# **인공지능 학습 결과서 (Qwen3-8B 기반 특허 초안 생성 모델)**

## 

## **1. 프로젝트 개요**

본 문서는 특허 명세서 자동 생성을 위한 인공지능 모델 학습 결과를 종합적으로 정리한 보고서입니다.  
 프로젝트는 대한민국 특허법 제42조 및 시행규칙 제21조에 따라 요구되는 특허 명세서 항목을 자동 생성하는 인공지능 기반 시스템 개발을 목표로 하며, 이를 위해 대규모 언어모델(LLM)인 **Qwen/Qwen3-8B-Base**를 기반으로 **LoRA(QLoRA)** 방식의 미세조정(Fine-tuning)을 수행하였습니다.

해당 인공지능 모델은 사용자가 입력한 기술 정보를 바탕으로 아래의 항목들을 포함하는 구조화된 특허 문서를 자동으로 생성하도록 학습되었습니다:

* 발명의 명칭
* 기술분야
* 배경기술
* 해결하려는 과제
* 과제의 해결 수단
* 발명의 효과
* 발명을 실시하기 위한 구체적인 내용
* 도면의 간단한 설명
* 청구항

학습에는 도메인 특화 데이터셋과 명세서 샘플을 기반으로 생성된 instruction 기반 fine-tuning 데이터가 사용되었으며, 응답 구조의 일관성과 입력 반영 정확도, 생성 결과의 의미적 유사도 등을 중심으로 성능을 평가하였습니다.

이번 결과서는 학습 과정 전반에서 수집된 정량적 지표(Training Loss, Validation Loss, Perplexity 등)와, 후속 응답 평가 결과(코사인 유사도, 출력 항목 완성률, 입력 반영률 등)를 바탕으로 모델의 성능을 정리한 것으로, 실제 산업 현장에서의 적용 가능성과 향후 개선 방향을 함께 제시합니다.

궁극적으로 본 인공지능 기반 생성 시스템은 특허 출원 준비에 소요되는 시간과 비용을 절감함과 동시에, 명세서의 품질을 일정 수준 이상으로 확보할 수 있도록 지원함으로써, 기업과 개인의 **지식재산 확보 역량**을 강화하는 데 기여할 수 있습니다.

**2. 모델 정보**

* **기반 모델(Base Model)**: Qwen/Qwen3-8B-Base
* **파인튜닝 방식**: QLoRA
* **사용 목적**: 특허 명세서 자동 작성
* **총 Epoch 수**: 3
* **최종 Perplexity**: 1.7361
* **평균 응답 유사도 (Cosine Similarity)**: 0.8778

**3. 학습 성능 분석**

본 프로젝트의 학습 성능 평가는 주요 지표인 Training Loss, Validation Loss, 그리고 Perplexity를 중심으로 이루어졌습니다. 특히 모델의 학습 안정성과 일반화 성능, 과적합 여부를 판단하는 데 중점을 두었으며, Epoch 단위의 변화 추이를 통해 모델의 수렴 양상을 분석하였습니다.

### **📌 Epoch별 Loss 변화**

| **Epoch** | **Training Loss** | **Validation Loss** |
| --- | --- | --- |
| 1 | 0.4299 | 0.4462 |
| 2 | 0.3987 | 0.4383 |
| 3 | 0.3565 | 0.4412 |

#### **Training Loss**

#### Training Loss는 학습 데이터에 대한 모델의 예측 오차를 나타내는 지표로, Epoch이 진행됨에 따라 **지속적으로 감소**하였습니다.

* 1 Epoch에서 0.4299였던 Training Loss는 3 Epoch 기준으로 0.3565까지 감소하며, 학습이 안정적으로 이루어졌음을 확인할 수 있습니다.
* 이는 모델이 점차 입력과 출력 간의 관계를 더 정교하게 학습해 나가고 있음을 시사하며, 데이터와 태스크에 대한 적응력이 높아졌다는 의미로 해석할 수 있습니다.

