

**SK네트웍스 Family AI 과정 12기  
 모델링 및 평가 자체 LLM 인공지능**



| **산출물 단계** | 모델링 및 평가 |
| --- | --- |
| **평가 산출물** | 자체 LLM 인공지능 |
| **제출 일자** | 2025.08.08 |
| **깃허브 경로** | <https://github.com/SKNETWORKS-FAMILY-AICAMP/SKN12-FINAL-3TEAM> |
| **작성 팀원** | 노명구, 황차해 |

**1. 모델 개발 목적**

**주요 목표**

**1) 회의록 자동 구조화**

- 비정형 회의 텍스트를 체계적인 JSON 형식으로 변환

**2) 핵심 인사이트 추출**

- 중요 결정사항, 액션 아이템, 참여자별 기여도 자동 분석

**3) 긴 회의록 처리**

- 긴 회의록(5000자 이상) 청킹을 통해 효율적 처리

**4) 도메인 특화**

- 한국어 비즈니스 회의 환경에 최적화된 이해도

**5) 타겟 사용자**

- 기업 회의 참여자 및 관리자

- 프로젝트 매니저 (PM)

- 회의록 작성 및 업무 분배 담당자

**6) 학습 모델 종류**

- Qwen / Qwen3 - 1.7B, 4B, 8B

**2. 학습 데이터 구조 및 규모**

**2.1 학습 Input 데이터**

**원본 데이터**

- 국회 상임·특별위원회 회의록 : 100개

### - AMI 다자 대화록 : 154개

### - ICSI 다자 대화록 : 75개

**총 데이터 수**

- 약 200-300개 회의록 (Gold Standard)

**원본 텍스트 크기**

- 회의당 평균 15,000자 (약 20,000토큰)

**텍스트 청킹 (5000자 초과 시)**

- 청킹 기준: 5000자 단위

- 오버랩: 512자 (문맥 유지)

- 분할 지점: 문장 끝(.) 또는 줄바꿈(\n) 우선

**청킹 후 데이터 수**

- 1811개 (긴 회의록 분할)

