개체명 기반 컨텍스트 추출을 통한 의료 분야 검색 증강 질의응답 시스템

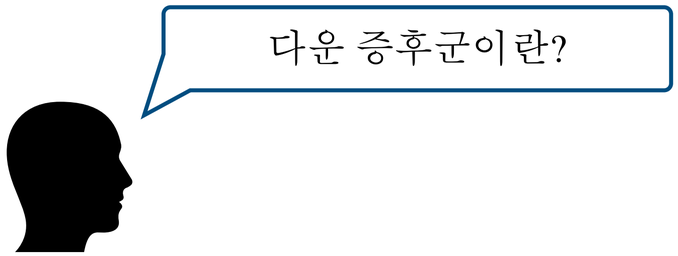
조정훈 이근배

POSTECH

June 28, 2024 | KCC 2024

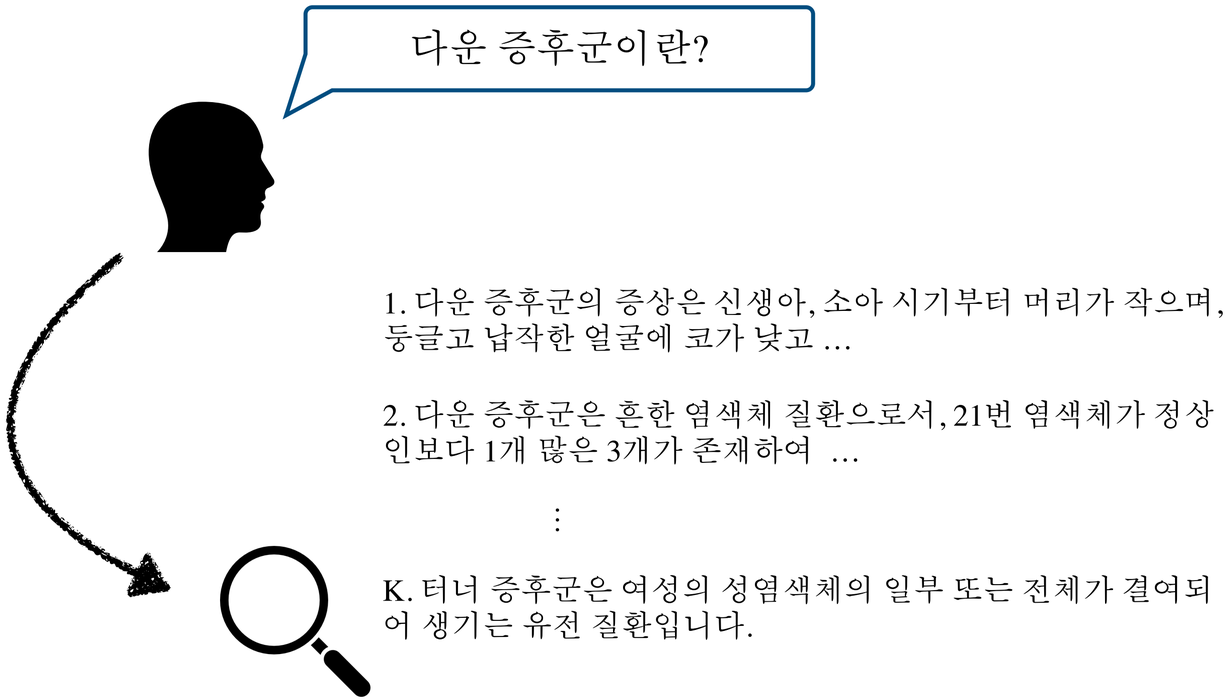
Retrieval-Augmented Generation (RAG)

Retrieval-Augmented Generation (RAG)



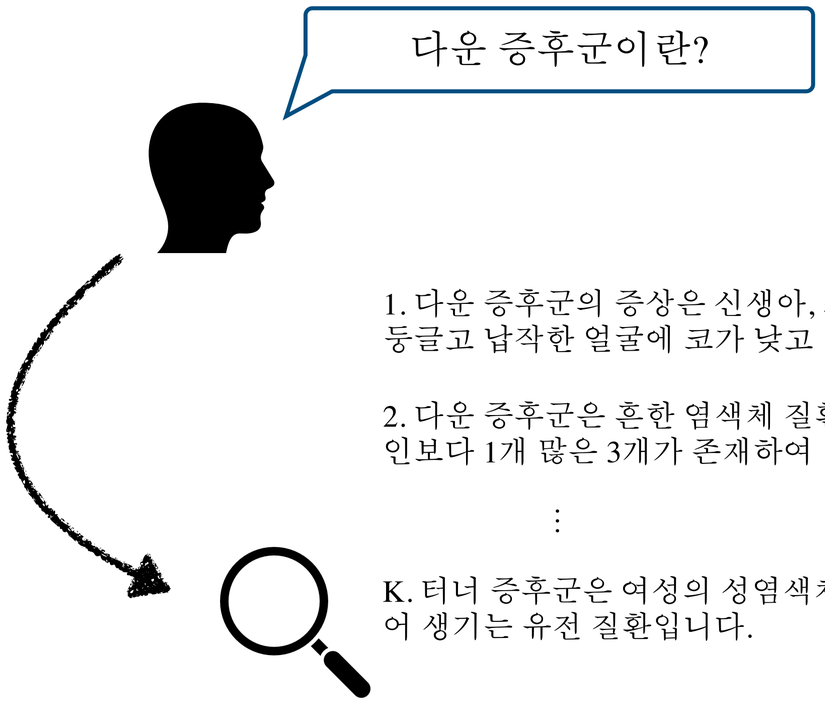
Retrieval-Augmented Generation (RAG)

Step 1: Retrieve K documents



Retrieval-Augmented Generation (RAG)

Step 1: Retrieve K documents

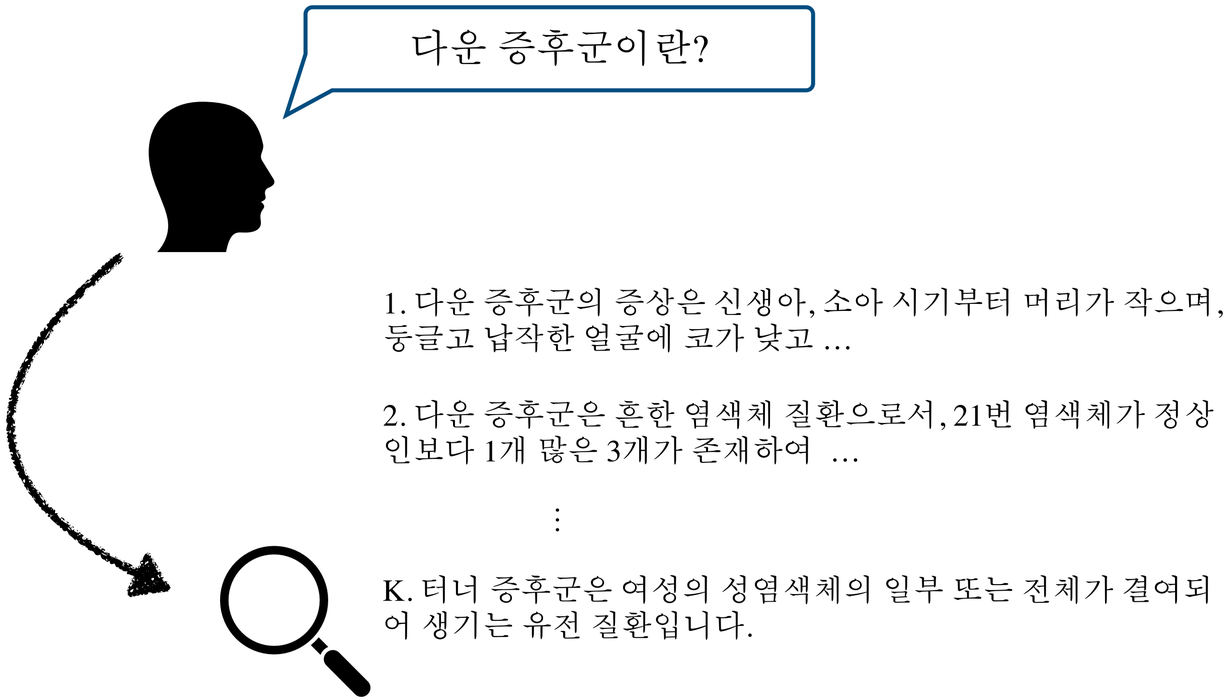


Step 2: Prompt augmentation with K docs

문서를 참고하여 질문에 대한 답을 하시오.
{1부터 K개의 문서}
질문: 다운 증후군이란?

Retrieval-Augmented Generation (RAG)

Step 1: Retrieve K documents

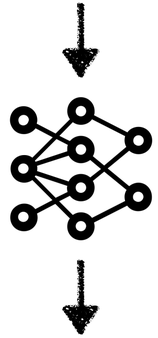


Step 2: Prompt augmentation with K docs

문서를 참고하여 질문에 대한 답을 하시오.
{1부터 K개의 문서}
질문: 다운 증후군이란?

Step 3: Generate

augmented prompt

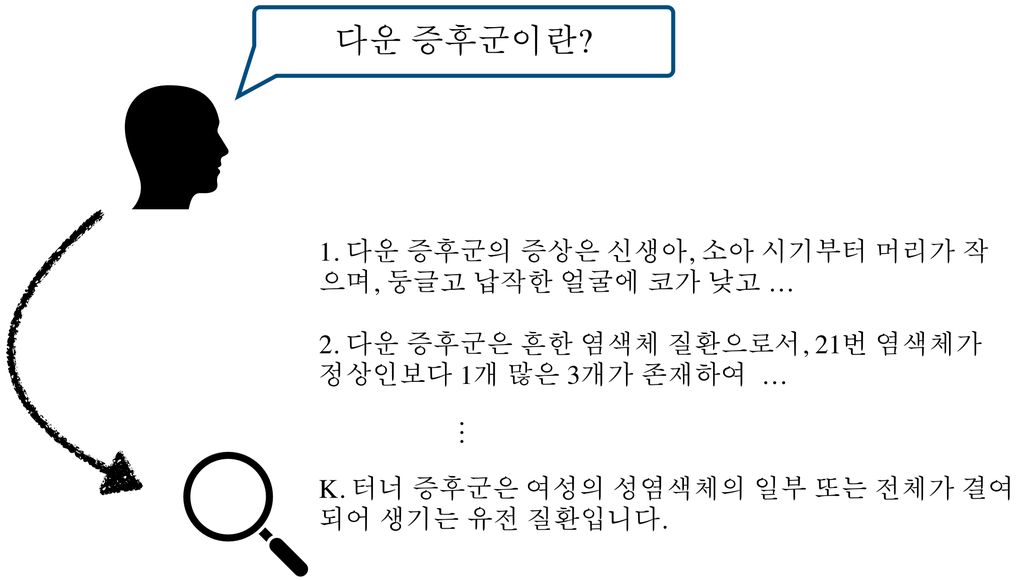


다운 증후군은 가장 흔한 염색체 질환으로서, 여성의 염색체의 일부 또는 전체가 결여되어 정신 지체, 신체 기형, 전신 기능 이상, 성장 장애 등을 일으키는 유전 질환이다.

RAG is not always efficient!

RAG is not always efficient!

Step 1: Retrieve K documents



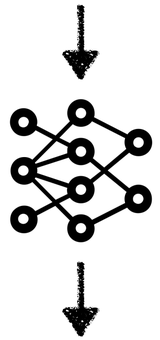
Retriever
(e.g., Google, BM25)

Step 2: Prompt augmentation with K docs

문서를 참고하여 질문에 대한 답을 하시오.
{1부터 K개의 문서}
질문: 다운 증후군이란?

Step 3: Generate

augmented prompt



다운 증후군은 가장 흔한 염색체 질환으로서, 여성의 염색체의 일부 또는 전체가 결여되어 정신 지체, 신체 기형, 전신 기능 이상, 성장 장애 등을 일으키는 유전 질환이다.

RAG is not always efficient!

Step 1: Retrieve K documents

다운 증후군이란?

1. 다운 증후군의 증상은 신생아, 소아 시기부터 머리가 작으며, 둥글고 납작한 얼굴에 코가 낮고 …

2. 다운 증후군은 흔한 염색체 질환으로서, 21번 염색체가 정상인보다 1개 많은 3개가 존재하여 …

⋮

K. 터너 증후군은 여성의 성염색체의 일부 또는 전체가 결여되어 생기는 유전 질환입니다.

Retriever
(e.g., Google, BM25)

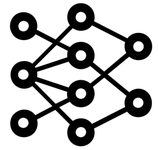
Retriever가 항상 관련 문서만을 내뱉는다는 보장이 없음

Step 2: Prompt augmentation with K docs

문서를 참고하여 질문에 대한 답을 하시오.
{1부터 K개의 문서}
질문: 다운 증후군이란?

