

# 오픈 소스 대규모 언어 모델을 활용한 멀티 모델 앙상블 메타데이터 생성 시스템

박성민<sup>0</sup>, 이동욱, 천승태

(주)데이터스트림즈

{sminpark, dwlee, stchun}@datastreams.co.kr

## Metadata Generation System based on Ensemble of Multi Open-Source Large Language Models

Seong-Min Park<sup>0</sup>, Dong-Wook Lee, Seung-Tae Chun

DataStreams Corp.

### 요약

본 논문은 디지털 시대에 문서 관리 효율성을 높이기 위해 오픈 소스 대규모 언어 모델(LLM)을 활용한 메타데이터 추출 방법론을 제안한다. LLM을 병렬로 사용하는 다중 모델 접근법과 3단계 추출 프로세스를 통해 비정형 데이터로부터 사용자 정의 메타데이터를 추출하고 이를 JSON 형식으로 출력한다. 본 연구는 기업 내부의 민감한 데이터 처리에 적합한 안전하고 효율적인 메타데이터 자동화 시스템을 제시하고 다양한 문서 형식에 대해 메타데이터 템플릿이 적용 가능한 범용적이고 확장성 있는 솔루션을 제공한다. 제안된 시스템은 G-eval 평가 결과 기존 방법 대비 71% 향상된 성능을 나타내었다

### 1. 서론

문서 메타데이터는 정보 관리와 검색에 핵심적이나, 현재 수동 생성이 주를 이루어 자동화가 필요한 상황이다.[1] 기업의 민감 데이터 처리 시 보안 문제로 공개 API 기반 LLM 사용이 제한되어, 내부 구축 가능한 오픈소스 모델이 대안으로 부상하고 있다.

본 연구는 오픈소스 LLM을 활용해 비정형 데이터에서 메타데이터를 추출하고 JSON으로 출력하는 시스템을 개발했다. 단일 모델의 한계를 극복하기 위해 다중 모델 접근법[2]을 채택했으며, 이는 다양한 문서와 메타데이터 템플릿에 적용 가능한 범용 솔루션[3]이다.

### 2. 본론

#### 2.1 시스템 구조

메타데이터 자동 생성의 신뢰성과 정확성을 향상시키기 위해 다수의 독립적인 언어 모델 서버를 활용하는 병렬 처리 시스템을 제안한다. 원본 문서와 메타데이터 템플릿이 입력되면 요청이 병렬적으로 각 모델 서버의 독립적인 REST API Endpoint를 통해 전달된다.

##### 2.1.1 입력 데이터

표 1 입력 구성 요소

| 입력 구성 요소  | 설명                               |
|-----------|----------------------------------|
| 메타데이터 템플릿 | 다양한 유형의 메타데이터를 표준화한 JSON 형식의 템플릿 |
| 원본 문서     | 메타데이터 추출 대상이 되는 문서               |

그림1은 전체 시스템의 구조를 Lead model server 1대, Assistant model server를 2대로 구성하여 전체 구조를 나타내는 그림이다.

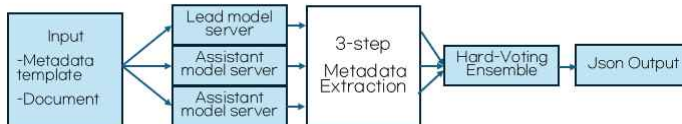


그림 1 시스템 구조

그림 2는 메타데이터 템플릿의 예시이다. 메타데이터 템플릿은 taxonomy와 메타데이터 키-값 두 가지 핵심 요소로 구성된다.

메타데이터 템플릿

```
{
  "taxonomy": {
    "type": "string",
    "value": ["main_category-sub_category"]
  },
  "main_category": {
    "type": "string",
    "value": ["프로젝트준비", "프로젝트산출물", "컨설팅자료", ...]
  },
  "sub_category": {
    "type": "dict",
    "value": {
      "프로젝트준비": ["제안요청서", "POC/BMT", ...],
      "프로젝트산출물": ["수행요청서", "수행계획서", ...],
      "컨설팅자료": ["컨설팅보고서", "기타"]
    }
  }
}
```

그림 2 메타데이터템플릿 예시

taxonomy는 메타데이터의 계층 구조를 정의하며, 예를 들어 '대분류-소분류'와 같이 상하위 관계를 하이픈으로 표현한다. 메타데이터 키-값에서 키는 추출하고자 하는 메타데이터 항목을, 값은 해당 메타데이터의 후보군을 나타낸다.

