M.A.R.S. 2025
SNUBH 의무기록 생성 데이터톤

Medical **A**uto-documentation with **R**eal-world **S**tructuring:
LLM Clinical Note Generation Challenge

예선 결과 보고서 [원피스]

주최/주관 의료인공지능센터 Center for Artificial Intelligence in Healthcare, SNUH 분당서울대학교병원 빅데이터센터, AI NATION

후원 aws AWS코리아, GBSA 경기도경제과학진흥원, IITP 정보통신기획평가원, KHIDI 한국보건산업진흥원

1. 팀 소개

- 팀 소개 및 역할

  김충만 : 팀원, Task C

  신대현 : 팀장, Task B

  정승환 : 팀원, Task A

- 잦은 API오류로 인하여 Task A,B,C를 각자가 맡아서 담당하였습니다.

2. 문제 정의 & 접근 개요

- Task A/B/C 접근 방법 요약

Task A :
환자의 긴 의료 기록을 기반으로 입원부터 퇴원까지의 중요한 임상 사건과 치료 과정을 간결하고 밀도 높게 요약하는 작업입니다. 이 요약문은 주어진 환자 기록에만 의존하며, 기록되지 않은 정보는 절대 추가하지 않습니다.

이를 통해 의료 현장에서 효율적이고 정확한 환자 입원 경과 파악을 지원합니다.
본 과제는 특히 복잡한 의료 데이터를 체계적으로 압축하고 의료용어를 적절히 사용함으로써 의료 전문가가 신속하게 환자 상태를 이해할 수 있도록 돕는 데 중점을 둡니다.

2. 문제 정의 & 접근 개요

- Task A/B/C 접근 방법 요약

Task B :

Radiology Impression 요약의 목표는 주어진 방사선 검사 보고서의 FINDINGS 섹션을 철저히 분석하여, 임상적으로 타당하고 의학적으로 정확하며 간결한 Radiology Impression을 생성하는 것입니다.
이를 통해 임상의가 이해하기 쉽고, 감별 진단 및 후속 검사에 대한 권고를 내릴 수 있도록 핵심 정보를 정확히 전달하는 것을 목적으로 접근 했습니다.

필요한 핵심 임상 소견을 추출하고, 불필요한 정보나 노이즈를 제거하여 정확한 요약문을 생성해
프롬프트에 정교한 지침을 제공함으로써 환각을 방지하고 핵심 내용 중심의 요약을 유도하고 전문 용어 사용과 임상적 의미 반영에 중점을 두어 실제 의료 현장에서 활용 가능한 결과를 도출하는 것을 목표로 합니다.

2. 문제 정의 & 접근 개요

- Task A/B/C 접근 방법 요약

Task C :
Task C는 환자의 퇴원 요약 정보를 바탕으로 ICD-10 코드를 정확히 예측하는 과제입니다.
특히 주 진단(Principal Diagnosis)과 부 진단(Secondary Diagnoses)을 구분하여 적절한 코드를 생성해야 합니다. 이는 의료 기관의 진단 코딩 업무를 자동화하고, 정확한 의료 기록 관리 및 보험 청구 과정을 효율화하는 데 중요한 역할을 합니다.

이 과제의 핵심은 비정형 텍스트에서 질병, 증상, 합병증 등의 정보를 정확히 추출하고, 이를 표준화된 ICD 코드 체계로 변환하는 능력에 있습니다. 이는 단순한 텍스트 요약이 아닌, 의학적 지식과 코딩 규칙에 대한 깊은 이해를 필요로 합니다.

3. 프롬프트 설계 방법

- 사용 모델 : LGAI-EXAONE/EXAONE-3.5-7.8B-Instruct-AWQ, meta-llama/Llama-3.1-8B-Instruct
- 프롬프트 아이디어 및 전략
  1. '의료관련 전문가' 역할설정
  2. 명확한 목표 설정과 단계별 프로세스 진행
  3. 샘플 제시를 통한 출력 포맷을 설정 후 결과 유사성 증가 기대

Task A :

LGAI-EXAONE/EXAONE-3.5-7.8B-Instruct-AWQ 모델을 채택하였으며, 약 0.74의 BertScore를 기록하여 요약 작업에 최적화되어 있는 것을 확인했습니다. 프롬프트 설계 시에는 경험이 풍부한 의료 전문가 역할을 부여함으로써 전문적인 분석이 이루어지도록 했고, 외부 추론이나 추가 정보의 사용을 명확히 금지하여 환각 현상을 방지하는 데 중점을 두었습니다.