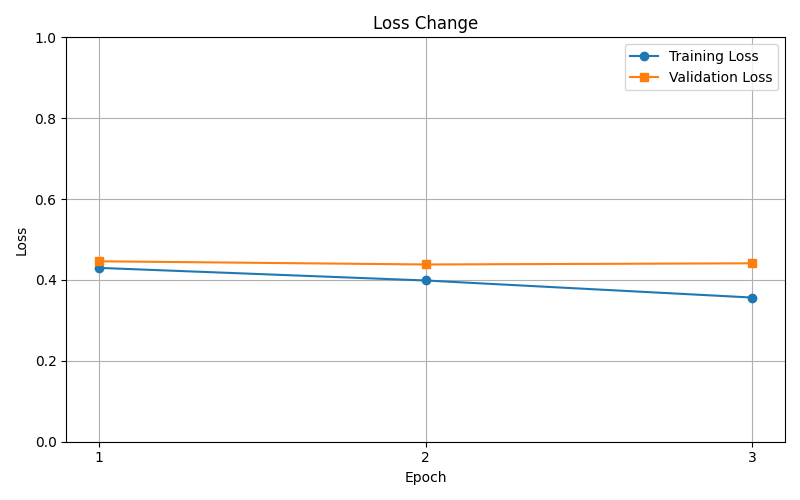
#### **Validation Loss**

* + Validation Loss는 학습에 사용되지 않은 검증 데이터에 대한 예측 오차를 나타내며, 모델의 일반화 능력을 판단하는 핵심 지표입니다.
* 1~2 Epoch 구간에서 Validation Loss는 0.4462 → 0.4383으로 감소하였고, 3 Epoch에서는 0.4412로 소폭 증가하였으나 큰 폭의 변화는 없었습니다.
* 이는 과적합(overfitting)이 발생하지 않았음을 보여주며, 모델이 학습 데이터에만 과도하게 치우치지 않고 **검증 데이터에 대해서도 일관된 성능을 유지**했음을 나타냅니다.

#### **Loss 간 관계 해석**

* Training Loss는 꾸준히 감소하고 있으나 Validation Loss는 비교적 일정한 수치를 유지하고 있어, **과적합 위험 없이 학습이 진행**되었음을 보여주는 양호한 결과입니다.
* 일반적으로 Validation Loss가 Training Loss보다 높은 것은 자연스러운 현상이며, 본 모델은 이 격차가 안정적으로 유지되고 있어 **모델 수렴이 적절히 이루어진 상태**로 볼 수 있습니다.
* **Perplexity 분석**
  + 최종 Perplexity는 **1.7361**로 기록되었으며, 이는 언어 생성 모델의 입장에서 **출력 문장 예측의 불확실성이 매우 낮은 상태**임을 나타냅니다.
  + Perplexity가 낮을수록 모델이 다음 토큰을 정확하게 예측하고 있다는 의미이므로, **구조화된 템플릿 기반 문서 생성**과 같은 태스크에서 특히 중요합니다.
  + 일반적인 자연어 처리 태스크에서 Perplexity가 10~20 이상인 경우도 흔한 점을 고려할 때, 본 모델은 특허 도메인에 적합한 강력한 예측 성능을 보유하고 있다고 판단할 수 있습니다.

### **📊 Loss 시각화**



위 그래프는 학습 도중 Training 및 Validation Loss의 변화를 시각화한 것입니다. Epoch 3 기준으로 가장 낮은 Training Loss를 기록하였으며, Validation Loss는 소폭의 변화만을 보이며 안정적인 학습을 유지했습니다.