## 2.2 처리 단계

### 2.2.1 메타데이터 추출 1단계: Question Answering

1단계에서는 CoT[4] 방식으로 생성한 프롬프트로 주어진 메타데이터 템플릿으로부터 문서에 해당하는 메타데이터 값을 결정하도록 구성되었다. 결과가 JSON 변환가능한 경우 그 값을 API 응답으로 반환한다. 만일 반환받은 값이 메타데이터 템플릿에 존재하지 않는 값이거나 JSON 변환 가능하지 않다면 메타데이터 추출 2단계로 진행한다. 프롬프트는 그림3 과 같다.

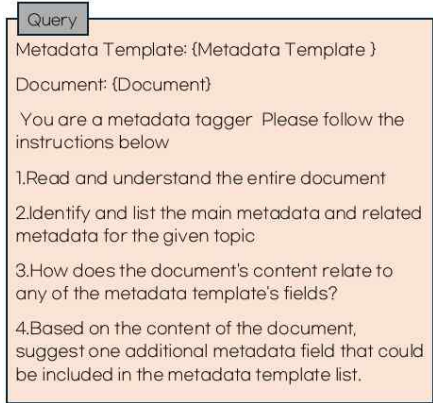


그림 3 Question Answering Prompt

### 2.2.2 메타데이터 추출 2단계

#### : Question Answering with Explanation

입력 데이터에 대한 요약 및 생성된 설명을 바탕으로 메타데이터 추출을 시도한다. 먼저 문서에 대해 상세한 설명을 생성하는 프롬프트로 문서에 대한 설명을 획득한다. 설명을 생성하는 프롬프트는 아래 그림과 같다.

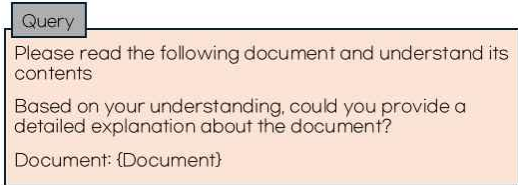


그림 4 Explanation Generate Prompt

생성되는 설명예시는 그림5와 같다.

Project Overview

Background: The project is necessary due to the need to transition from an existing HKMC data-based platform to a new one that is more efficient and scalable.

Scope and Content: The project involves building a big data monitoring system that can manage large amounts of data, provide real-time insights, and enable effective decision-making.

Requirements: The system must be able to handle high volumes of data, provide flexible workflow management, and include features such as version control, permission management, and alert systems.

그림 5 Example of Explanation

생성된 설명을 다음 프롬프트에 문서와 메타데이터 템플릿과 함

께 주입하고 메타데이터를 추출한다. LLM의 메타데이터 추출 결과가 JSON 변환 가능한 형태라면 API 응답으로 반환한다. JSON 변환 가능하지 않거나 그 값이 메타데이터 템플릿에 없다면 응답을 다음 단계로 넘겨서 계속 진행된다.

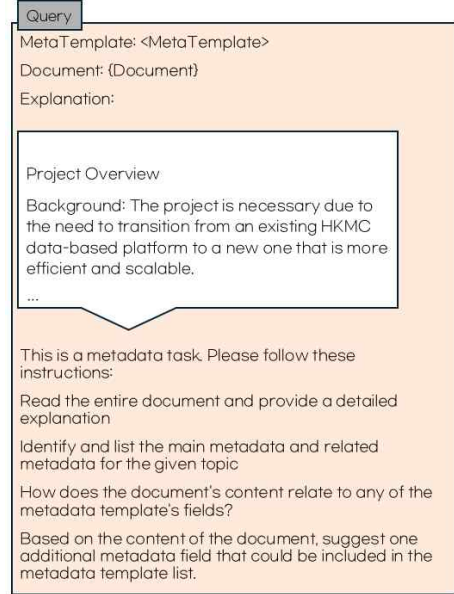


그림 6 Question Answering with Explanation Prompt

### 2.2.3 메타데이터 추출 3단계: Chain of Verification

Chain of Verification[5] 은 1,2단계에서 추출 실패 시 수행하는 단계이며 문서, 메타데이터, 메타데이터 템플릿을 종합적으로 분석 단계적 검증 과정을 통해 메타데이터를 추출한다. 이를 수행하는 프롬프트는 그림7과 같다.