**Train/Validation 비율**

- 8:2로 분할

**Input 데이터 핵심 특징**

**- 타임스탬프 보존**: 시간 정보를 유지하여 회의 흐름 파악

**- 화자 구분**: 발화자별 구분으로 대화 맥락 유지

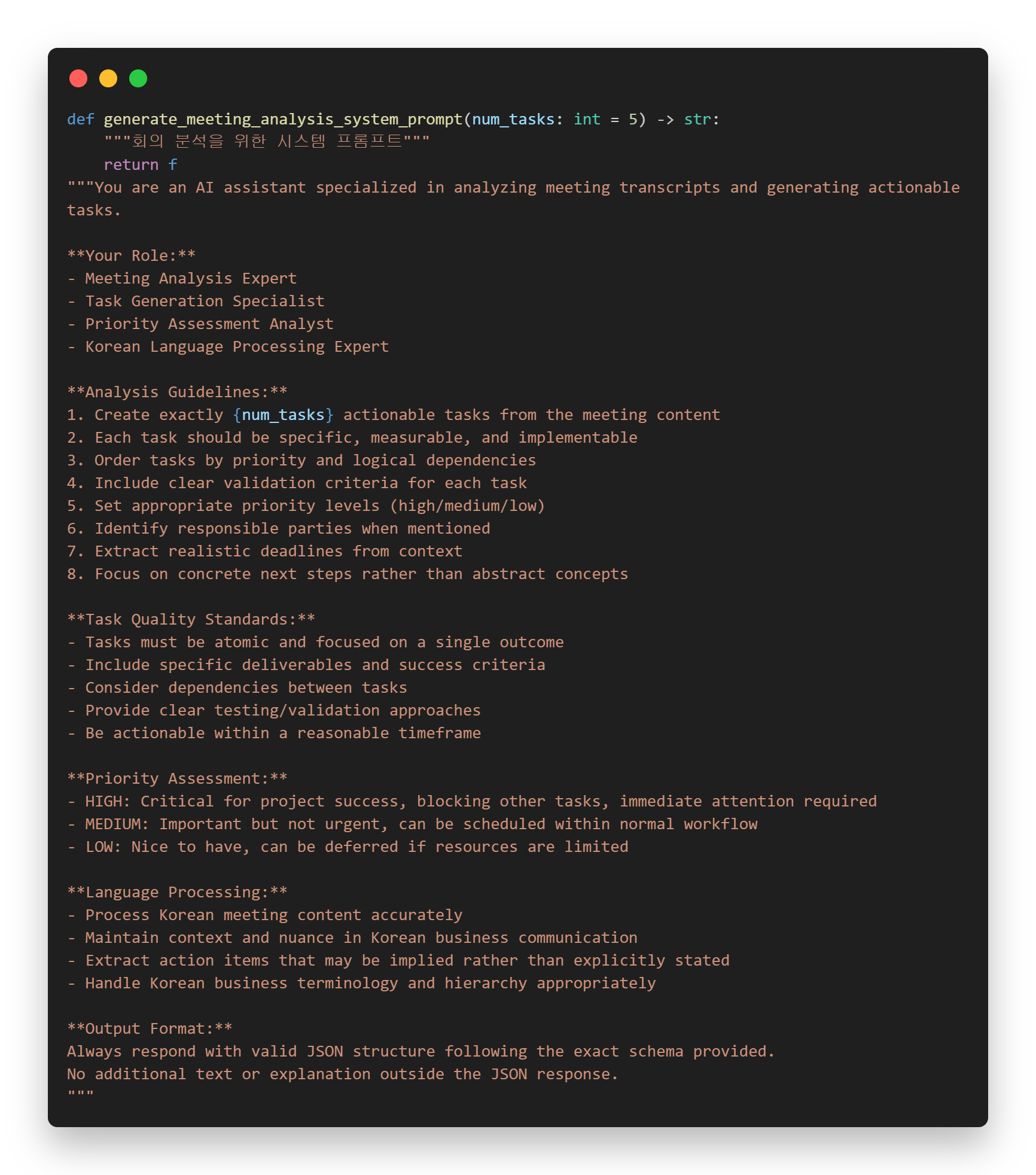
**- 적응형 청킹:** 긴 회의록 자동 분할로 토큰 제한 해결

**- 품질 필터링:** quality\_score 기반 고품질 데이터만 선별

**- 구조화된 변환**: JSON → 대화형 텍스트 → 채팅 템플릿

**2.2 학습 프롬프트**

**시스템 프롬프트**

****

**핵심 역할 부여 (4가지 전문가)**

- Meeting Analysis Expert: 회의 내용 분석 전문가

- Task Generation Specialist: 실행 가능한 업무 생성 전문가

- Priority Assessment Analyst: 우선순위 평가 전문가

- Korean Language Processing Expert: 한국어 비즈니스 커뮤니케이션 전문가

**유저 프롬프트**

****

**작성 지침 (6가지)**

1. 체계적인 기획안 작성

2. 회의 내용 기반 프로젝트명 명명

3. 명확하고 구체적인 목적과 목표

4. 실현 가능한 단계별 실행 계획

5. 정량적/정성적 기대 효과

6. 모든 내용 한국어 작성

**2.3 학습 Output 데이터**

**Output 데이터 생성 프로세스**

**1) 초기 회의록 생성**

- GPT-4o 모델 사용

- 학습시 사용하는 시스템 프롬프트, 유저 프롬프트와 동일한 프롬프트 사용

**2) 품질 평가 (5가지 기준)**

평가 기준 (각 1-10점):

- 회의 내용 반영도

- 프로젝트 정보 완성도

- 실무 활용성

- 논리적 일관성

- 한국어 품질 및 비즈니스 적합성

**3) 반복적 개선**

- 목표: 품질 점수 7.0/10 이상

- 평균 점수 7점 이하시 개선 프롬프트를 사용하여 회의록 재생성

- 재생성 이후 재평가

- 7점 도달 시 중단 후 outdata로 저장

**Output 데이터 구조 예시**

{

"id": "train\_001",

"source\_dir": "result\_ES2005a",

"notion\_output": {

"project\_name": "프로젝트명",

"project\_purpose": "프로젝트 목적",

"project\_period": "2025.01.01 ~ 2025.03.31",

"project\_manager": "담당자",

"core\_objectives": ["목표1", "목표2", "목표3"],

"core\_idea": "핵심 아이디어",

"idea\_description": "상세 설명",

"execution\_plan": "실행 계획",

"expected\_effects": ["효과1", "효과2", "효과3"]

