

# HyperCLOVA 요약 서비스: Long-Form Open Domain Dialogue Summarization

김형준, 정승환

NAVER Cloud

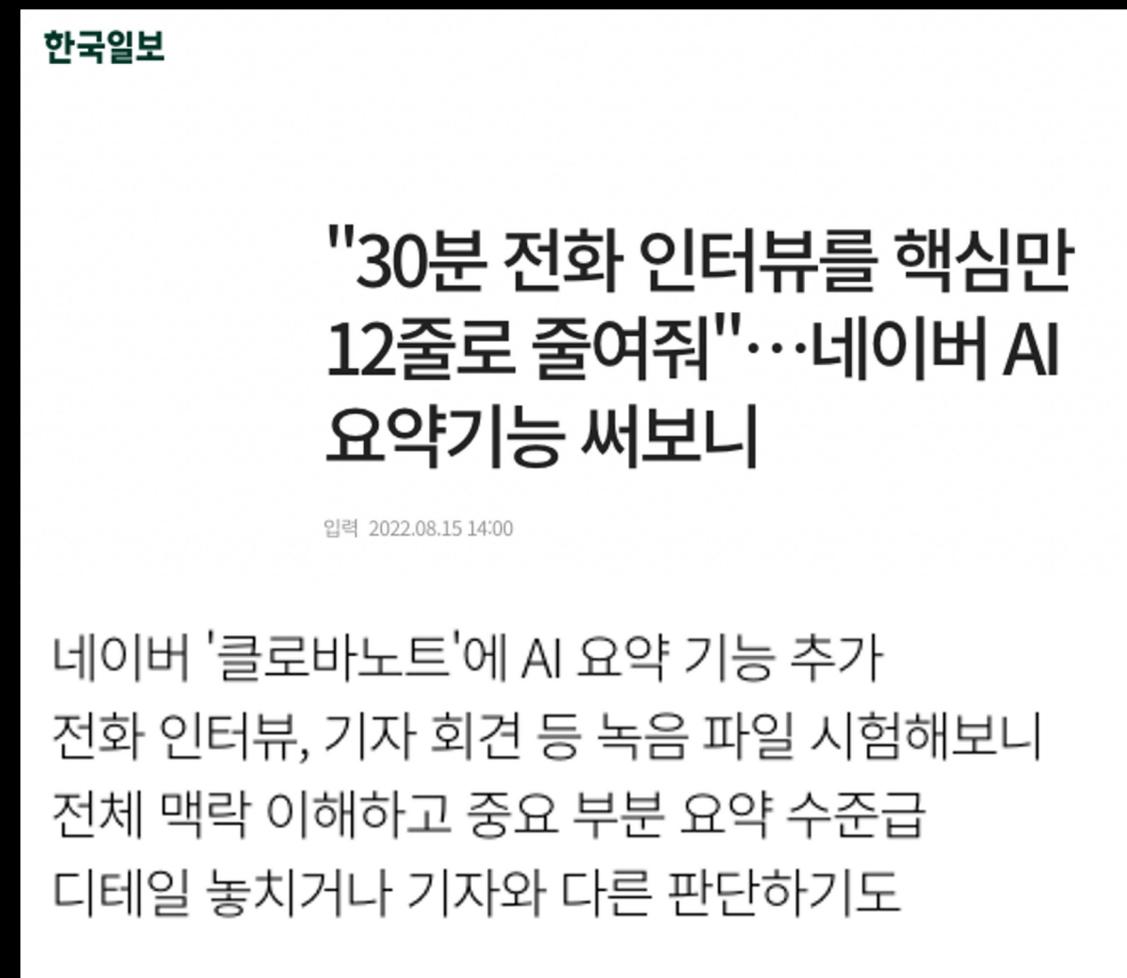
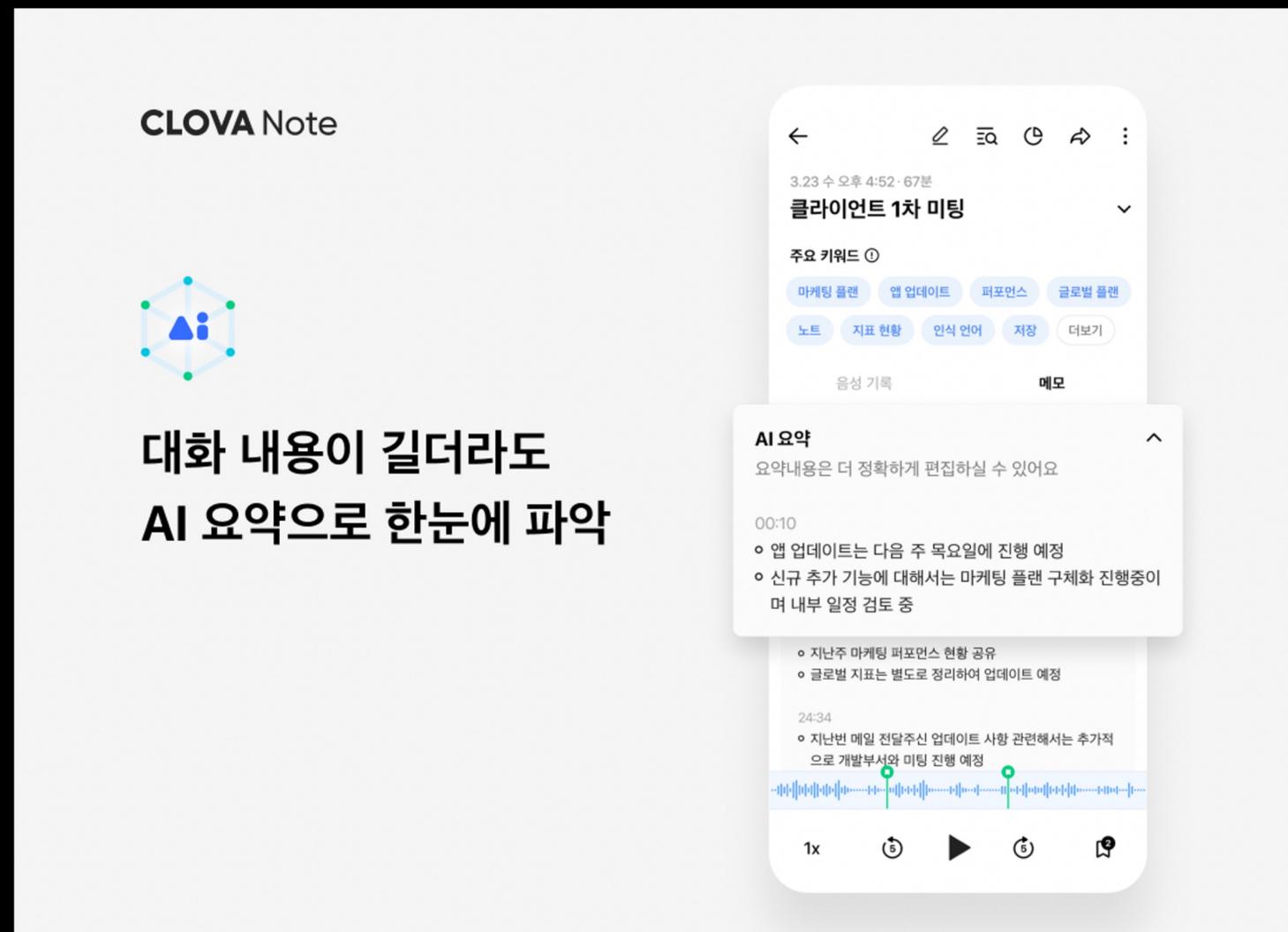
# CONTENTS