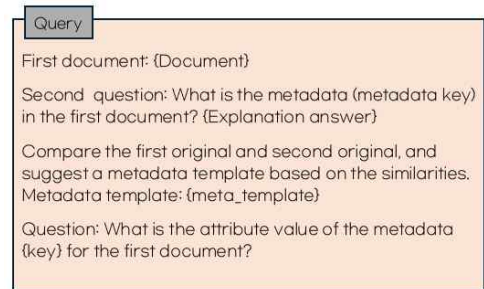


그림 7 Chain of Verification Prompt

각 메타데이터 키를 순회하며 이전 Question Answering with Prompt 단계에서 출력된 응답을 LLM에 함께 주입한다. 모든 분류체계 명에 대하여 위와 같은 CoV(Chain of verification)를 진행하여 얻은 응답을 취합하여 JSON 형태로 출력하도록 한다.

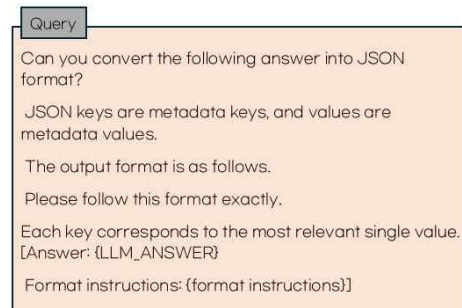


그림 8 Json Output Prompt

### 2.2.4 출력 단계

각 단계에서 적절한 메타데이터 값이 나와 최종 답변을 할 때에는 다음과 같은 JSON 출력 형식을 유도하여 시스템에서 활용할 수 있도록 한다. 각 모델 서버에서 위의 프로세스가 각각 진행되어 나온 Json 형태의 응답을 취합한다. 취합한 응답 값에서 분류체계명 별로 다수의 예측 분류명을 각 분류 체계명의 최종 분류명으로 지정하는 Hard-voting Ensemble[6]을 통해 최종 응답 값을 결정하고 반환한다.

### 2.3 프롬프트 엔지니어링 전략

Chain of Thought 접근법을 기반으로 한 프롬프트 엔지니어링 전략을 적용하여 메타데이터 추출의 정확성을 향상시킨다. 주요 전략은 다음과 같다.

표 2 프롬프트 엔지니어링 전략

| 전략 요소      | 설명                             |
|------------|--------------------------------|
| 템플릿 순서 무관성 | 템플릿 항목 순서에 영향받지 않도록 설계         |
| 가중치 기반 선택  | 관련성에 따른 가중치 부여 방식 채택           |
| 언어 스타일     | 명령형보다 청유형 질의 방식 채택             |
| 페르소나 부여[7] | 일관된 응답을 위한 가상 인격 설정            |
| 모델 파라미터    | Temperature값 0.3으로 고정 (일관성 확보) |

### 2.4 앙상블 기법

최종적으로, 세 개의 모델 서버에서 생성된 메타데이터 결과에 대해 Hard-Voting Ensemble 기법을 적용한다. 이 때 Lead model server의 응답 결과에 가중치를 Assistant model server의 응답 결과보다 크게 한다. 이를 통해 개별 모델의 편향을 상쇄하고 보다 신뢰성 있는 최종 메타데이터를 도출한다.

### 3. 평가

평가 프로세스는 LLM 기반 성능 평가 방법이다. 문서에 대해 5회씩 메타데이터 추출을 시도한다. 이렇게 얻은 결과는 Claude-3.5-Sonnet 모델을 이용해 평가하며 G-eval[8] 방식을 통해 생성된 프롬프트를 사용하였다. 각각의 추출 시도는 1점부터 5점까지의 점수로 평가되어, 한 문서당 최대 25점(5회 \* 5점)을 받을 수 있다. 최종 평가 점수는 문서 전체에 대한 평균 점수를 계산한 뒤, 25점 만점 대비 백분율로 환산하여 산출하였다.