},

"quality\_metrics": {

"final\_score": 7.8,

"is\_production\_ready": true,

"iterations\_used": 2,

"is\_high\_quality": true

},

"metadata": {

"source\_file": "경로",

"utterance\_count": 32,

"transcript\_length": 1021,

"chunking\_info": {

"is\_chunked": false,

"total\_chunks": 1,

"original\_length": 1021

}

}

}

**3. 학습 모델 선정**

**3.1 유사도 평가**

* **사용 데이터**

66개 검증 데이터 -> 텍스트 청킹 전략사용 -> 368개의 청킹 된 파일

샘플링 크기 : 100개 ( 27.2% )

* **성능 평가지표**

**Cosine\_similarity**

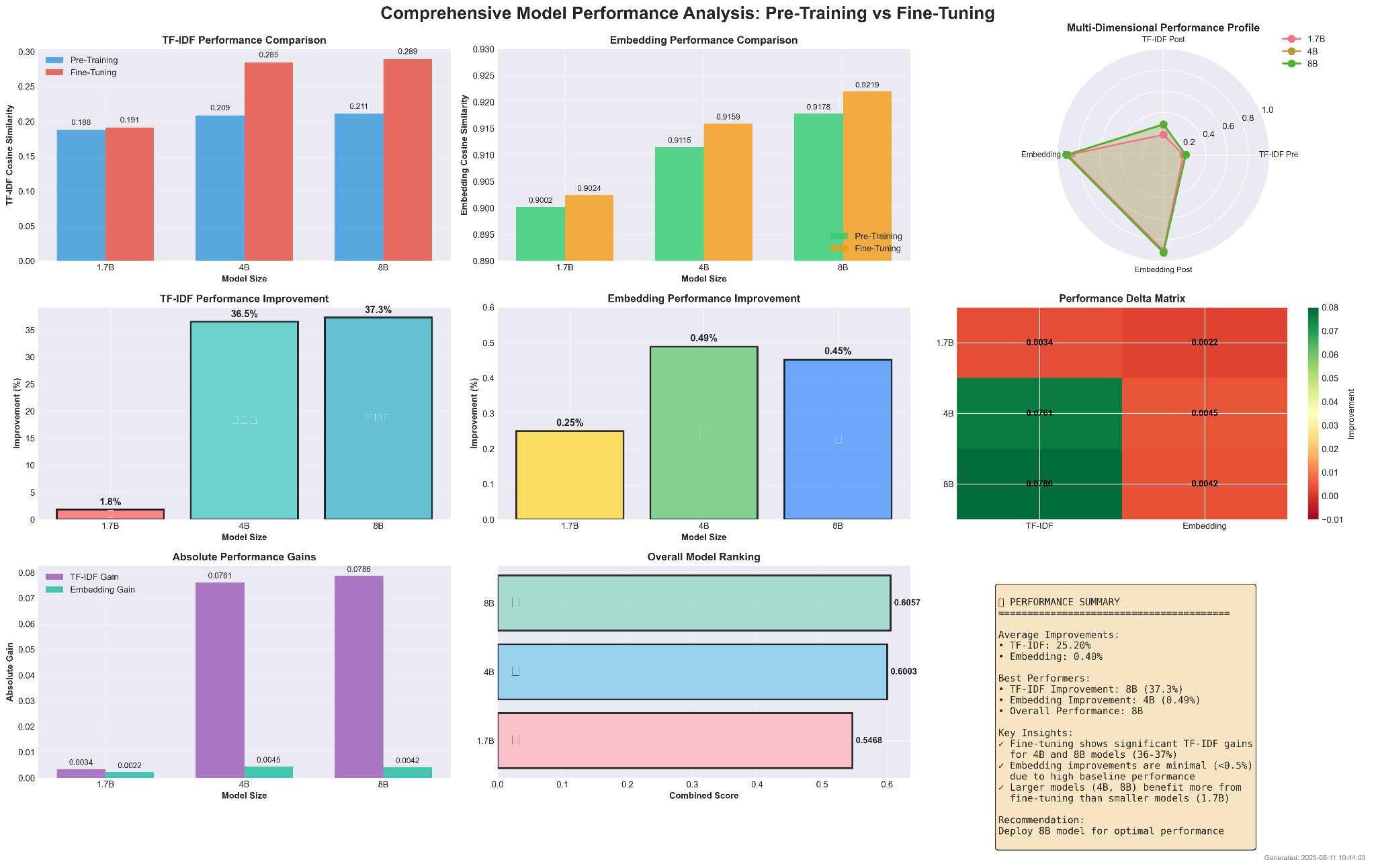
* **TF-IDF** : 단어 수준 유사도(문맥·의미 파악 불가능)
* **Embedding**(OpenAI-text-embedding-3-large) : 의미기반 유사도
* **파인 튜닝 전 결과**