Step 3: Generate

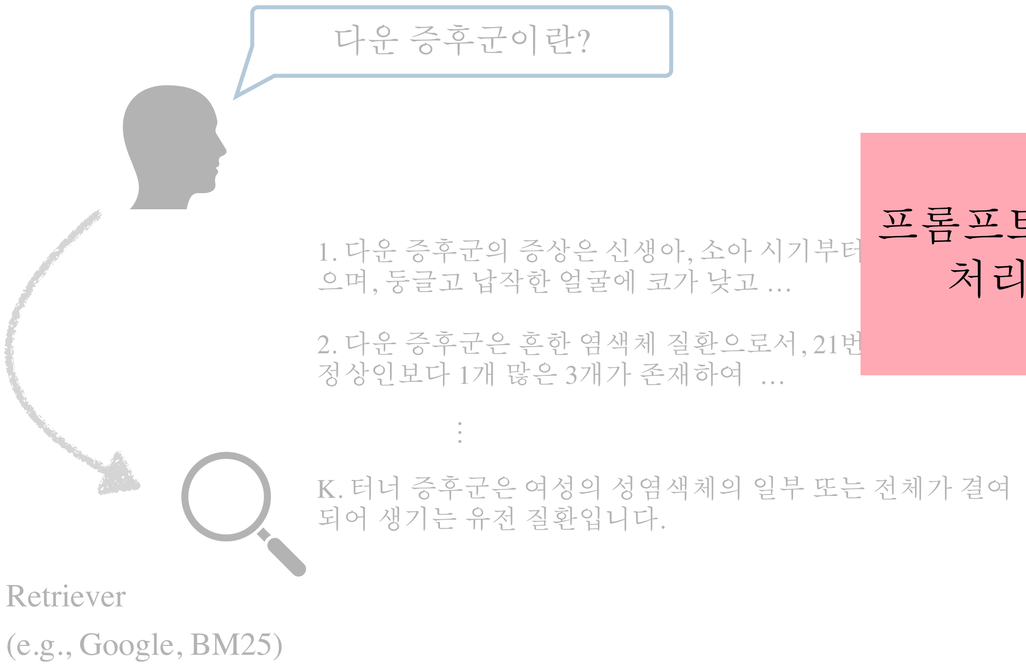
augmented prompt



다운 증후군은 가장 흔한 염색체 질환으로서, 여성의 염색체의 일부 또는 전체가 결여되어 정신 지체, 신체 기형, 전신 기능 이상, 성장 장애 등을 일으키는 유전 질환이다.

RAG is not always efficient!

Step 1: Retrieve K documents

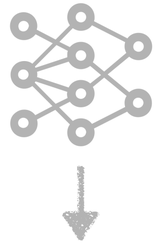


Step 2: Prompt augmentation with K docs

문서를 참고하여 질문에 대한 답을 하시오.
{1부터 K개의 문서}
질문: 다운 증후군이란?

프롬프트의 길이가 증가하므로, **LLM**이 처리해야 하는 토큰 수가 많아짐

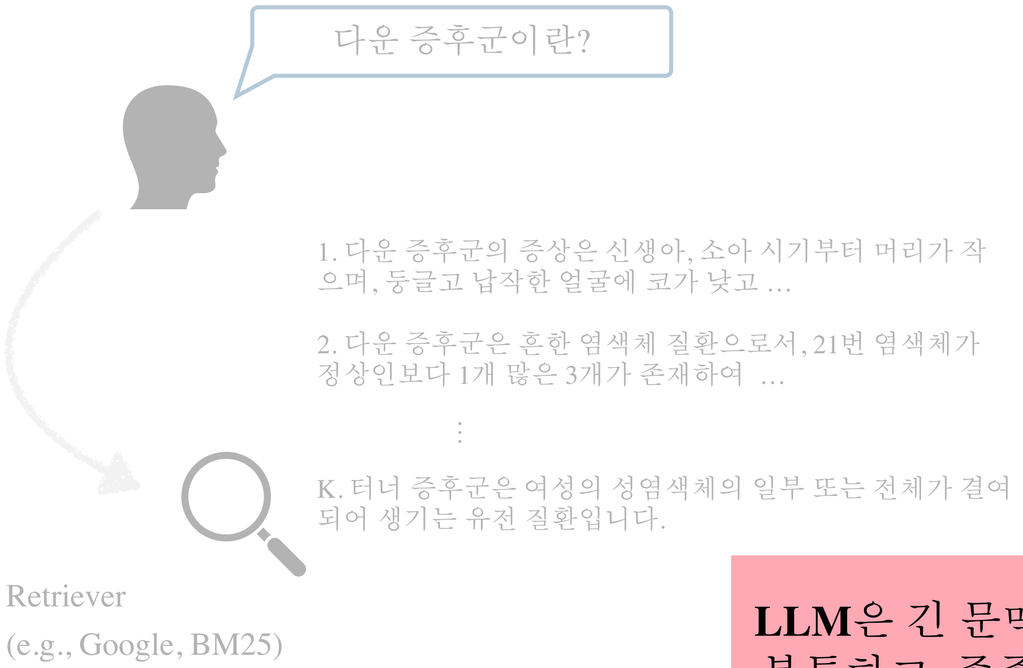
augmented prompt



다운 증후군은 가장 흔한 염색체 질환으로서, 여성의 염색체의 일부 또는 전체가 결여되어 정신 지체, 신체 기형, 전신 기능 이상, 성장 장애 등을 일으키는 유전 질환이다.

RAG is not always efficient!

Step 1: Retrieve K documents



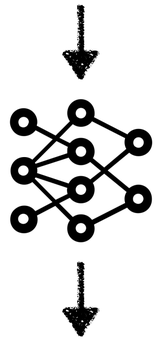
Step 2: Prompt augmentation with K docs

문서를 참고하여 질문에 대한 답을 하시오.
{1부터 K개의 문서}
질문: 다운 증후군이란?

LLM은 긴 문맥을 이해하기 위해 고군분투하고, 종종 주요한 단어나 문장의 정보가 누락됨

Step 3: Generate

augmented prompt



다운 증후군은 가장 흔한 염색체 질환으로서, 여성의 염색체의 일부 또는 전체가 결여되어 정신 지체, 신체 기형, 전신 기능 이상, 성장 장애 등을 일으키는 유전 질환이다.

RAG is not always efficient!

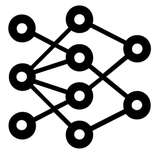
Retriever가 항상 관련 문서만을 내뱉는다는 보장이 없음

프롬프트의 길이가 증가하므로, **LLM**이 처리해야 하는 토큰 수가 많아짐

LLM은 긴 문맥을 이해하기 위해 고군분투하고, 종종 주요한 단어나 문장의 정보가 누락됨

Document
{document_1}
{document_2}
…
{document_k}

Question
{question}



answer

RAG is not always efficient!

Retriever가 항상 관련 문서만을 내뱉는다는 보장이 없음

프롬프트의 길이가 증가하므로, **LLM**이 처리해야 하는 토큰 수가 많아짐

LLM은 긴 문맥을 이해하기 위해 고군분투하고, 종종 주요한 단어나 문장의 정보가 누락됨

Document
{document_1}
{document_2}
…
{document_k}

Question
{question}

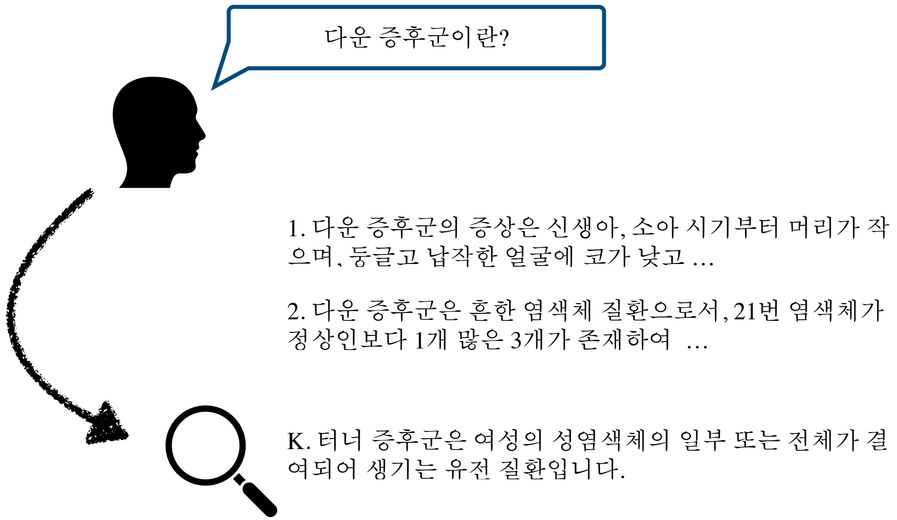
Document
{summary}

Question
{question}

answer

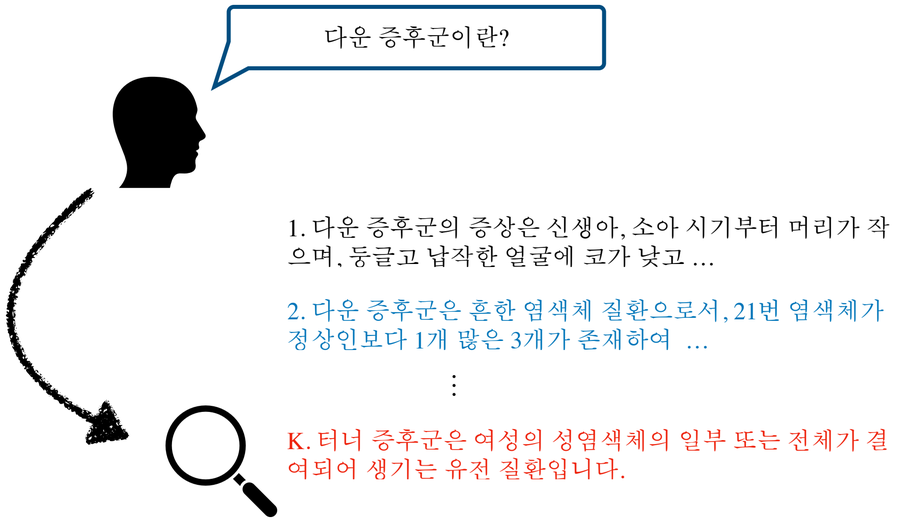
Overview

Step 1: Retrieve K documents

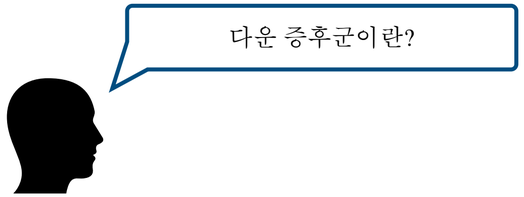


Overview

Step 1: Retrieve K documents



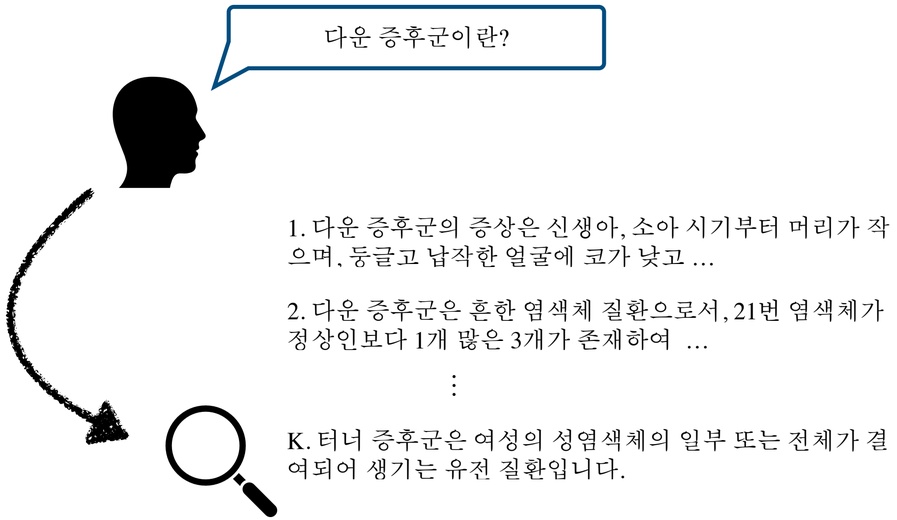
Step 2: Summarization



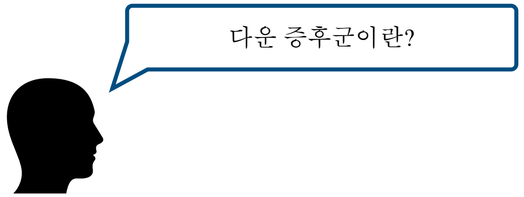
다운 증후군은 21번 염색체가 하나 더 많아 발생하며, 정신 지체와 다양한 신체 기형을 동반하는 유전 질환이다. 특징적인 얼굴 모습과 저지능, 발달 지연, 선천성 심장 기형 등 여러 증상이 나타나며, 출생 후에도 다양한 장기 기능 이상이 발생한다. 정기적인 의료 및 사회적 지원이 필요하며, 증상만으로는 진단이 어려워 의료진과의 상담 및 정밀 검사가 중요하다.

