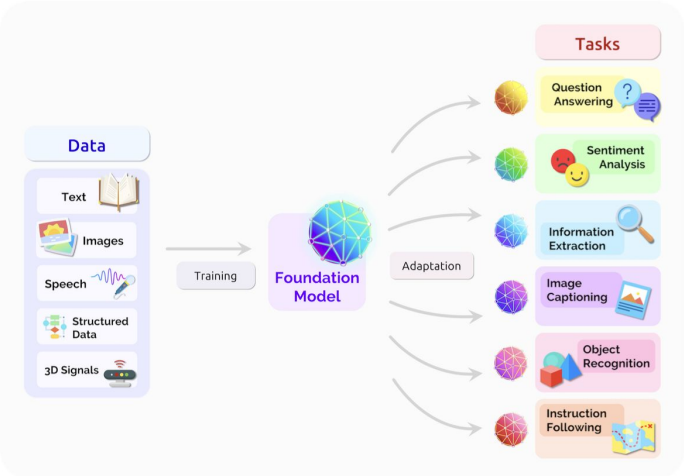
- AI 연구는 각 단계가 진행될수록 Emergence(창발성)와 Homogenization(획일화)이 증가함
- Foundation 모델 단계에서 Emergence와 Homogenization이 가장 두드러짐

	머신러닝	딥러닝	Foundation Models
Emergence	학습 알고리즘을 통해 "어떻게" 작업을 수행할지	"어떤" 고차원 특징을 추출할지	In-context-learning과 같이 훈련 시 의도되지 않은 "새로운 능력"이 나타남
Homogenization	로지스틱 회귀와 같은 일반적인 알고리즘이 다양한 작업에 사용됨	CNN, RNN과 같은 일반적인 모델 아키텍처가 다양한 작업에서 사용됨	소수의 대규모 모델이 다양한 작업의 기반으로 사용됨 (GPT 등)

Foundation Models



Foundation Models



- 광범위한 데이터로 훈련되어 다양한 downstream 작업에 적용 가능한 모델인 Foundation Model의 등장
- BERT, GPT-3, CLIP 등
- 최신 NLP 모델의 대부분은 현재 BERT, RoBERTa, BART, T5 등과 같은 몇 가지 기본 모델 중 하나에서 파생되어 사용됨 (2022년 당시 기준)
- Foundation 모델의 기반이 된 트랜스포머 모델은 텍스트, 이미지, 음성, 테이블 데이터 등 다양한 분야에서 사용되어 자연어처리에만 적용되는 것이 아닌 AI의 패러다임으로 고려됨
- Transfer Learning and Scale

Transfer Learning and Scale

- **전이 학습(Transfer Learning):** 한 작업(e.g., 이미지에서 객체 인식)에서 학습한 "지식"을 다른 작업(e.g., 비디오에서 활동 인식)에 적용하는 것
- Foundation 모델에서의 활용
  - 대규모 데이터셋에서 일반적인 지식을 학습
  - 특정 작업에 필요한 소량의 데이터로 미세조정 (Fine-Tuning)
- **규모(Scale):** 모델의 크기, 데이터셋의 크기, 컴퓨팅 자원의 규모
  - 모델, 데이터, 자원의 규모의 상승 = 모델의 성능, emergence가 향상되는 경향
  - Transformer와 같은 아키텍처의 발전 → 모델의 크기 확장에 기여
  - Self-Supervised Learning → 라벨이 없는 대규모 데이터셋 활용
  - 컴퓨팅 자원 → GPU와 같은 하드웨어의 발전

Transfer Learning and Scale

- **Data.** Compared to prior versions of Llama (Touvron et al., 2023a,b), we improved both the quantity and quality of the data we use for pre-training and post-training. These improvements include the development of more careful pre-processing and curation pipelines for pre-training data and the development of more rigorous quality assurance and filtering approaches for post-training data. We pre-train Llama 3 on a corpus of about 15T multilingual tokens, compared to 1.8T tokens for Llama 2.
- **Scale.** We train a model at far larger scale than previous Llama models: our flagship language model was pre-trained using 3.8×10^{25} FLOPs, almost 50× more than the largest version of Llama 2. Specifically, we pre-trained a flagship model with 405B trainable parameters on 15.6T text tokens. As expected per

3.2 Model Architecture

Llama 3 uses a standard, dense Transformer architecture (Vaswani et al., 2017). It does not deviate significantly from Llama and Llama 2 (Touvron et al., 2023a,b) in terms of model architecture; our performance gains are primarily driven by improvements in data quality and diversity as well as by increased training scale.