또한 최대 6문장으로 요약을 제한해 간결함을 유지했으며, 의료 기록 내 핵심 용어 중 다섯 개(예: 심부정맥 혈전증, 신부전, 간경화, 고혈압, 고지혈증)의 예시를 넣어 최대한 포함하도록 하여 내용의 정확성과 전문성을 높였습니다. 여기에 구체적인 입력 및 출력 예시를 함께 제공함으로써 대형 언어 모델이 생성 방향을 명확히 이해하고 따라가도록 유도하였습니다.

3. 프롬프트 설계 방법

- 사용 모델
- 프롬프트 아이디어 및 전략

Task B :

meta-llama/Llama-3.1-8B-Instruct 모델을 활용하여 Impression 요약 작업을 수행하였다. 프롬프트는 대형 언어 모델에 "환자 진료에 필수적인 통찰력을 제공하는 최고 수준의 방사선과 전문의 AI"라는 역할을 부여하여, 임상적으로 타당하고 정확하며 간결한 Radiology Impression을 생성하도록 설계하였습니다.

구체적인 전략으로는, 보고서의 FINDINGS 섹션에 명시된 사실만을 분석하도록 하여 환각과 외부 정보 개입을 철저히 차단하였으며, Impression은 진단적 가치가 높은 1~2개의 핵심 소견만으로 구성되도록 제한하여 불필요한 설명이나 장황한 문장을 제거하였습니다.

또한, 의학적으로 정확한 전문 용어를 필수적으로 사용하고 모호한 주관적인 표현을 배제하여 임상적 의미와 진단적 중요도가 명확히 드러나도록 하였습니다. 아울러 FINDINGS 섹션에서 진단에 결정적인 영향을 미치는 소견을 선별해 우선순위에 따라 기술하도록 하였으며, 최종 결과물의 일관성과 가독성을 보장하기 위해 "최우선 핵심 소견", "추가 또는 동반 소견", "감별 진단 또는 후속 검사 권고"와 같은 세부 항목을 포함한 출력 형식을 엄격히 준수하도록 하였습니다.

3. 프롬프트 설계 방법

- 사용 모델
- 프롬프트 아이디어 및 전략

Task C :

Task C에서는 의료 용어,핵심 의학 정보를 추출하는 텍스트 분석 및 코드 매핑 작업에 적합한 ICD 코드 예측을 위해 Llama-3.1-8B-Instruct 모델을 활용하였습니다. 프롬프트 설계 시에는 "경험이 풍부한 의료 코딩 전문가"라는 역할을 부여하여 전문적인 관점에서 진단 코드를 예측하도록 유도했습니다.

핵심 전략으로 퇴원 요약에서 주요 질병 및 증상을 식별하고 적절한 ICD 코드로 변환하는 과정을 안내하는 단계적 분석 유도, 주 진단과 부 진단 간의 분류 기준을 명확히 제시하여 코드 분류의 정확성을 높이는 주/부 진단,구분을 명확히 적용하여 불필요한 설명 없이 코드만 출력,구분하여 형식을 엄격 지정했으며, 실제 사례를 통해 퇴원 요약과 예측된 ICD 코드의 관계를 명확히 보여주는 구체적 예시 제공으로 모델의 이해를 돕고 일관되게 유도했습니다.

마지막으로, 퇴원 요약에 언급되지 않은 코드는 예측하지 않도록 명시적 제한 설정하여 환각(hallucination) 방지에 중점을 두었습니다.

4. 실험 결과 & 분석

- 리더보드 성적, 자체 지표 요약
- 시도한 방법의 효과와 한계
- 임상적 활용 가능성 및 개선/응용 아이디어

Task A :
Task A 실험 결과를 종합하면 BERTScore는 약 0.74로 상당히 준수한 성과를 보였습니다. 그러나 LLM 평가에서는 전반적으로 절반 이하의 점수를 받아 일부 지표에서 개선의 여지가 크게 나타났습니다. 이 차이는 모델이 텍스트의 의미적 유사성은 비교적 잘 포착하고 있으나, 임상적 정확성, 간결성, 요약의 완성도 측면에서는 아직 부족함을 시사합니다. 따라서 향후 작업에서는 핵심 임상 정보의 누락을 줄이고, 표현의 명확성과 간결성을 더 강화하는 프롬프트 설계가 필요합니다. 또한 LLM 평가 결과를 면밀히 분석하여 구체적인 오류 유형과 패턴을 파악함으로써 보다 효과적인 개선 전략을 수립할 계획입니다.