Overview

Step 1: Retrieve K documents



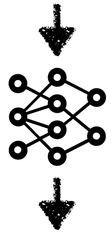
Step 2: Summarization



다운 증후군은 21번 염색체가 하나 더 많아 발생하며, 정신 지체와 다양한 신체 기형을 동반하는 유전 질환이다. 특징적인 얼굴 모습과 저지능, 발달 지연, 선천성 심장 기형 등 여러 증상이 나타나며, 출생 후에도 다양한 장기 기능 이상이 발생한다. 정기적인 의료 및 사회적 지원이 필요하며, 증상만으로는 진단이 어려워 의료진과의 상담 및 정밀 검사가 중요하다.

Step 3: Prompt LM with summarized context and generate

문서를 참고하여 질문에 대한 답을 하시오.
{요약된 문서}
질문: 다운 증후군이란?



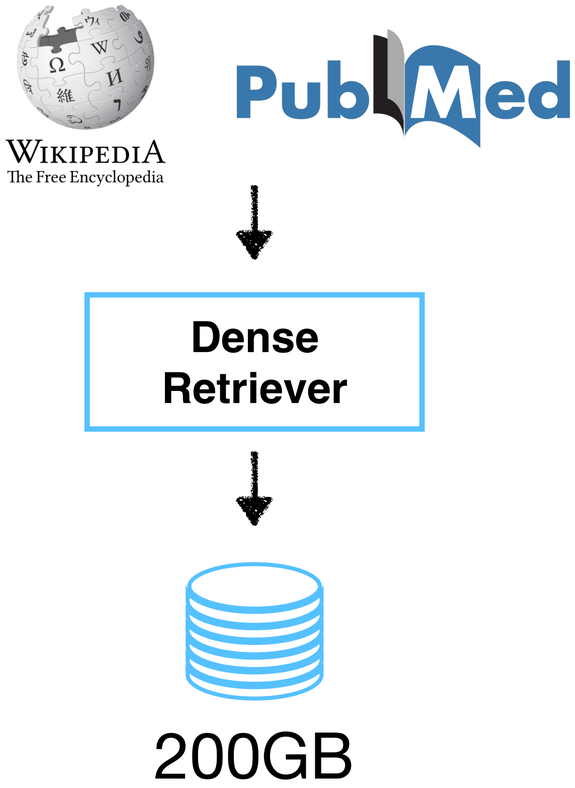
다운 증후군은 21번 염색체가 정상보다 하나 더 많은 3개가 존재하여 발생하는 유전 질환으로, 정신 지체, 신체 기형, 성장 장애 등을 일으킨다. 특징적인 얼굴 모습과 낮은 지능을 가지며, 다양한 신체적 이상이 동반된다. 출생 전후로 다양한 장기 기능 이상이 나타나며, 평생에 걸쳐 지속적인 의료 및 사회적 지원이 필요하다.

Methodology - data

- **Vector DB** 구축
- **Extract summary**
- 각 토큰 별 **B-I-O label** 할당

Methodology - data

- **Vector DB** 구축

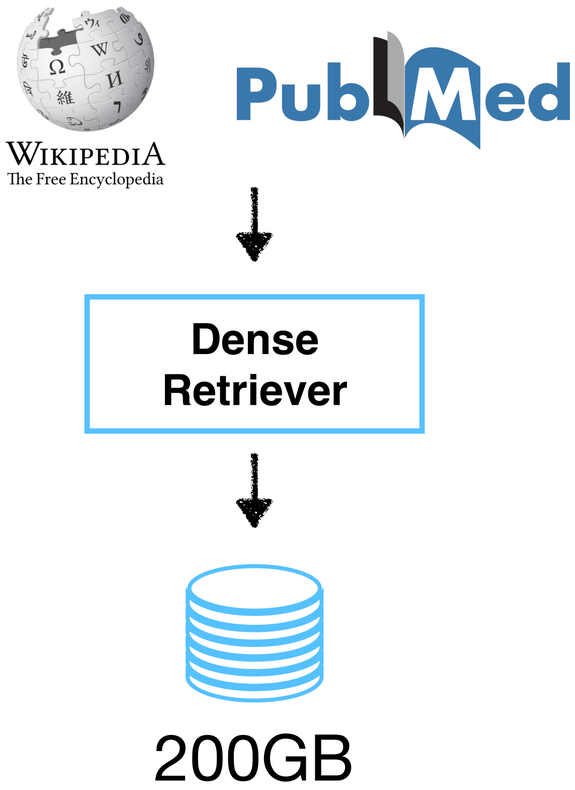


- **Extract summary**

- 각 토큰 별 **B-I-O label** 할당

Methodology - data

- **Vector DB** 구축



- **Extract summary**

x

1. 다운 증후군의 증상은 신생아, 소아 시기부터 머리가 작으며, 둥글고 납작한 얼굴에 코가 낮고 …

2. 다운 증후군은 흔한 염색체 질환으로서, 21번 염색체가 정상인보다 1개 많은 3개가 존재하여 …

5. 터너 증후군은 여성의 성염색체의 일부 또는 전체가 결여되어 생기는 유전 질환입니다.

gpt-3.5

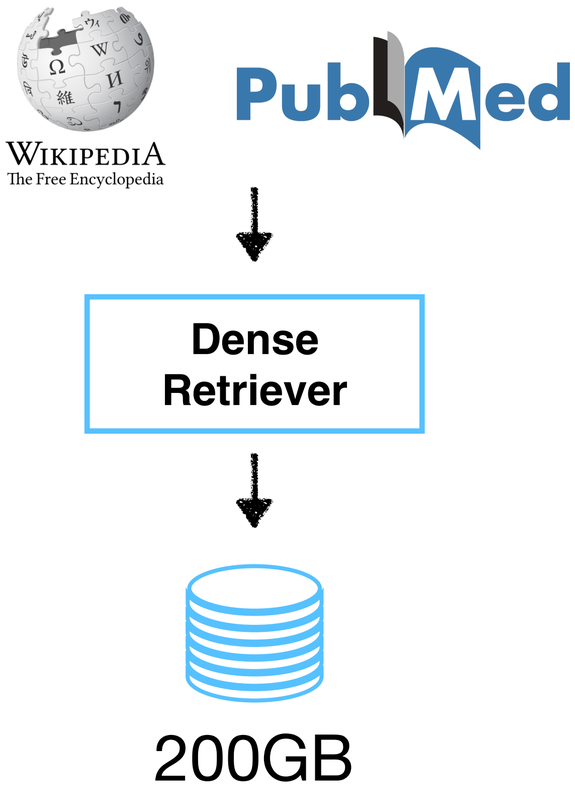
y

다운 증후군은 21번 염색체가 하나 더 많아 발생하며, 정신 지체와 다양한 신체 기형을 동반하는 유전 질환이다. 특징적인 얼굴 모습과 저지능, 발달 지연, 선천성 심장 기형 등 여러 증상이 나타나며, 출생 후에도 다양한 장기 기능 이상이 발생한다. 정기적인 의료 및 사회적 지원이 필요하며, 증상만으로는 진단이 어려워 의료진과의 상담 및 정밀 검사가 중요하다.

- 각 토큰 별 **B-I-O label** 할당

Methodology - data

- **Vector DB 구축**



- **Extract summary**

x

1. 다운 증후군의 증상은 신생아, 소아 시기부터 머리가 작으며, 둥글고 납작한 얼굴에 코가 낮고 …

2. 다운 증후군은 흔한 염색체 질환으로서, 21번 염색체가 정상인보다 1개 많은 3개가 존재하여 …

5. 터너 증후군은 여성의 성염색체의 일부 또는 전체가 결여되어 생기는 유전 질환입니다.

gpt-3.5

y

다운 증후군은 21번 염색체가 하나 더 많아 발생하며, 정신 지체와 다양한 신체 기형을 동반하는 유전 질환이다. 특징적인 얼굴 모습과 저지능, 발달 지연, 선천성 심장 기형 등 여러 증상이 나타나며, 출생 후에도 다양한 장기 기능 이상이 발생한다. 정기적인 의료 및 사회적 지원이 필요하며, 증상만으로는 진단이 어려워 의료진과의 상담 및 정밀 검사가 중요하다.

- 각 토큰 별 **B-I-O label** 할당

x

다운 증후군의 증상은 신생아, 소아 시기부터 머리가 작으며, 둥글고 납작한 얼굴에 코가 낮고 …

다운 증후군은 흔한 염색체 질환으로서, 21번 염색체가 정상인보다 1개 많은 3개가 존재하여 …

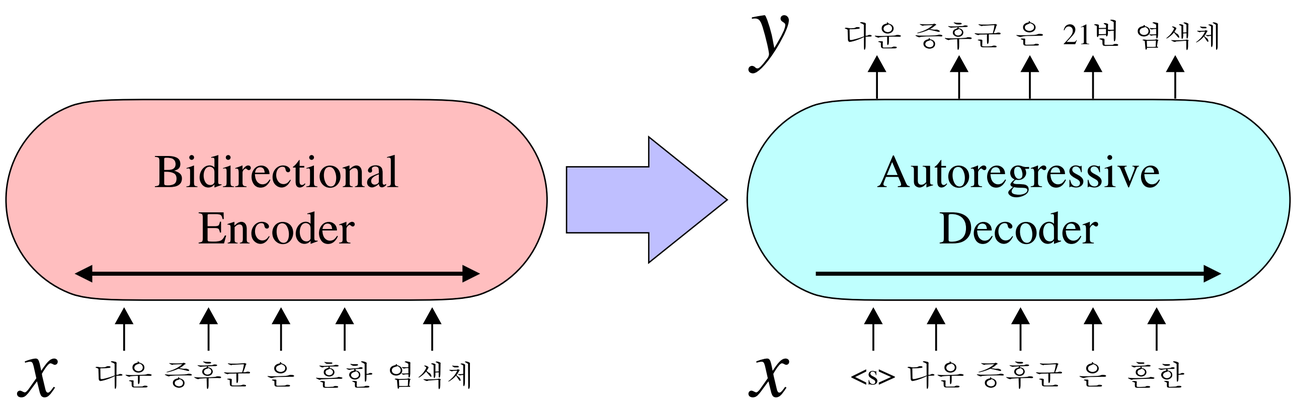
터너 증후군은 여성의 성염색체의 일부 또는 전체가 결여되어 생기는 유전 질환입니다.