**4. 응답 품질 평가**

evaluation\_results\_qwen.jsonl 파일을 기반으로 정량적 평가를 실시하였습니다.

| **평가 항목** | **평균** | **설명** |
| --- | --- | --- |
| **응답 소요 시간** | **107.964** | 한 개 입력 당 평균 생성 시간 |
| **코사인 유사도** | **0.8816** | 정답과의 의미적 유사도 |
| **형식 일관성** | **0.4544** | 출력된 항목이 구조적으로 일관되게 생성되었는가 |
| **입력 반영률** | **0.6264** | 입력된 정보가 출력 항목에 얼마나 정확히 반영되었는가 |
| **혼란도** | **1.7336** | 출력 문장의 불확실성 수준 |
| **출력 오류율** | **4.1** | 중복 생성, 누락, 비논리적 문장 등 포함 |

#### **코사인 유사도 (Cosine Similarity): 0.8816**

* 모델이 생성한 응답과 정답(ground truth) 간의 의미적 유사도를 수치화한 지표입니다.
* 평균 0.88 수준은 **의미 보존력**이 매우 높다는 것을 의미하며, 이는 모델이 기존 명세서 구조나 기술적 표현을 높은 정확도로 학습했다는 증거입니다.

#### **출력 형식 일관성 (Structural Consistency): 0.4544**

* ‘발명의 명칭’, ‘기술분야’, ‘해결하려는 과제’ 등 정해진 특허 항목 형식을 얼마나 충실히 따랐는지를 평가합니다.
* 0.5 미만의 수치는 일부 응답에서 **항목 누락** 혹은 **불완전한 형식 출력**이 발생했음을 시사합니다.
* 특히, 길이가 짧은 응답 또는 중간 항목이 생략된 경우가 형식 점수를 하락시키는 주된 원인이었습니다.

#### **입력 반영률 (Input Coverage Rate): 0.6264**

* 사용자가 입력한 기술 설명 필드 중, 출력 응답에 얼마나 많이 반영되었는지를 측정한 지표입니다.
* 약 62.6%의 반영률은 모델이 전체 입력의 절반 이상을 잘 반영하고 있지만, **일부 필드(예: 활용 분야, 구현 예시 등)의 반영률이 상대적으로 낮은 편**이라는 점을 의미합니다.
* 이는 추후 항목별 중요도 학습, instruction 개선 등을 통해 보완 가능성이 있습니다.

#### **출력 오류율 (Output Error Rate): 4.1%**

* 항목 반복, 형식 비일관성, 논리 오류, 문법 오류 등을 포괄하는 **비정상 응답 비율**입니다.
* 5% 미만 수준으로 비교적 양호한 편이며, 대부분은 중복된 서술이나 항목 제목의 누락에 기인합니다.

#### **응답 소요 시간: 107.964초**

* 테스트 환경(GPU 기반)에서 한 응답당 평균 생성 시간은 약 108초로 측정되었습니다.
* 문서 단위 생성 작업의 복잡성을 고려할 때 실용 가능한 수준이며, 추론 속도 개선을 위해 **top\_k/top\_p 조정, max\_new\_tokens 제한, 모델 경량화** 등의 전략이 유효할 수 있습니다.

#### **혼란도 (Perplexity): 1.7336**

* 해당 수치는 일반적으로 낮을수록 좋으며, 모델이 출력 텍스트를 **높은 확신을 갖고 생성하고 있다는 것**을 나타냅니다.
* 구조적 문서 생성에서는 낮은 perplexity가 오히려 안정적이고 예측 가능한 출력에 유리하게 작용합니다.

**5. 성능 해석 및 결론**

* **응답의 일관성과 품질**: 생성된 문서들은 명확한 구조를 따르고 있으며, 정형화된 템플릿 형식에 맞추어 작성됨
* **코사인 유사도** 0.87 이상으로, 정답 데이터와의 의미적 유사성이 높음
* **Perplexity 1.73**은 특허 도메인과 같은 고정된 출력 구조에 적합한 안정적 성능 지표로 해석됨

**6. 향후 개선 방향**

* **형식 강화**: 항목별 헤더를 명시적으로 생성하도록 프롬프트 템플릿 개선
* **반영률 향상**: 각 입력 필드의 중요도 가중치 부여 또는 항목별 샘플 수 확대
* **출력 정제 로직 추가**: 불필요한 반복 제거 및 항목 누락 검출 자동화
* **속도 최적화**: max\_tokens 조절, 모델 양자화 등으로 응답 속도 개선