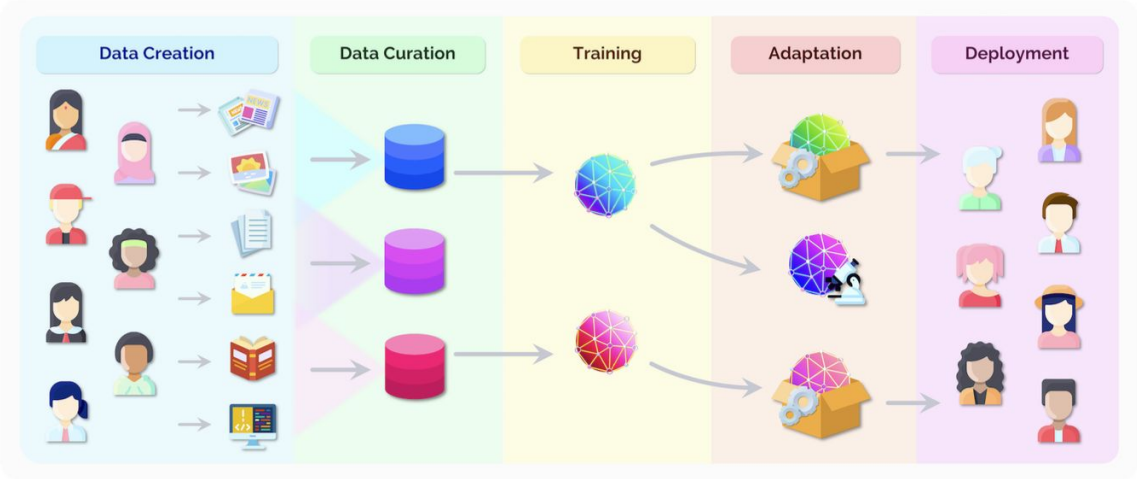
- 2024년까지도 이 특성은 이어지고 있음
- Llama 3(Dubey et al., 2024)에서도 Scale을 확장하는 방식으로 훈련을 진행함

Dubey, Abhimanyu, et al. "The llama 3 herd of models." arXiv preprint arXiv:2407.21783 (2024).

Social Impact

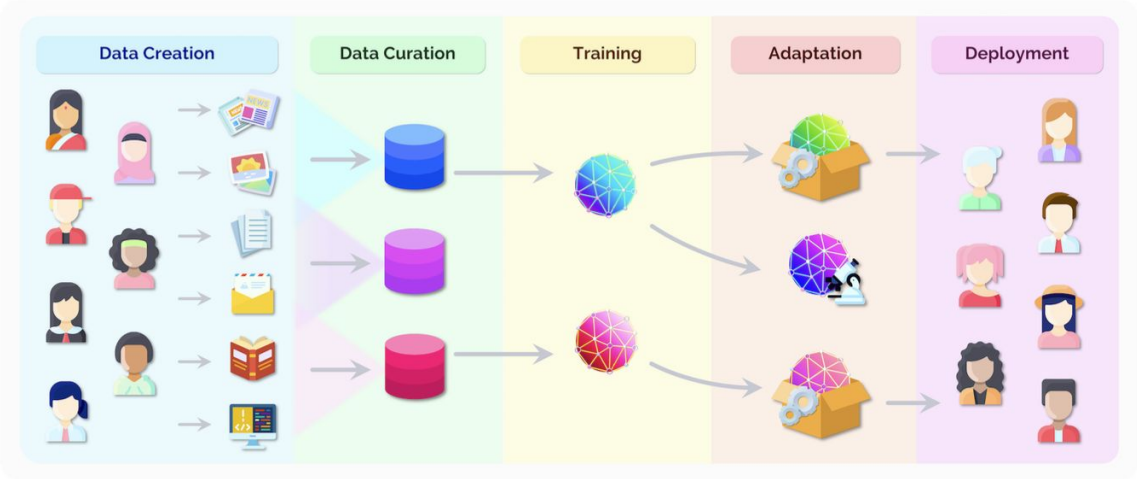
- Foundation 모델은 사회 전반에 광범위한 영향을 줄 수 있음
  - 사회적 불평등 심화 가능성
  - 경제적 영향 (생산성 증가와 격차 확대)
  - 컴퓨팅 자원의 증가로 인한 환경적 영향
  - 잘못된 정보의 확산 가능성
  - 법적, 윤리적 문제
- 연구와 배포를 구분해야 함 → 연구 모델은 경고 라벨을 붙여 신중히 사용, 배포 모델은 철저히 검증 후 활용
- 이를 이해하기 위해서는 모델을 종합적으로 살펴야 함
- Data Creation → Data Curation → Training → Adaptation → Deployment 과정을 거침