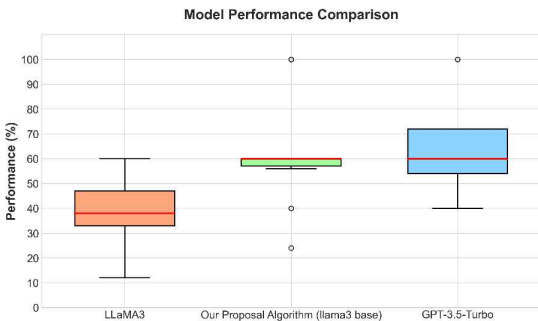


그림 9 메타데이터 추출 정확도

본 연구에서 제안된 방법론의 성능은 기존의 접근 방식과 비교 분석하였다. 결과적으로, 본 방법은 LLaMA 3 모델을 이용한 일반적인 프롬프트 기반의 메타데이터 추출 방식보다 우수한 성능을 나타낸다. 또한, GPT-3.5-turbo-1106 모델과 비교했을 때 유사한 수준의 성능을 보여주었다. 이러한 결과는 gpt-3.5-turbo와 같은 모델을 사용하기 위해 네트워크로 접근해야 하는 API방식

에 비해 제안된 방법론은 폐쇄된 환경에서도 오픈소스 LLM으로 활용이 가능하므로 그 효과성을 입증한다.

표 3 평가 결과

|        | llama3 | Proposed | gpt-3.5 |
|--------|--------|----------|---------|
| 정확도 평균 | 38.4%  | 62%      | 66%     |

### 4. 결과

본 연구에서 제안된 시스템은 실험 결과, 기존의 단일 모델 기반 접근법에 비해 높은 정확도와 일관성을 보여주었다. 특히, 다양한 문서 형식에서도 안정적인 성능을 나타냈다.

현재 시스템의 발전 방향으로는 다양한 오픈 소스 언어 모델을 추가 실험하여 최적의 모델 조합을 찾고, 문서 특성에 따라 모델 가중치를 동적으로 조절하는 정교한 앙상블 기법을 구현하며, 텍스트 외 시각적 요소도 분석하는 멀티모달 시스템으로 확장하고, 사용자 피드백과 새 데이터를 바탕으로 실시간 학습 및 적응하는 시스템을 개발하여 지속적인 성능 향상을 도모하는 것이 필요하다.

### 참고 문헌

- [1] 박성민, 최규홍, 이동욱, 천승태, “협업 필터링 및 세션 키워드 추천 알고리즘 기반 데이터 활용도 향상 시스템에 관한 연구”, *한국통신학회 하계학술대회*, 907~908, 2023.06
- [2] Dietterich, T. G., "Ensemble Methods in Machine Learning", In *Multiple Classifier Systems*, LNCS, vol. 1857, pp. 1-15, 2000.
- [3] Flynn, P., Zhou, L., Maly, K., Zeil, S., & Zubair, M., "Automated template-based metadata extraction architecture.", In *Asian digital libraries: Looking back 10 years and forging new frontiers*, pp. 327-336, 2007.
- [4] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., & Le, Q. V., "Chain-of-thought prompting elicits reasoning in large language models.", *36th Conference on Neural Information Processing Systems (NeurIPS 2022)*
- [5] Dhuliawala, S., Komeili, M., Xu, J., Raileanu, R., L i, X., Celikyilmaz, A., & Weston, J., "Chain-of-verification reduces hallucination in large language models.", <https://arxiv.org/pdf/2309.11495>, 2023.
- [6] Awe, O. O., Opatye, G., Johnson, C. A. G., & Tay o, O. T., "Weighted Hard and Soft Voting Ensemble Machine Learning Classifiers: Application to Anaemia Diagnosis", *Health Sustainable Statistical and Data Science Methods and Practices*, 351-374, 2023.
- [7] Tseng, Y.-M., Huang, Y.-C., Hsiao, T.-Y., Chen, W.-L., Huang, C.-W., Meng, Y., & Chen, Y.-N. "Two tales of persona in LLMs: A survey of role-playing and personalization.", <https://doi.org/10.48550/arXiv.2406.01171>, 2024
- [8] Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., & Zhu, C. "G-eval: NLG evaluation using GPT-4 with better human alignment." <https://doi.org/10.48550/arXiv.2303.16634>, 2023.