4. 실험 결과 & 분석

- 리더보드 성적, 자체 지표 요약
- 시도한 방법의 효과와 한계
- 임상적 활용 가능성 및 개선/응용 아이디어

Task B :
리더보드 성적에서 Quantitative 점수는 총 6점 만점 중 4.3점을 기록하였으며, 세부적으로 BERTScore는 2.3/3, Fairness는 2.0/3으로 나타났다. LLM Evaluation에서는 총 10점 만점 중 6.6점을 획득했으며, 항목별로는 Summary 1.3/3, Clinical 2.4/3, Concise 1.2/2, Error 1.7/2로 평가되었다.

적용한 방법의 효과로는 전문가 역할 부여와 명확한 정보 원칙 지시를 통해 환각 발생을 최소화했으며, 체계적인 프롬프트 설계를 통해 임상적 핵심 소견을 효과적으로 요약할 수 있었다.
이를 통해 결과물 전반의 임상적 신뢰성과 정확성을 확보할 수 있었다. 그러나 요약문의 간결성과 요점 압축 능력은 여전히 개선이 필요하며, 낮은 Summary 점수를 감안할 때 더 효과적인 핵심 내용 추출 전략이 요구된다. 또한 후속 권고 및 감별 진단과 관련된 지침을 강화하고, 데이터 노이즈 및 다양한 표현 패턴에 대한 대응력을 높이는 것이 필요하다.

후속 검사 권고와 임상 워크플로우 자동화를 환자 관리 시스템과 연계하는 방안도 기대된다. 향후에는 추가 데이터 확보와 실제 현장 피드백을 반영하여 모델을 고도화하는 노력이 필요하다.

4. 실험 결과 & 분석

- 리더보드 성적, 자체 지표 요약
- 시도한 방법의 효과와 한계
- 임상적 활용 가능성 및 개선/응용 아이디어

Task C :

Task C의 ICD 코드 예측 과제에서는 F1 스코어 약 0.20이라는 매우 저조한 결과를 기록하여 주/부 진단 코드 예측의 정확도와 재현율 모두에서 근본적인 한계를 보여주었습니다. 이는 모델에 의료 코딩 전문가 역할을 부여하고 단계별 분석을 지시했음에도 불구하고, 현재 접근법이 복잡한 의료 텍스트를 ICD 코드로 정확히 매핑하는 데 충분치 않았음을 시사합니다.

의료 용어의 다양성, 임상적 판단의 어려움, ICD 코드 체계의 복잡성에 대한 모델의 심층적 이해 부족과 함께, 코드 누락 및 환각 현상 제어에 실패한 것이 주요 원인으로 분석됩니다. 현재의 F1 스코어로는 즉각적인 임상 활용이 어렵지만, 이를 통해 ICD 코드 예측 자동화의 난이도와 개선점을 명확히 파악할 수 있었습니다.

5. 한계 & 향후 개선

- 현재 접근의 한계
- 본선 진출 시 보완 아이디어

전문가 역할 부여와 명확한 정보 원칙 지시를 통해 환각 발생을 최소화했으며, 체계적인 프롬프트 설계를 통해 임상적 핵심 소견을 효과적으로 요약할 수 있었다. 이를 통해 결과물 전반의 임상적 신뢰성과 정확성을 확보할 수 있었다. 그러나 요약문의 간결성과 요점 압축 능력은 여전히 개선이 필요하며, 낮은 Summary 점수를 감안할 때 더 효과적인 핵심 내용 추출 전략이 요구된다. 또한 후속 권고 및 감별 진단과 관련된 지침을 강화하고, 데이터 노이즈 및 다양한 표현 패턴에 대한 대응력을 높이는 것이 필요하다.

임상적 활용 가능성 측면에서는 다양한 의료기관과 영상검사 유형으로 확장 적용할 수 있는 가능성이 있으며, 후속 검사 권고와 임상 워크플로우 자동화를 환자 관리 시스템과 연계하는 방안도 기대된다. 향후에는 추가 데이터 확보와 실제 현장 피드백을 반영하여 모델을 고도화하는 노력이 필요하다

마지막으로 잦은 API오류로 인하여 다방면에 접근이 어려워 이로 인해 BERTScore에 의존하여 코드를 작성하였고 , 프롬프트 명령문에 대해서는 수정 작업을 진행하였으나 전처리 및 후처리 과정에 대해서는 접근이 어려웠다.

만약 다음 공모전이 열린다면 이런 보안점을 고려하여 주길 아쉬운 마음을 전달드립니다.