z

B, I, I, I, O, O, O, O ,,, B, O, O,

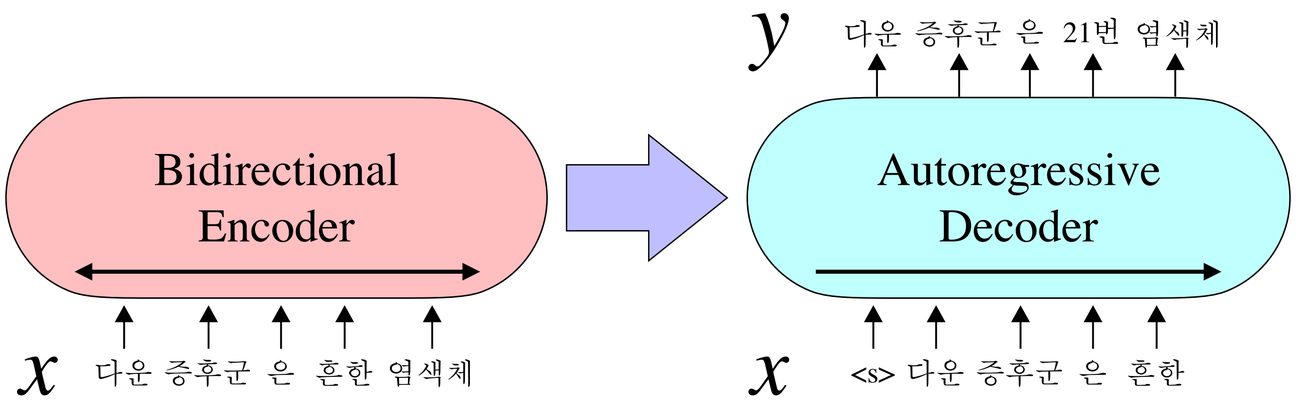
Methodology - training

- **Standard seq2seq**

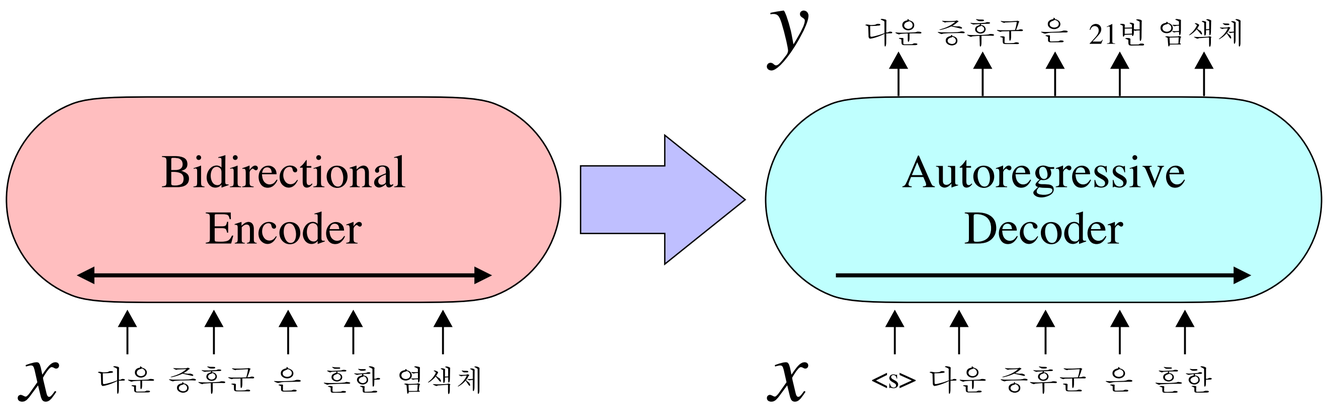


Methodology - training

- **Standard seq2seq**

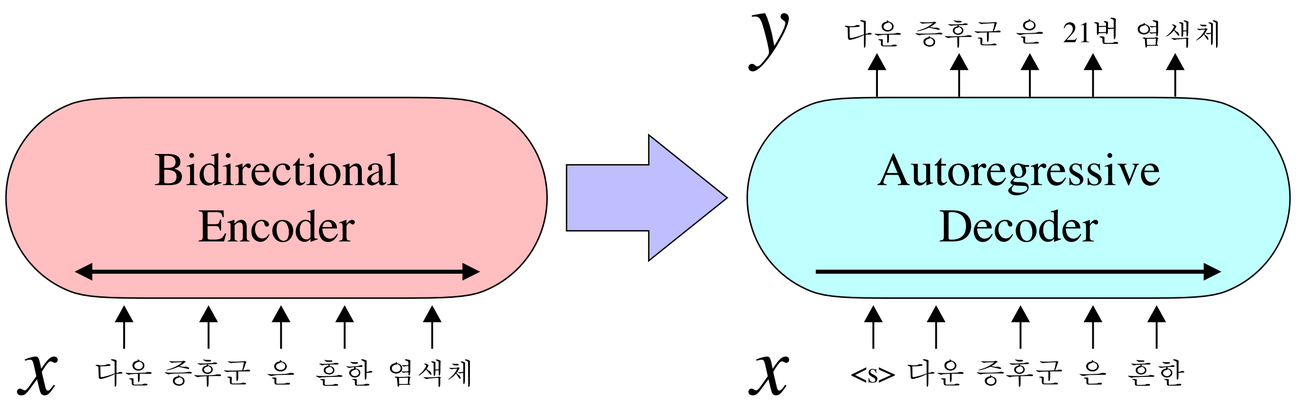


- **Ours**

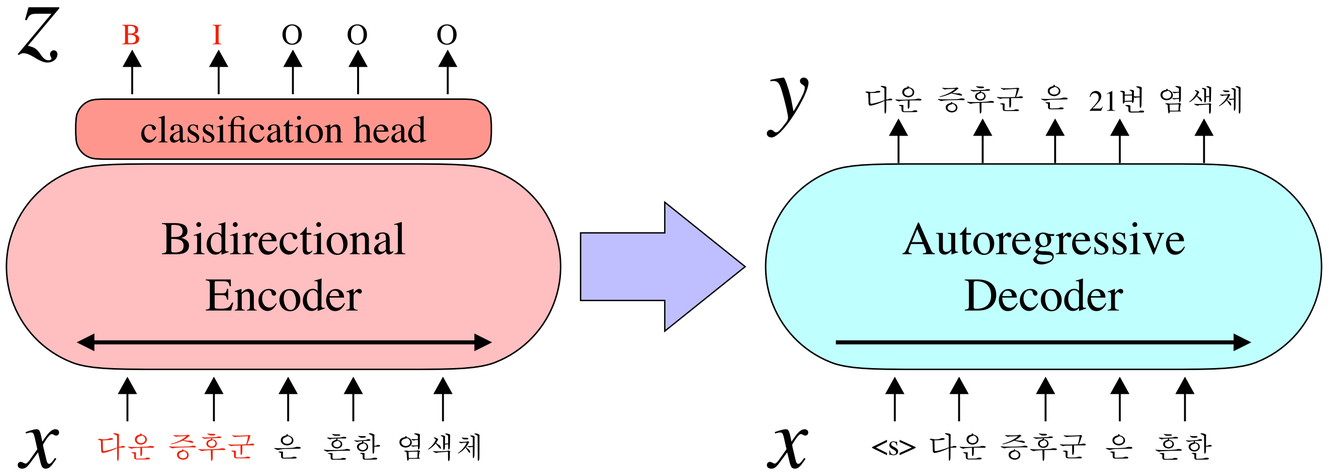


Methodology - training

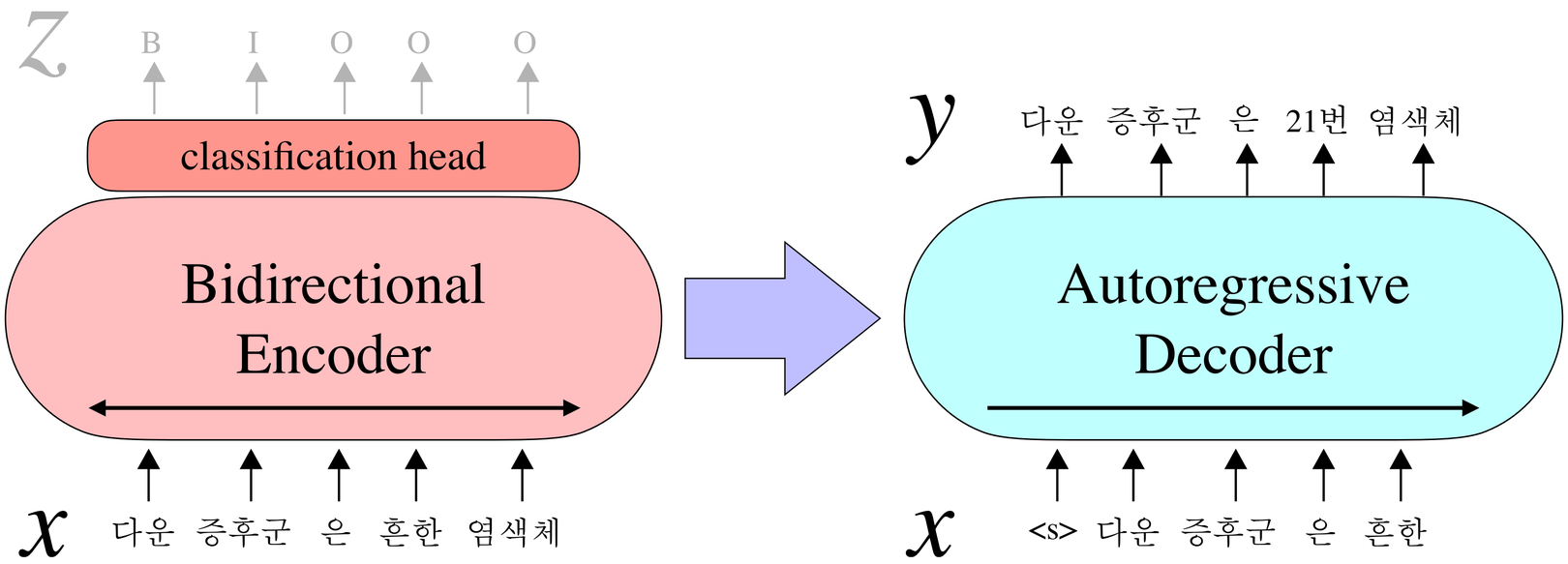
- **Standard seq2seq**



- **Ours**

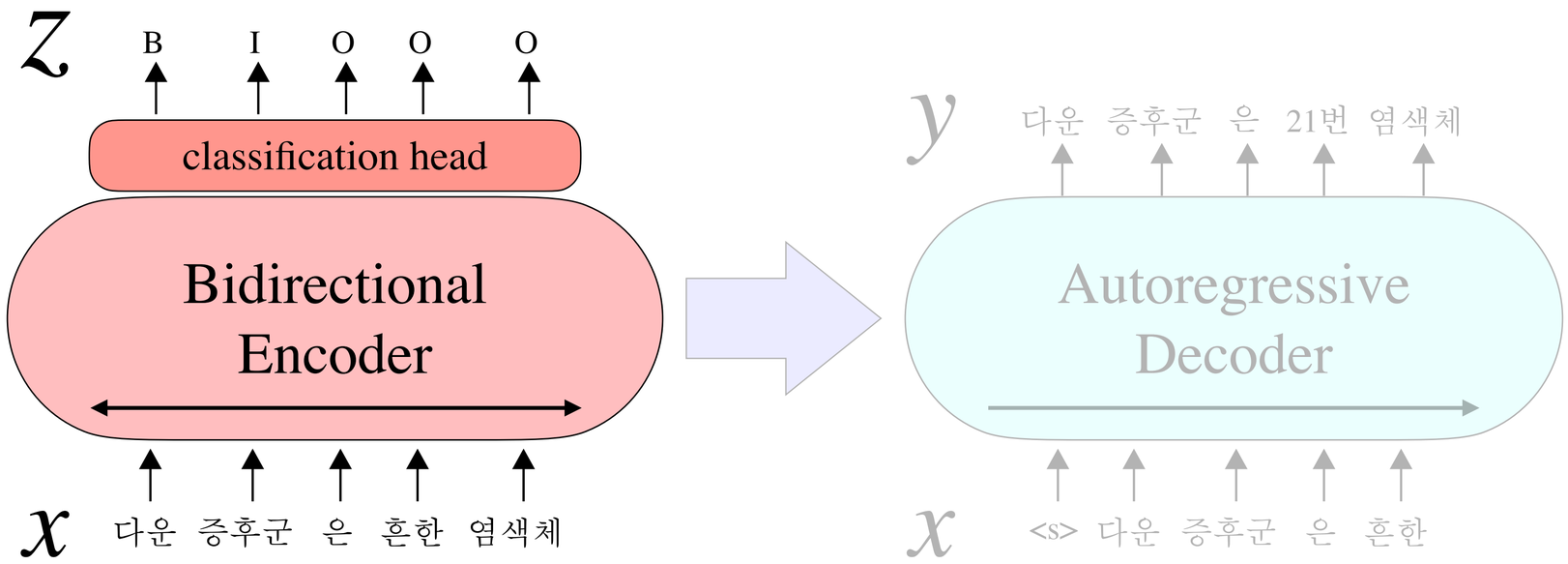


Methodology - training



$$L_i^{seq}(x_i, y_i, \theta) = -\sum_{t=1}^{m} logp_\theta(y_i^t \mid x_i, y_i^{<t})$$

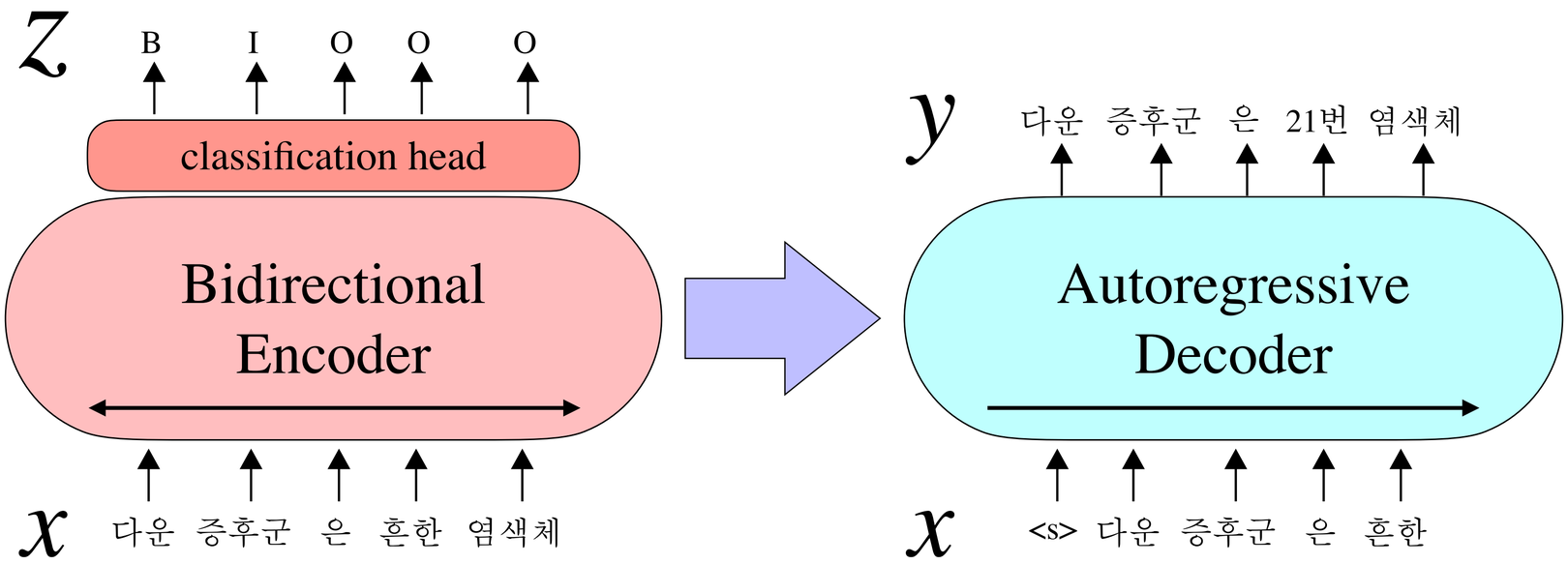
Methodology - training



$$L_i^{seq}(x_i, y_i, \theta) = -\sum_{t=1}^{m} logp_\theta(y_i^t \mid x_i, y_i^{<t})$$

$$L_i^{ner}(x_i, z_i, \theta_{enc}) = -\sum_{t=1}^{k} logp_{\theta_{enc}}(z_i^t \mid x_i)$$