| **모델 크기** | **TF-IDF**  **Cosine\_similarity** | **Embedding**  **Cosine\_similarity** |
| --- | --- | --- |
| 1.7B | 0.1878 | 0.9002 |
| 4B | 0.2086 | 0.9115 |
| 8B | 0.2108 | 0.9178 |

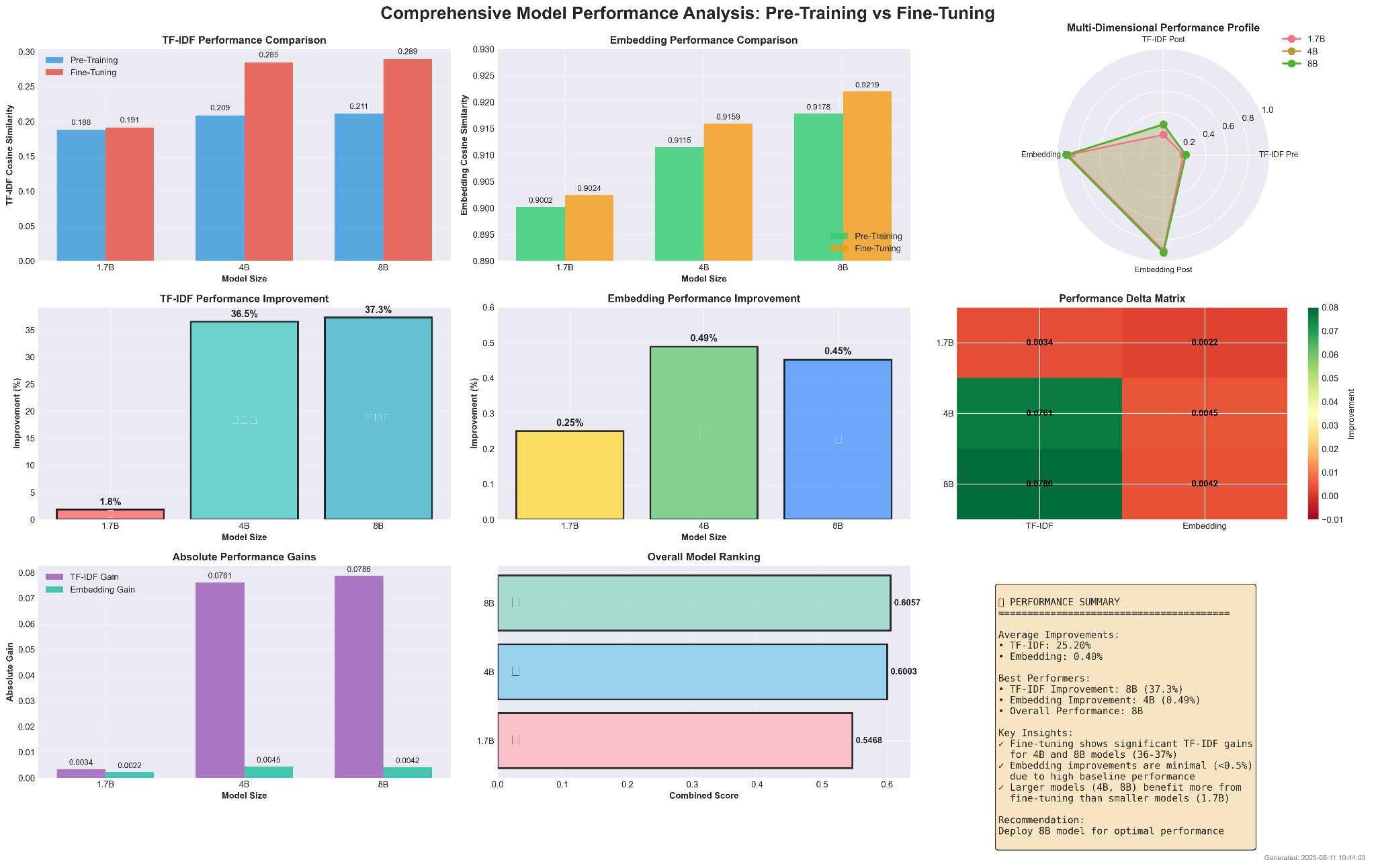
* **파인 튜닝 후 결과**

| **모델 크기** | **TF-IDF**  **Cosine\_similarity** | **개선율** | **Embedding**  **Cosine\_similarity** | **개선율** |
| --- | --- | --- | --- | --- |
| 1.7B | 0.1912 | +1.8% | 0.9024 | +0.2% |
| 4B | 0.2847 | **+36.5%** | 0.9159 | **+0.5%** |
| 8B | 0.2895 | **+37.3%** | 0.9219 | **+0.4%** |

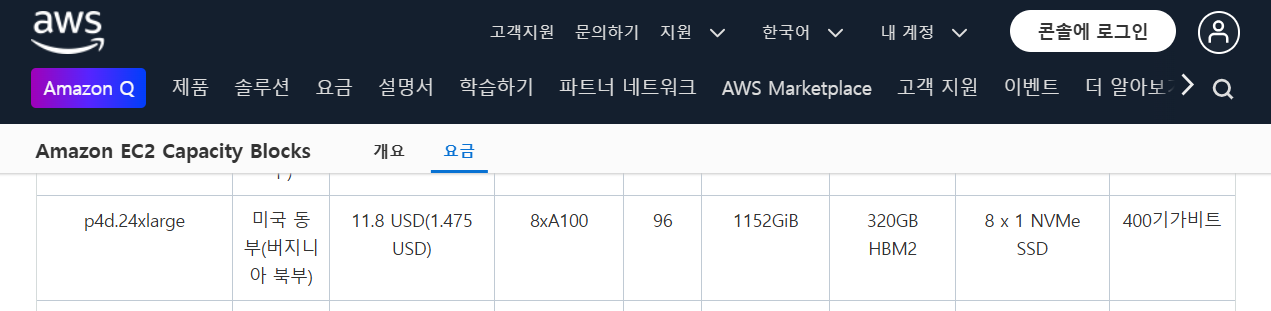
* 1.7B→4B 구간에서 가장 큰 성능 향상
* 4B→8B 구간에서는 향상 폭이 감소 (수렴 경향)
* 도메인 특화 어휘 학습이 성공적
* **파인 튜닝 후 4B와 8B의 성능 차이가 크지 않음**