Social Impact



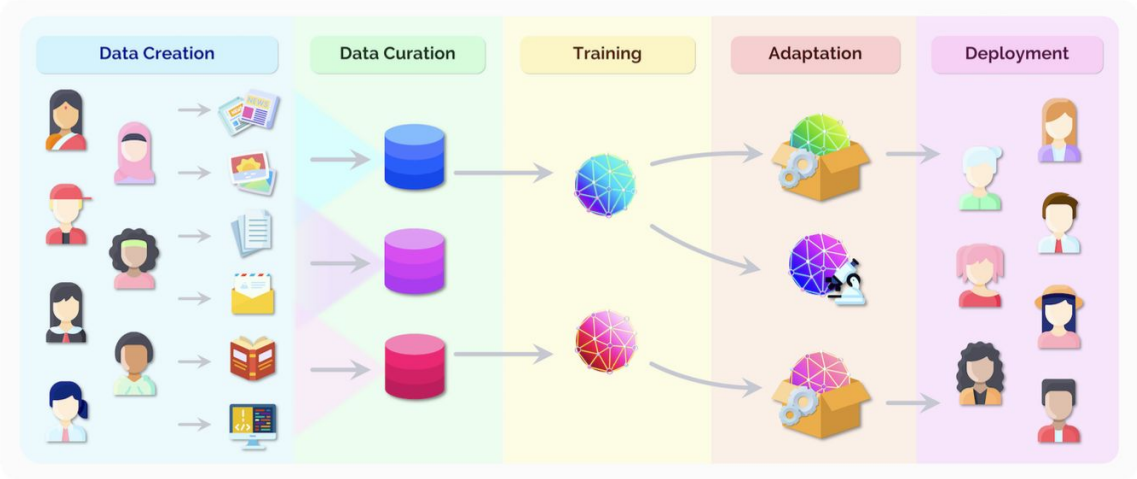
- **Data Creation**
  - 데이터는 인간에 의해 만들어지며 대부분 인간 또는 인간과 관련됨
  - 이메일, 기사, 등 사람이 사람을 위해 생성한 데이터
  - 데이터는 특정한 목적으로 생성되며, 그 목적이 반드시 AI 훈련을 위한 것은 아닐 수 있음
  - 데이터의 출처와 소유권은 반드시 고려되어야 함

Social Impact



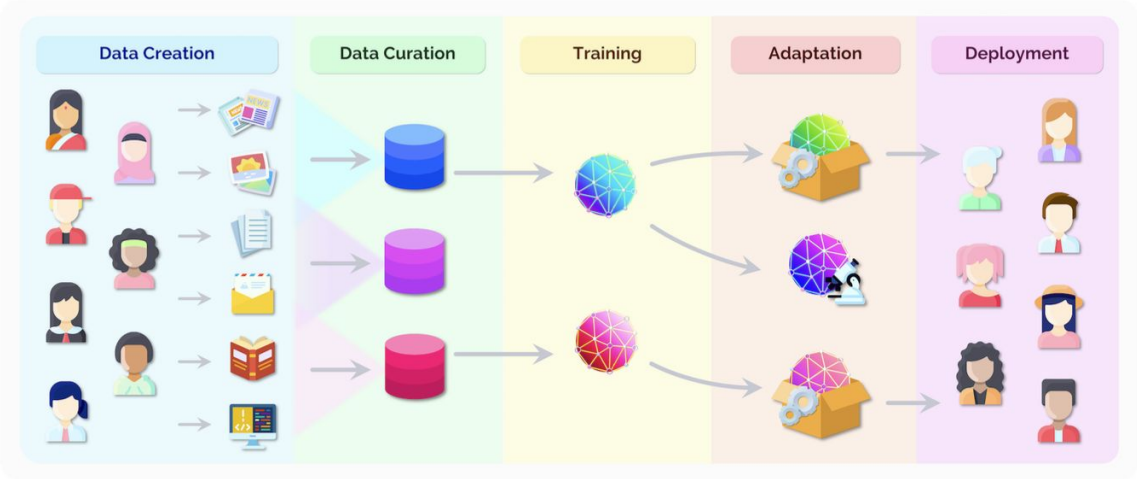
- **Data Curation**
  - 데이터는 특정 목적에 맞게 선별 및 정제됨
  - 인터넷 크롤링 등으로 수집된 데이터도 필터링과 선택이 필요
  - 데이터 선별 과정이 AI 연구에서 충분히 주목받지 못하고 있음