Methodology - training

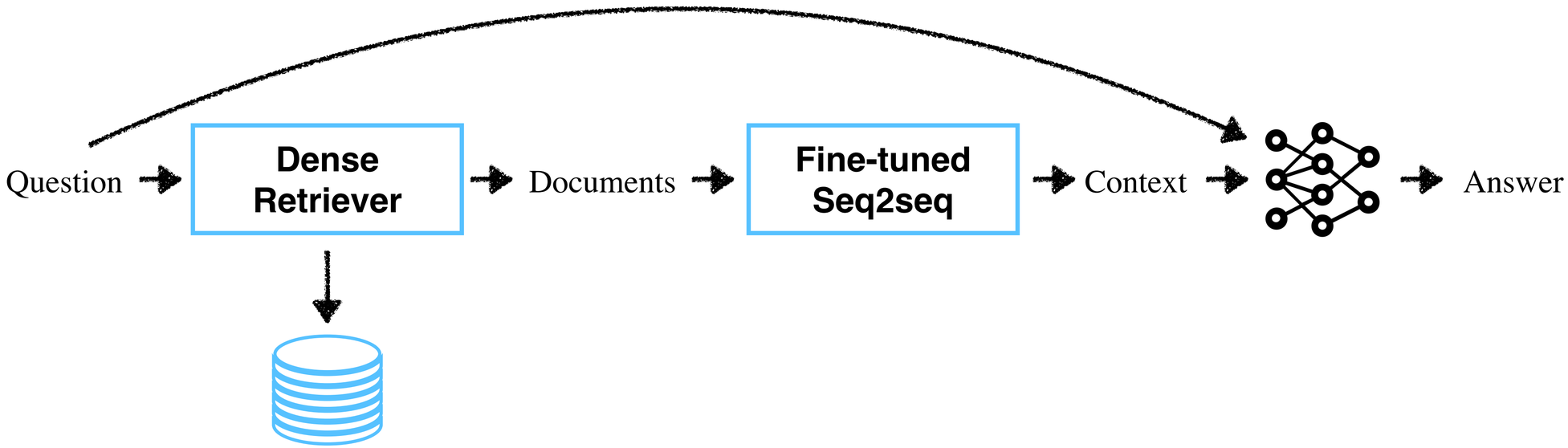


$$L_i^{seq}(x_i, y_i, \theta) = -\sum_{t=1}^{m} logp_\theta(y_i^t \mid x_i, y_i^{<t})$$

$$L_i^{ner}(x_i, z_i, \theta_{enc}) = -\sum_{t=1}^{k} logp_{\theta_{enc}}(z_i^t \mid x_i)$$

$$\therefore L_i = \alpha \times L_i^{seq} + \beta \times L_i^{ner}$$

Methodology - inference



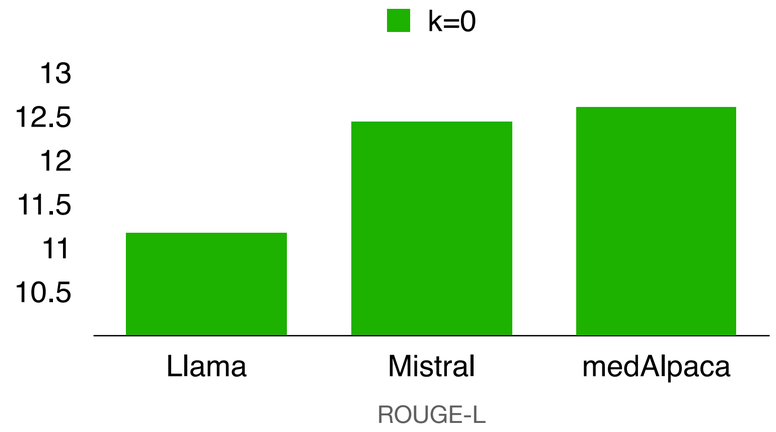
Experimental details

- **Specific-domain tasks** (medical QA)
  - MedQuAD (Ben Abacha et al., 2019)
  - Mash-QA (Zhu et al., 2020)
- **Seq2Seq model**
  - Bart (Lewis et al., 2019)
- **Retriever**
  - Contriever (Izacard et al., 2021)
- **Reader LMs**
  - Llama2-7B (Touvron et al., 2023)
  - Mistral-7B-v0.1 (Jiang et al., 2023)
  - medAlpaca-7B (Han et al., 2023)

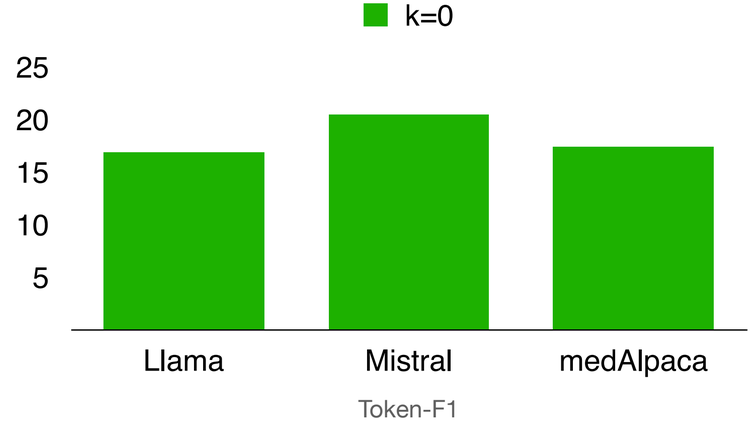
Experimental results

k=0: No Retrieval-augmentation

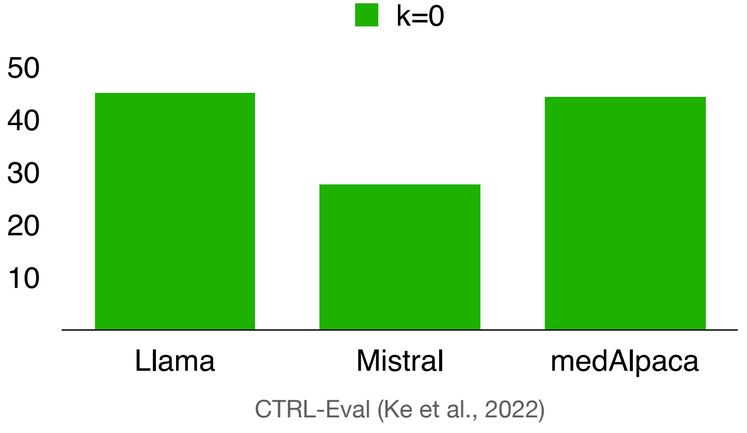
MedQuAD



ROUGE-L

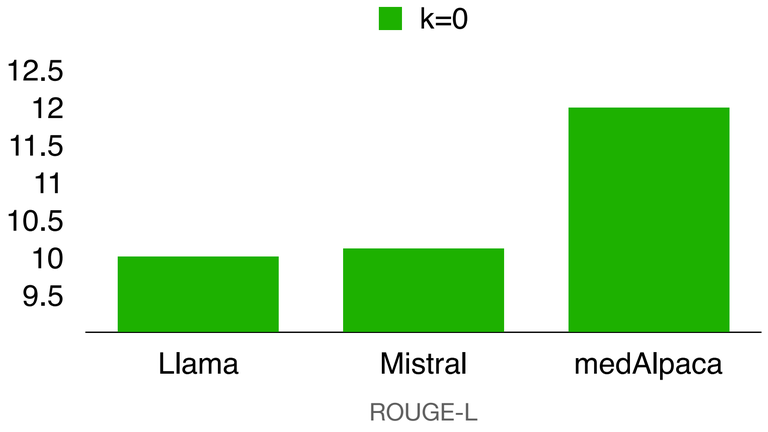


Token-F1

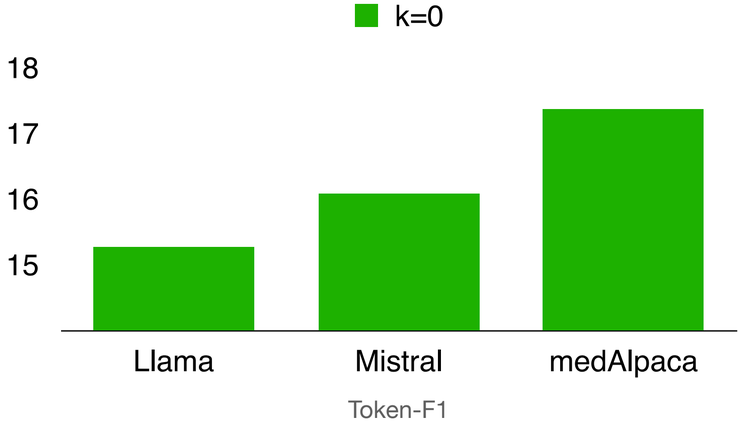


CTRL-Eval (Ke et al., 2022)

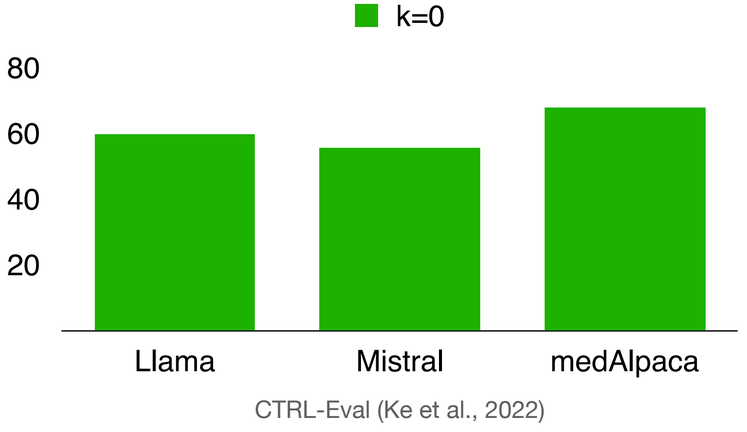
Mash-QA



ROUGE-L



Token-F1



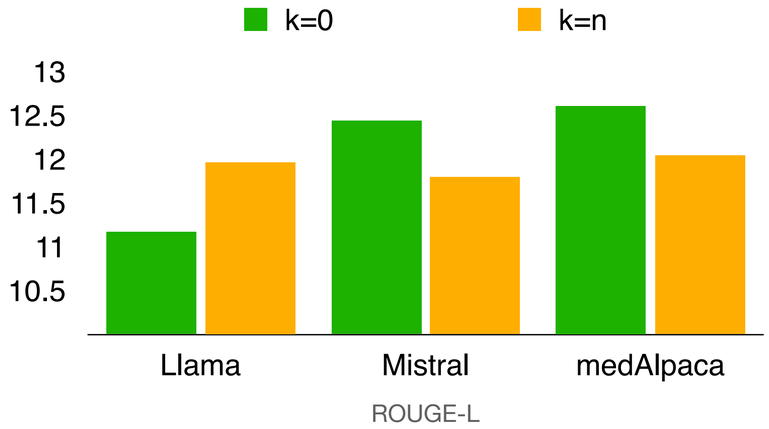
CTRL-Eval (Ke et al., 2022)