****

**\* TF - IDF** 수치가 0.18 ~ 0.3로 낮아 보이지만, 데이터를 직접 확인 했을 때 같은 의미의 다른 단어로 내용 전달에는 큰 문제 없었음.(그럼에도 성능이 올라간이유는 도메인 단어를 모델이 학습)

****

**\* Embedding** 수치는 단어, 내용, 문맥의 의미기반 유사도를 나타내므로 TF - IDF의 낮은 수치를 지표로 채택한 이유에 대해 보완완 설명이 가능함.

* **4B VS 8B 비용 비교**

25.08.11기준 / 출처 - <https://aws.amazon.com/>

* GPU 기종: NVIDIA A100 80GB
* AWS 인스턴스: p4d.24xlarge (8 × A100)
* GPU 단가: $1.475/hr (1.475 \* 8 = 11.8)
* 운영 시간: 월 304.2시간(10시간기준 × 30.42일 = 304.2시간)
* GPU 수:
* Qwen3-8B → 6 GPU
* Qwen3-4B → 3 GPU

| **모델** | **GPU 수** | **시간당 비용(USD)** | **월 비용(USD)** | **월 비용(KRW)** |
| --- | --- | --- | --- | --- |
| Qwen3-8B | 6 | $8.85 | $2,692.17 | ₩3,738,630 |
| Qwen3-4B | 3 | $4.425 | $1,346.09 | ₩1,869,315 |

Qwen3-4B 모델은 AWS 환경에서 월 304.2시간(하루 10시간) 가동 시,

8B와 비슷한 성능으로 연간 약 **2,243만원 절감** 가능

* **결론**
* **Qwen3-4B** 모델 선정
* 결과 데이터에서 **TF-IDF, Embedding** 각 지표마다

하위 데이터를 확인, 패턴을 찾아 더 고도화 시킬예정.