Social Impact



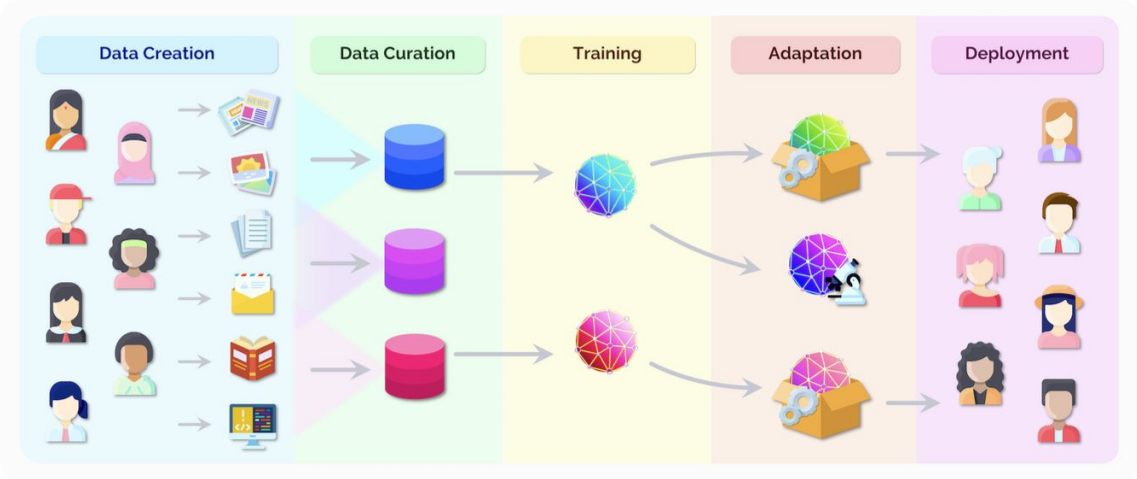
- **Training**
  - 선별된 데이터를 사용하여 기반 모델을 훈련
  - 대규모 데이터와 컴퓨팅 자원을 활용하여 모델의 기본 역량을 학습
  - 이 단계만이 중요한 것은 아니며, 전체 단계의 일부에 불과

Social Impact



- **Adaptation**
  - 훈련된 Foundation 모델을 특정 작업에 맞게 조정
  - 연구적 관점: 요약, 번역 등 특정 작업을 수행하도록 새로운 모델 생성
  - 배포적 관점: 사용자 환경에 맞춘 시스템 구축

Social Impact



- **Deployment**
  - AI 시스템을 사용자에게 직접 제공
  - 배포된 모델은 사람들에게 직접적인 사회적 영향을 미침
  - 연구 모델은 학문적 발전에 기여하나, 배포 모델은 철저한 검증과 점진적 출시 필요

Foundation 모델의 미래

- **학문적 다양성 (Disciplinary Diversity)**
  - Foundation 모델은 기술적 발전뿐 아니라 윤리적, 사회적, 정치적 문제를 포함한 다양한 학술적 접근이 필요
  - 현재의 기술 개발은 주로 Google, Facebook, OpenAI 등 대기업 중심으로 진행되고 윤리적, 사회적 영향을 사후적으로 평가하는 경우가 많음 → 초기 설계 단계부터 사회적, 윤리적 설계를 통합해야 함
  - 학계는 기술, 윤리, 법, 사회적 측면을 결합한 Foundation 모델을 개발해야 함