Experimental results

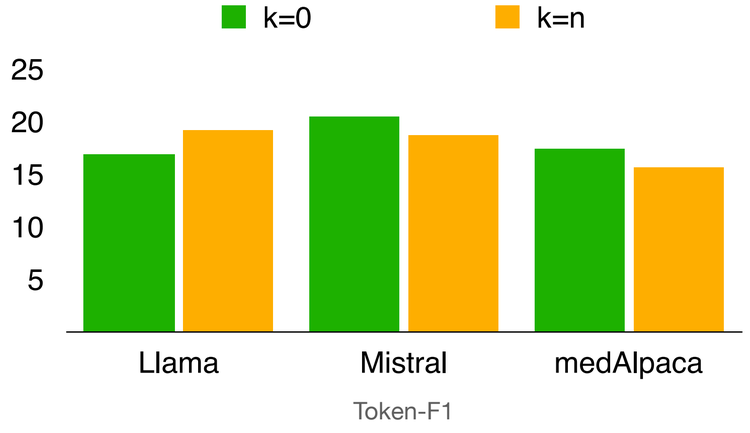
k=0: No Retrieval-augmentation

k=n: 1,5,10개의 검색 증강 질의응답 성능 중 높은 성능

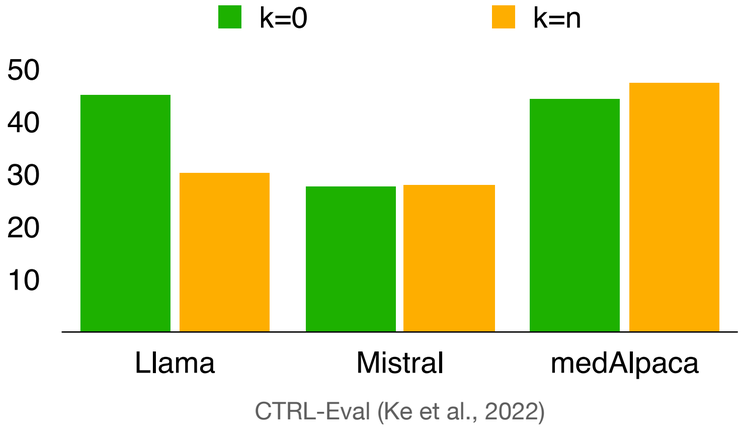
MedQuAD



ROUGE-L

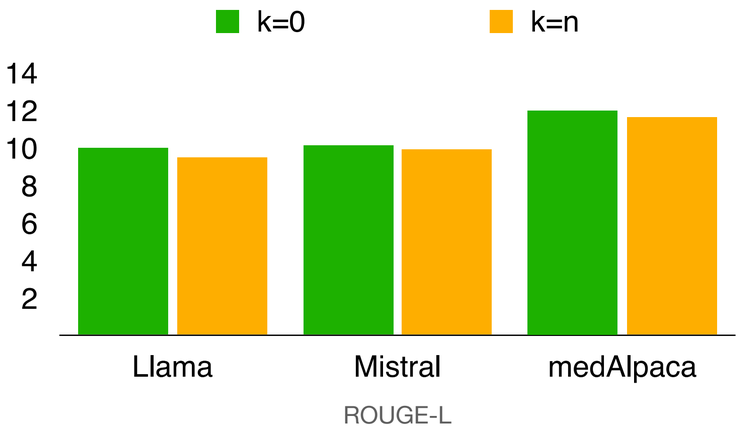


Token-F1

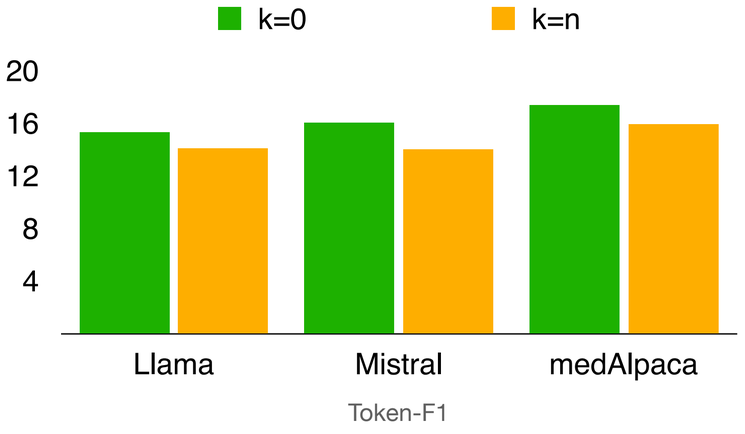


CTRL-Eval (Ke et al., 2022)

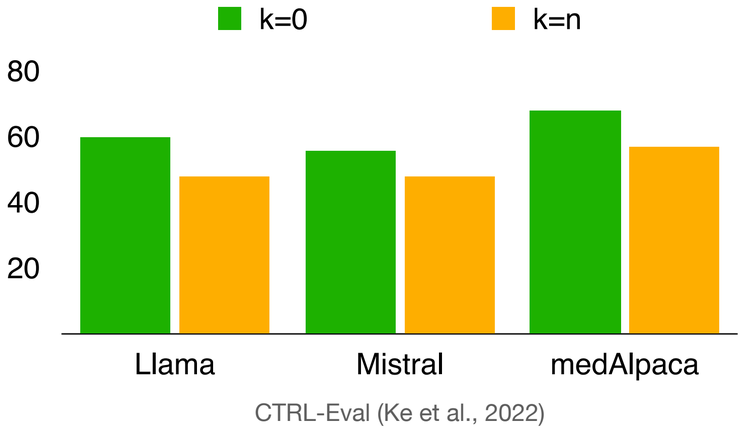
Mash-QA



ROUGE-L



Token-F1



CTRL-Eval (Ke et al., 2022)

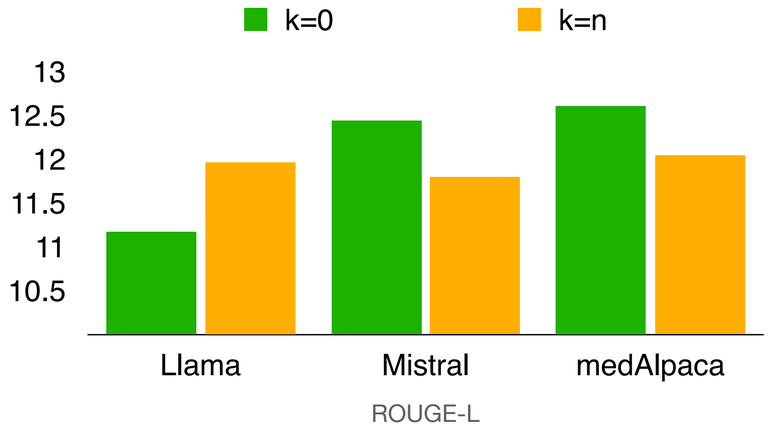
Experimental results

k=0: No Retrieval-augmentation

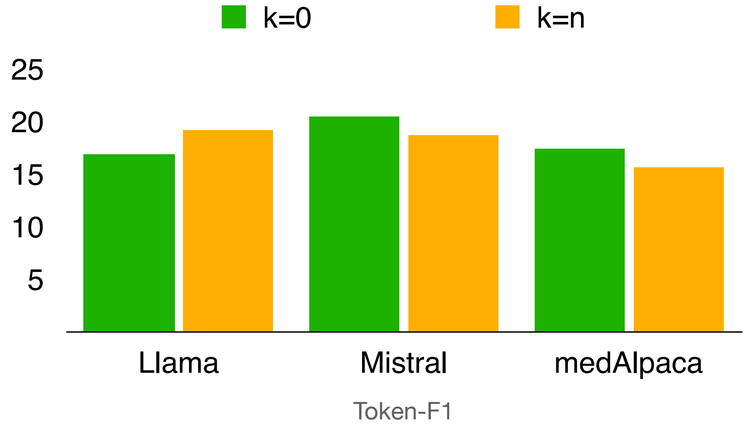
k=n: 1,5,10개의 검색 증강 질의응답 성능 중 높은 성능

의료 도메인에 있어서 검색증강 방식이 효과적이라고 보기 힘들다.

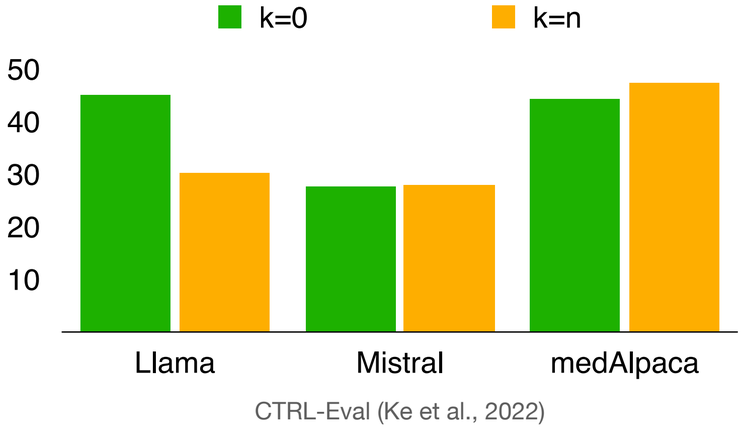
MedQuAD



ROUGE-L

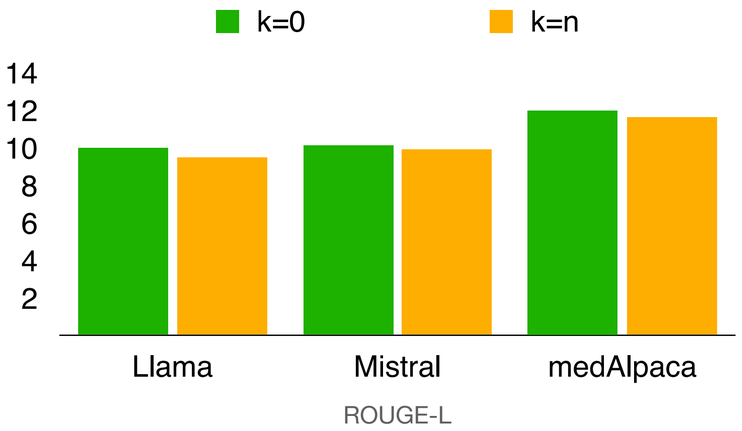


Token-F1

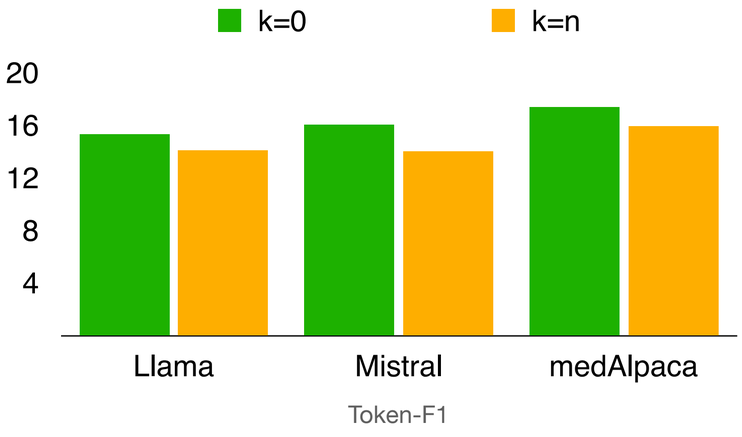


CTRL-Eval (Ke et al., 2022)

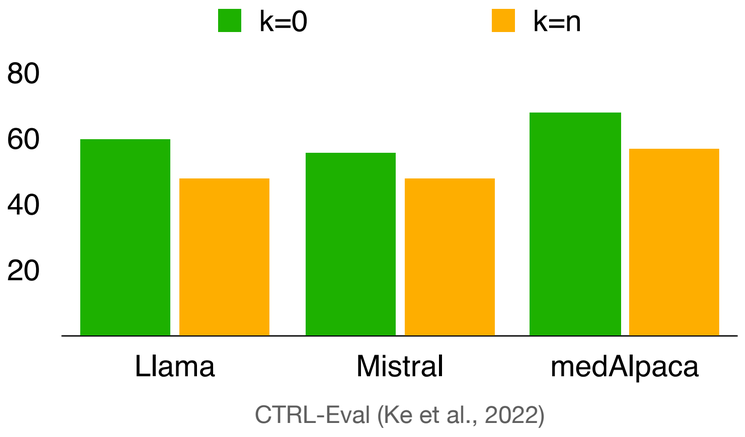
Mash-QA



ROUGE-L



Token-F1



CTRL-Eval (Ke et al., 2022)

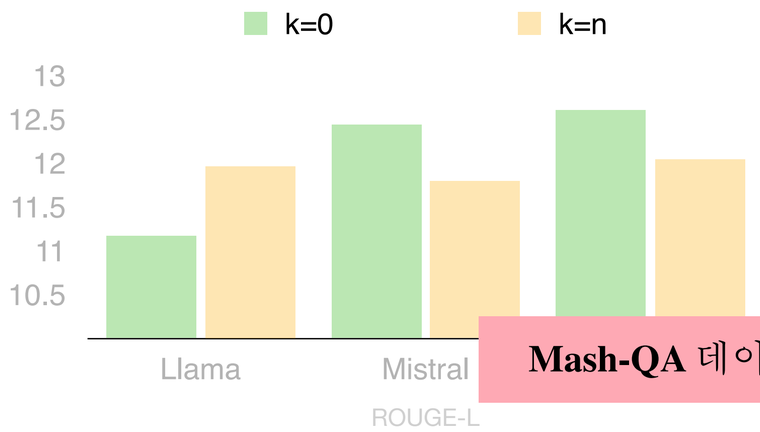
Experimental results

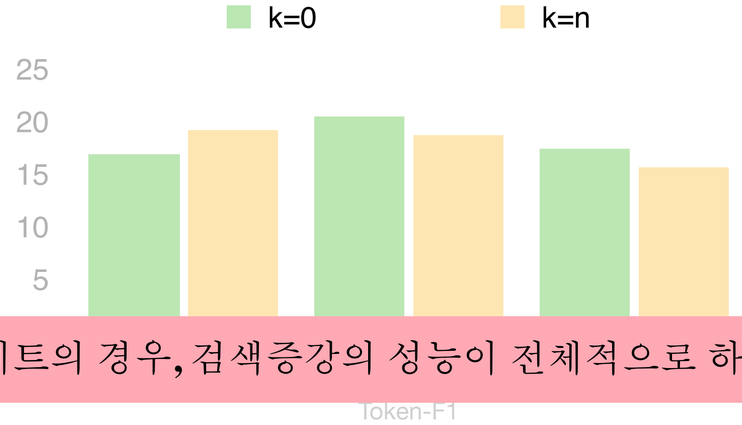
k=0: No Retrieval-augmentation

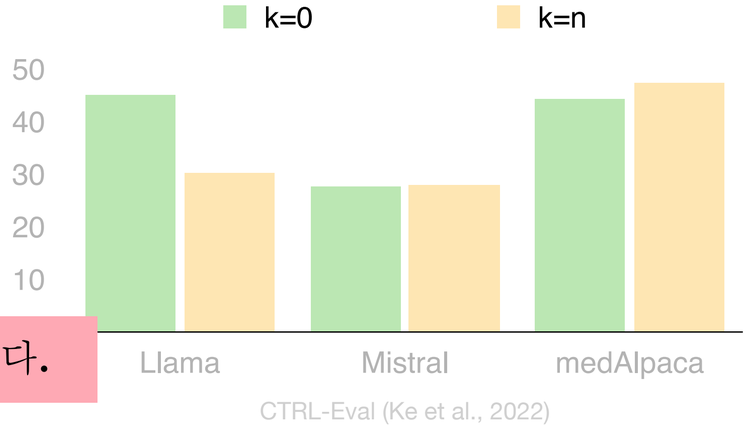
k=n: 1,5,10개의 검색 증강 질의응답 성능 중 높은 성능

의료 도메인에 있어서 검색증강 방식이 효과적이라고 보기 힘들다.

