# FlowMate 보고서: LLM 튜닝 및 모델 경량화 기술 분석

## 목차

1. 서론

2. 주제와 목적

3. 주요 내용 요약

4. 결론 및 권고사항

5. 참고문헌

---

**1. 서론**

이 보고서는 FlowMate가 LLM 튜닝과 모델 경량화 기술에 대한 분석을 통해 사내 업무 효율성을 높이는 방법을 탐색하고자 합니다. 주요 내용은 파라미터 효율적 튜닝, LoRA와 양자화, 프루닝 및 증류 등의 기법을 중심으로 진행됩니다.

**2. 주제와 목적**

**주제:**

- LLM 튜닝: LoRA를 활용한 LLM 튜닝

- 모델 경량화: DistilBERT와 같은 모델 축소 기술

**목적:**

- 다양한 기법의 효과와 한계를 이해하고 적용 가능성을 평가합니다.

- 실제 사례 연구를 통해 기술의 실무적 활용 방안을 탐색합니다.

**3. 주요 내용 요약**

**1주차: 개요 및 기본 개념 소개**

- **과정 오리엔테이션**: 파라미터 효율적 튜닝 개요와 LoRA 기본 개념 소개

- **사전 실습**: Hugging Face 모델 불러오기 및 테스트

**2주차: LoRA 동작 원리 심화 및 사례 연구**

- **LoRA 동작 원리**: 저랭크 행렬 업데이트와 Hugging Face PEFT 활용

- **사례 연구**: LoRA의 장점과 한계, BERT 모델에 LoRA 적용하여 문서 분류 미세조정

**10주차: 모델 경량화 종합 분석 및 사례 연구**

- **모델 경량화 기법 종합 정리**: LoRA, 양자화, 프루닝, 증류 각각의 장단점 비교

- **사례 연구**: DistilBERT가 BERT 대비 40% 가볍고 97% 성능을 달성한 사례

**11주차: LoRA 활용 사례 분석 및 실습**

- **실제 사례 1**: Alpaca 모델이 LLaMA 7B에 LoRA로 미세조정되어 ChatGPT 유사 성능을 얻은 과정

- **실습**: 소규모 LLM LoRA 튜닝 과정 실습

**12주차: DistilBERT 및 모델 최적화 사례 분석**

- **DistilBERT 상세 사례 분석**: BERT 대비 40% 파라미터 감소, 60% 추론 속도 향상

- **TinyBERT 사례**: 교사의 중간표현까지 학습하는 4층 학생모델

**4. 결론 및 권고사항**

**결론:**

- LoRA와 양자화, 프루닝, 증류 등의 기법은 대형 모델을 효과적으로 튜닝하고 경량화할 수 있는 중요한 도구입니다.

- 실제 사례 연구를 통해 이러한 기술의 적용 가능성을 확인할 수 있습니다.

**권고사항:**

1. **기술평가**: 다양한 기법을 평가하여 가장 적합한 방법을 선택합니다.

2. **실무적 적용**: 실제 업무 환경에서 LLM 튜닝과 모델 경량화를 시도해 보며 성능을 검증합니다.

3. **커뮤니티 활용**: LoRA와 같은 기술의 커뮤니티 활용을 통해 최신 동향을 파악하고, 실용적인 노하우를 얻습니다.

**5. 참고문헌**

- [참고문헌 목록]

---

이 보고서는 FlowMate가 LLM 튜닝과 모델 경량화 기술에 대한 체계적 분석을 통해 사내 업무 효율성을 높이는 방법을 탐색하는 데 도움을 드립니다. 다양한 사례 연구와 실습을 통해 실제 적용 가능성을 확인하고, 최적의 기법을 선택하여 사내 업무를 더욱 효과적으로 수행할 수 있도록 권고합니다.