Foundation 모델의 미래

- **인센티브 (Incentives)**
  - Foundation 모델 연구는 상업적 유인에 크게 의존 → 단기적 이익에 집중 (공익적 연구 소홀)
  - **균형 잡힌 인센티브:** 상업적 유인과 공익적 연구 사이의 균형을 맞추는 인센티브 구조를 설계해야 함
  - **장기적 연구 지원:** 장기적인 관점에서 기초 연구와 응용 연구를 모두 지원해야 함
  - **공익적 연구 장려:** 사회적 책임, 윤리적 고려, 공정성 등을 고려한 연구에 인센티브 제공

Foundation 모델의 미래

- **접근성 감소 (Loss in Accessibility)**
  - 대규모 기반 모델 연구의 접근성이 점점 줄어들고 있음
  - GPT-3와 같은 모델은 공개되지 않거나 제한된 API로 제공
  - 데이터셋 또한 비공개인 경우가 많음
  - 고가의 컴퓨팅 자원과 복잡한 엔지니어링 요구로 인해 연구자가 직접 훈련하기 어려움
  - 해결책
    - 정부의 대규모 컴퓨팅 인프라 지원
    - Volunteer Computing (Folding@home, Learning@home 등)

Capabilities

- Foundation 모델은 대규모 데이터와 Self-Supervised learning을 통해 다양한 능력을 가지며, 이것은 언어, 비전, 로보틱스, 추론, 상호작용 등 여러 분야에 걸쳐 나타남
- NLP 분야에서 혁신적인 변화를 주도하고 있음
  - 뛰어난 언어 생성 및 이해 능력을 가지며 다양한 작업 (e.g., 기계 번역, 텍스트 요약, 질의응답)에서 인간 수준의 성능을 보임
  - 언어의 다양성, 다국어 처리 능력, 상식 추론 능력 등은 여전히 개선이 필요함
  - Foundation 모델은 사람보다 훨씬 많은 양이 데이터가 필요하며 실세계와의 연결 문제도 존재함

Applications (의료, 법률, 교육)

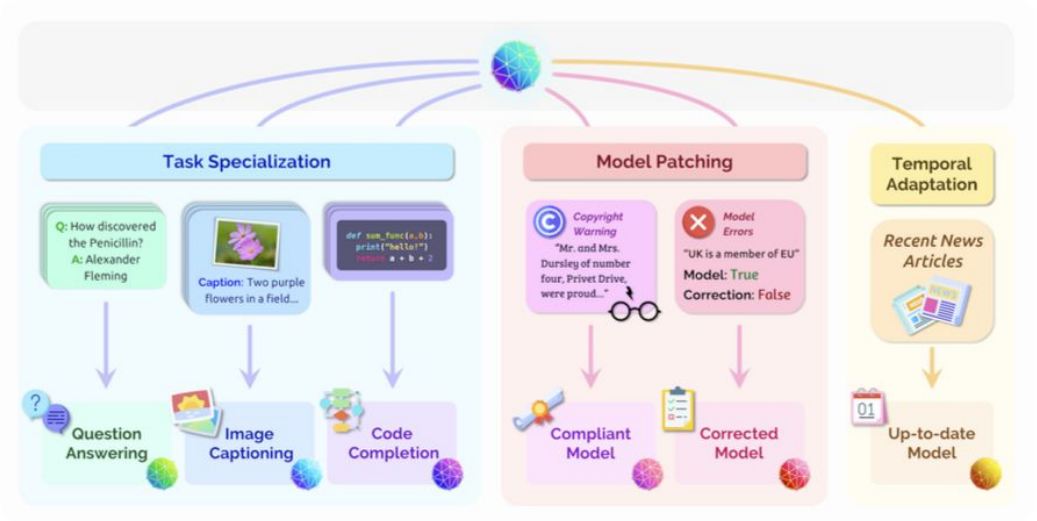
- **의료**
  - 진료 효율성 향상, 진료 접근성 개선, 개인 맞춤형 의학 제공 가능
  - 다양한 형태의 의료 데이터(텍스트, 이미지, 수치, 시계열 등)를 통합하고 품질 관리 하는 것이 과제
  - 모델의 의사 결정 과정을 투명하게 보기 위해 **설명 가능성**의 발전이 필요함
- **법률**
  - 법률 서비스 접근 향상, 법률 업무 효율화
  - 고품질 법률 데이터 부족
  - **설명 가능성**