MedQuAD

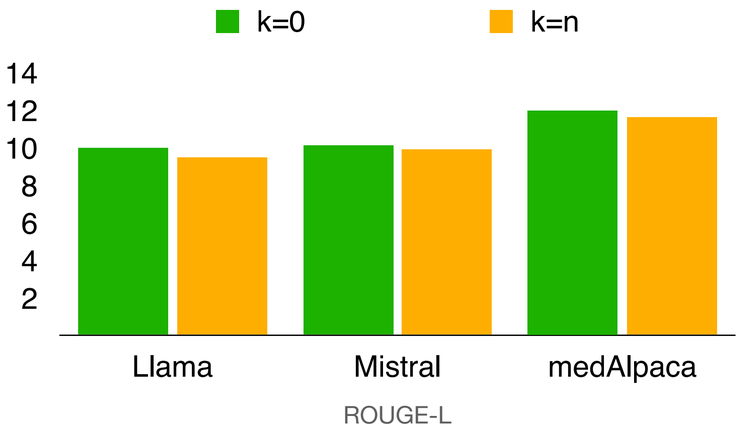


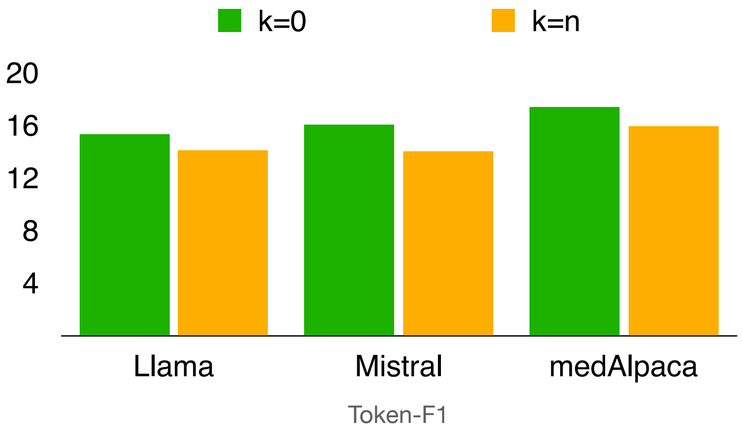


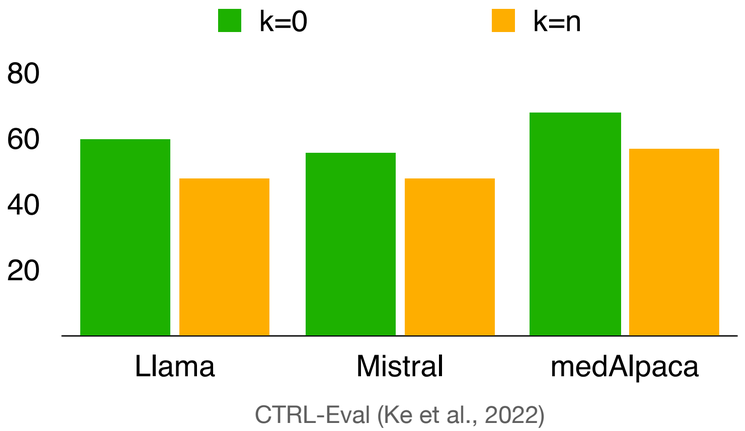


Mash-QA 데이터세트의 경우, 검색증강의 성능이 전체적으로 하락한다.

Mash-QA







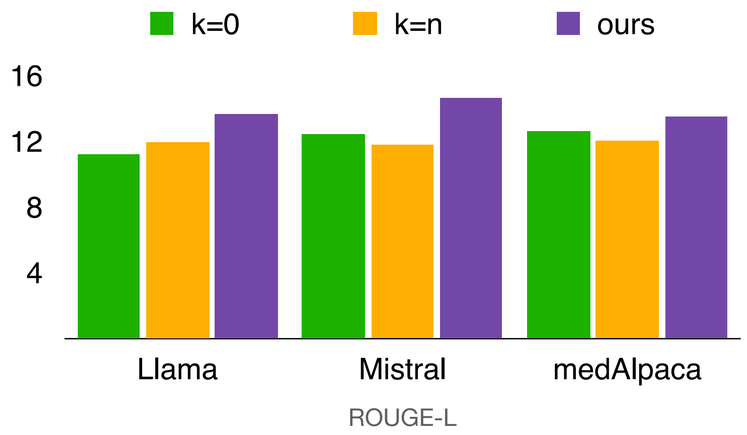
Experimental results

k=0: No Retrieval-augmentation

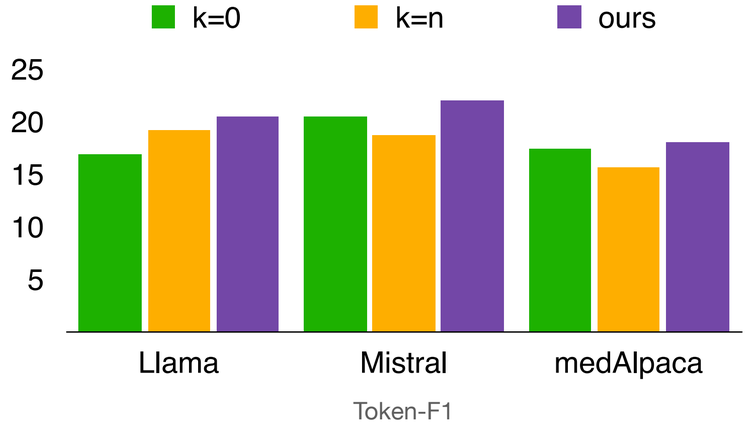
k=n: 1,5,10개의 검색 증강 질의응답 성능 중 높은 성능

모든 언어모델에서 베이스라인을 전반적으로 능가한다.

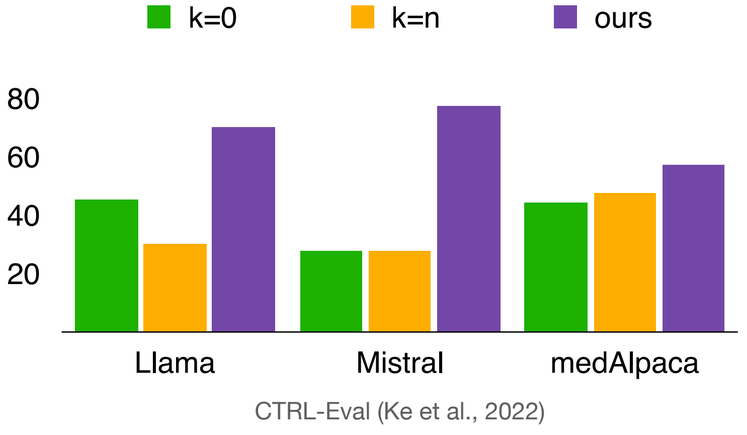
MedQuAD



ROUGE-L

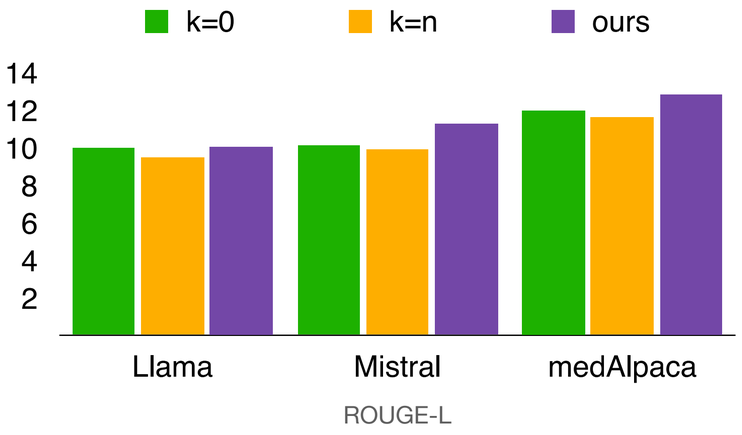


Token-F1

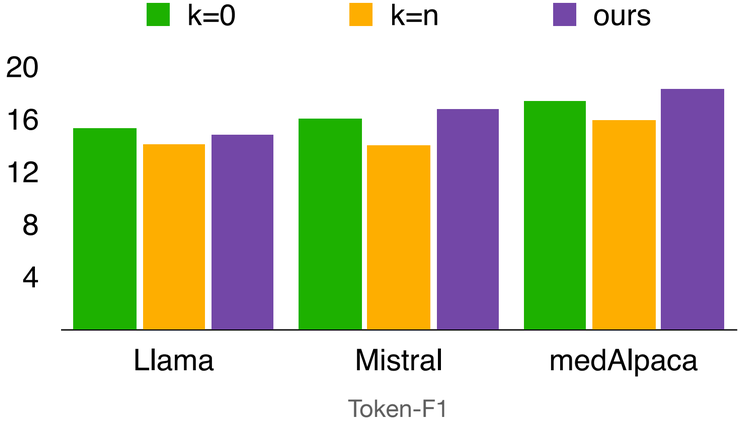


CTRL-Eval (Ke et al., 2022)

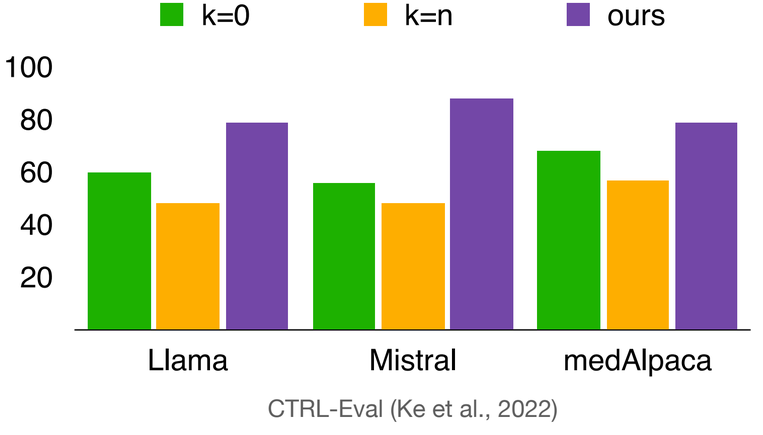
Mash-QA



ROUGE-L



Token-F1



CTRL-Eval (Ke et al., 2022)

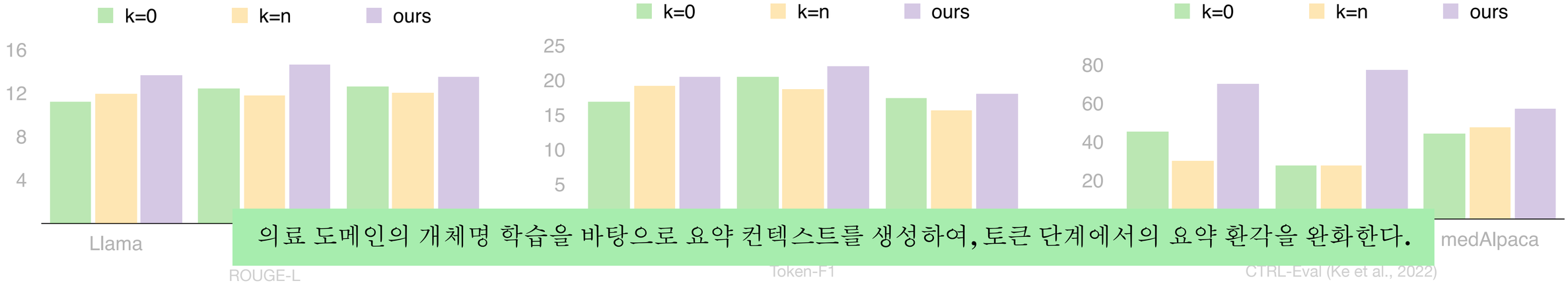
Experimental results

k=0: No Retrieval-augmentation

k=n: 1,5,10개의 검색 증강 질의응답 성능 중 높은 성능

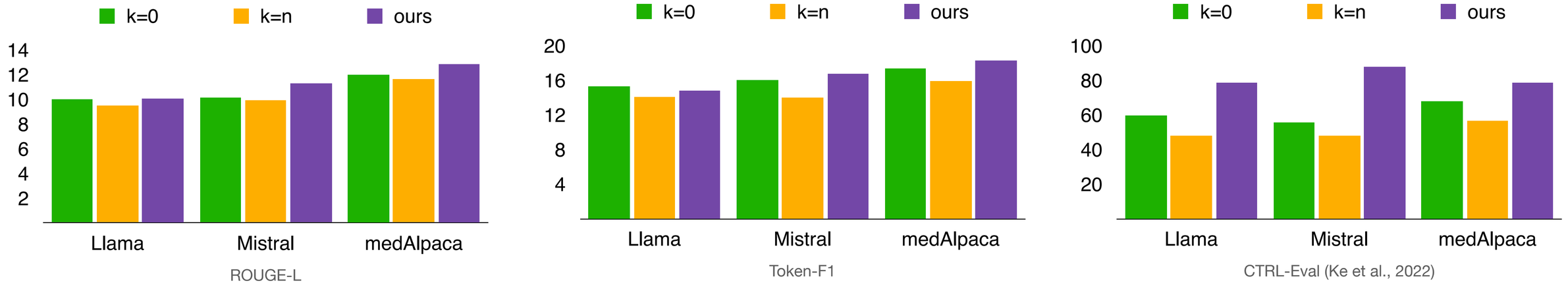
모든 언어모델에서 베이스라인을 전반적으로 능가한다.

MedQuAD



의료 도메인의 개체명 학습을 바탕으로 요약 컨텍스트를 생성하여, 토큰 단계에서의 요약 환각을 완화한다.

Mash-QA



Experimental results

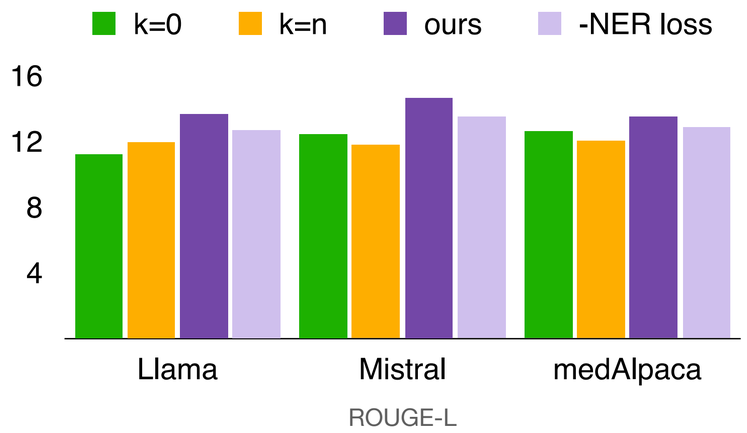
k=0: No Retrieval-augmentation

k=n: 1,5,10개의 검색 증강 질의응답 성능 중 높은 성능

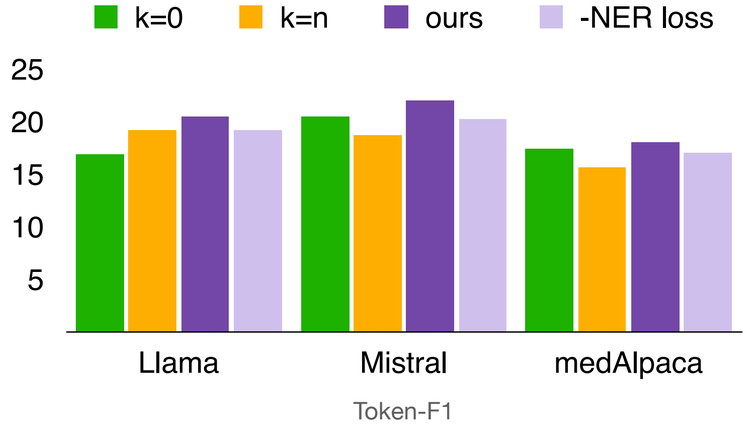
-NER loss: only summarization task

NER task의 학습의 효과를 관찰하기 위하여 **summarization task**만 학습한다.

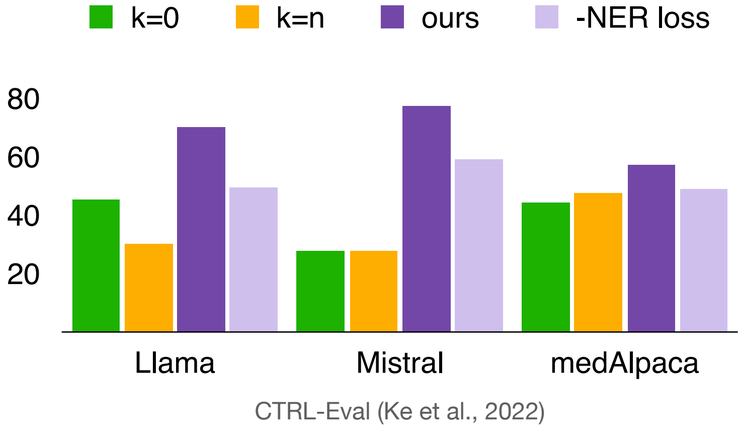
MedQuAD



ROUGE-L

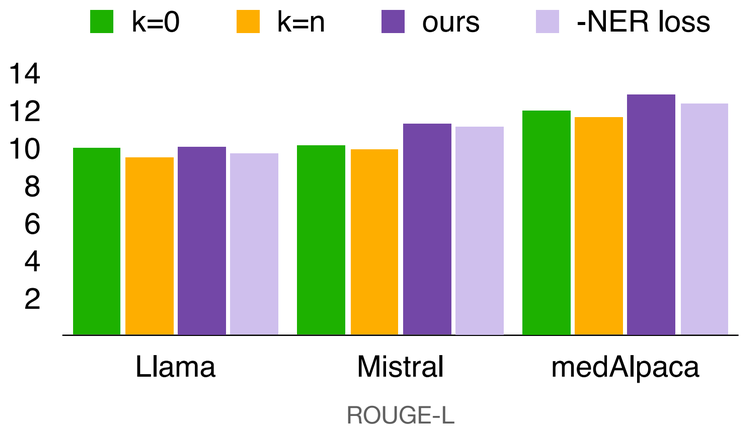


Token-F1

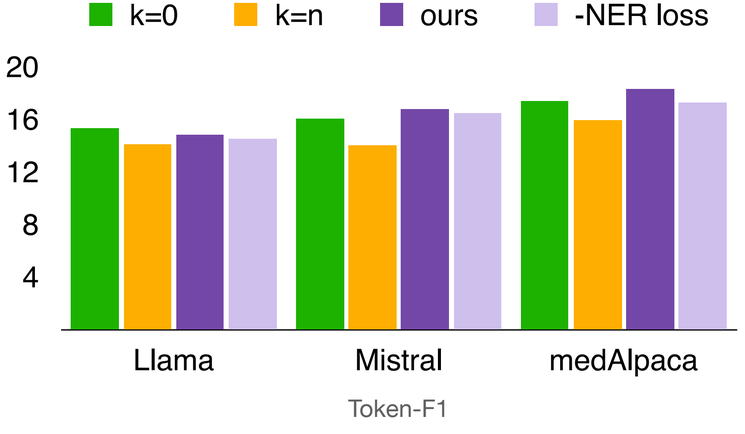


CTRL-Eval (Ke et al., 2022)

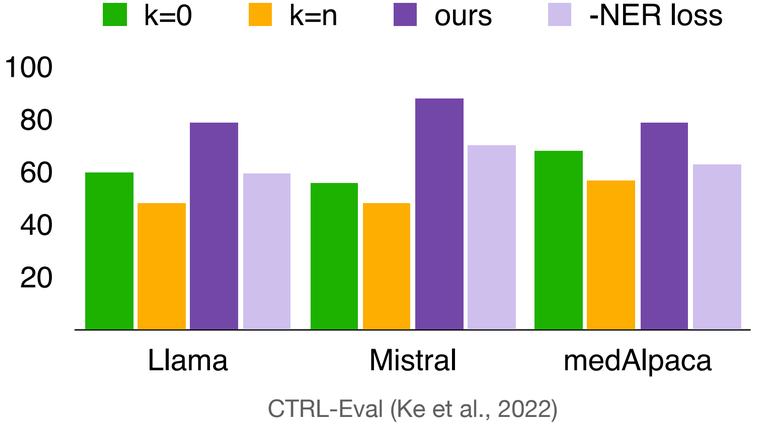
Mash-QA



ROUGE-L



Token-F1



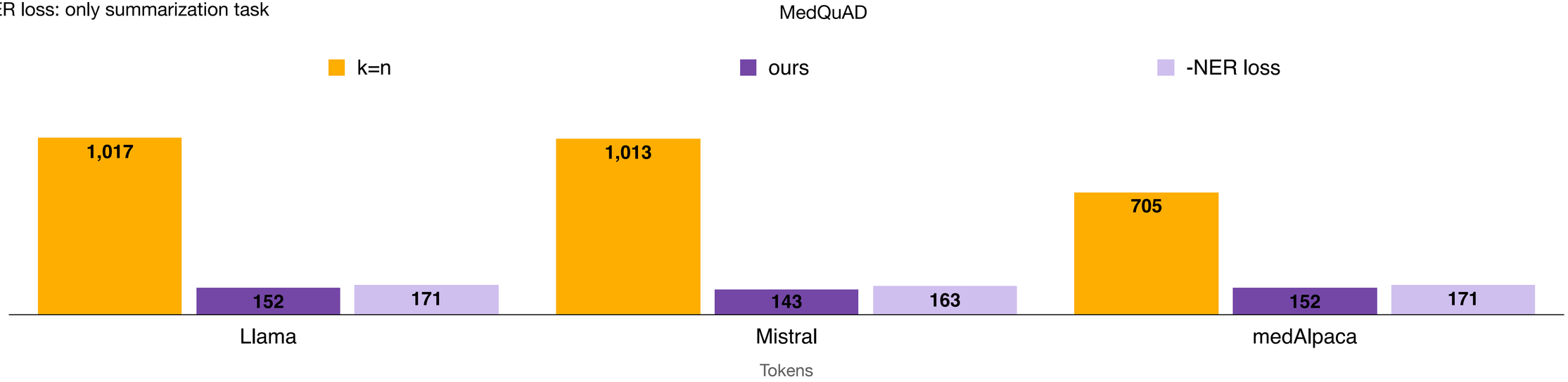
CTRL-Eval (Ke et al., 2022)

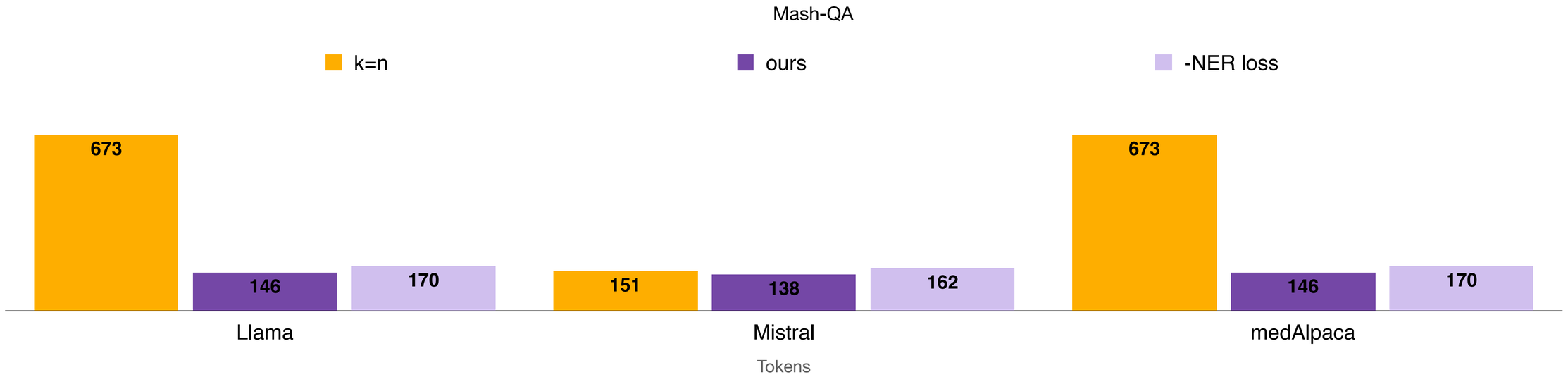
Experimental results

k=0: No Retrieval-augmentation

k=n: 1,5,10개의 검색 증강 질의응답 성능 중 높은 성능

-NER loss: only summarization task





Thank you so much for listening!