Applications (의료, 법률, 교육)

- **교육**
  - 학습 수준, 학습 스타일에 맞는 맞춤형 학습 제공
  - 학생별 개별화된 피드백
  - 학생의 인지 상태, 감정, 동기 등을 파악하고 이해하는 것이 중요
  - 교육 분야의 다양한 형태의 데이터(텍스트, 이미지, 영상, 센서 등)를 통합하고 품질 관리하는 것이 중요

Technology

- Expressivity, Scalability, Multimodal, Memory capacity 등 아키텍처 설계의 중요성
- 추론 능력(추론, 계획, 상식)을 강화할 수 있는 새로운 아키텍처 연구가 필요
- 사전 훈련된 파운데이션 모델을 특정 작업 또는 도메인에 맞게 조정 (Adaptation)

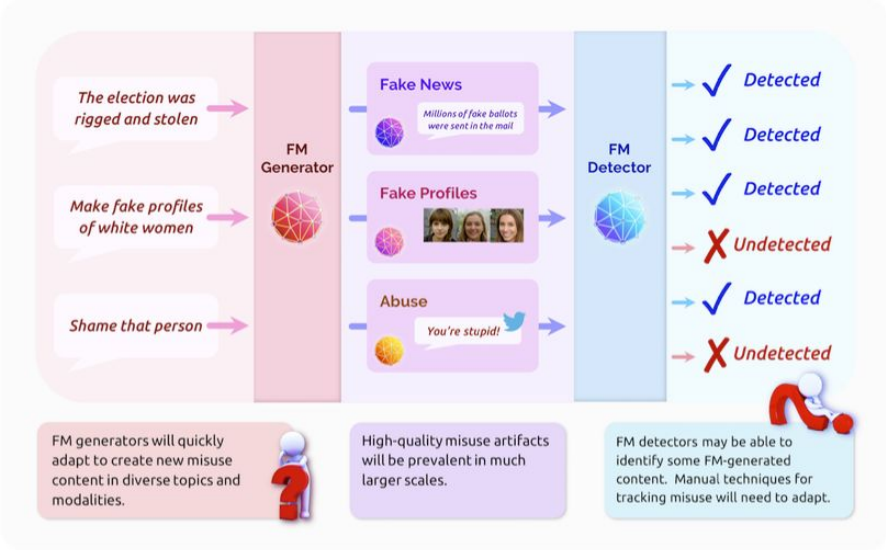


Society

- 불평등과 공정성, 오용, 환경, 경제, 법, 규모의 윤리 등 여러 측면에서 분석
- **불평등과 공정성**
  - 훈련 데이터에 존재하는 편향을 학습하고 증폭시켜, 사회적 불평등을 심화할 수 있음 (내재적 편향)
  - 외재적 피해는 모델의 적용 결과 발생하는 피해로 차별, 불공정, 사회적 배제가 있음
  - 해결 방법
    - 다양하고 대표성 있는 데이터 수집, 편향된 데이터 제거
    - 편향 완화 알고리즘 개발
    - 모델의 한계, 편향, 불확실성을 명확하게 공개

Society

- **Misuse (오용)**
  - 인간이 생성한 것과 구분이 어려운 가짜 뉴스, 딥페이크 등의 생성
  - 낮은 비용으로 대량의 콘텐츠 생성 가능
  - Foundation 모델이 이를 감지하도록 훈련 가능



정리

- Foundation 모델이라는 새로운 AI 패러다임의 등장과 그로 인한 기회와 위험을 종합적으로 분석하고, 미래 연구 방향을 제시
- Foundation 모델의 핵심 특징 2가지: Emergence와 Homogenization
- 기술, 사회, 윤리적 측면을 종합적으로 고려하여 다학제적 연구와 책임 있는 연구를 통해 잠재력 극대화 및 위험의 최소화 필요
- 연구자, 개발자, 사용자, 정부 등의 협력이 필요