1. Introduction
2. HyperCLOVA Summarization
3. Topic Segmentation
4. Deploying & Serving
5. Application
6. Epilogue

# 1. Introduction

# 1.1 Use Case - CLOVA Note

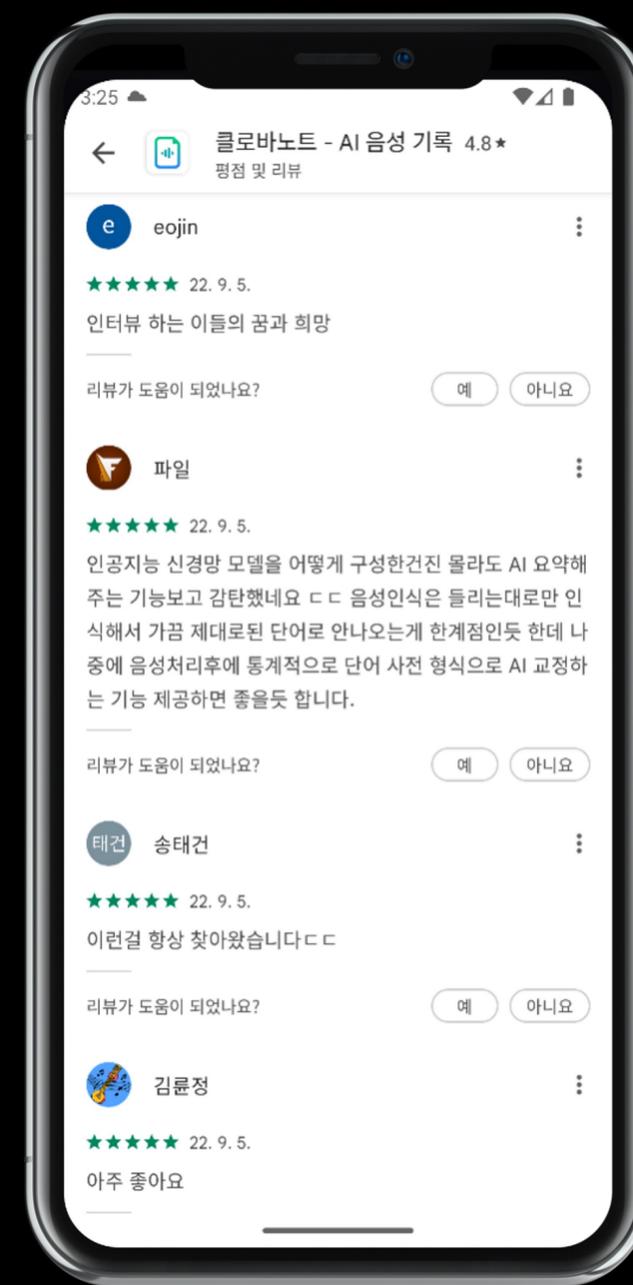
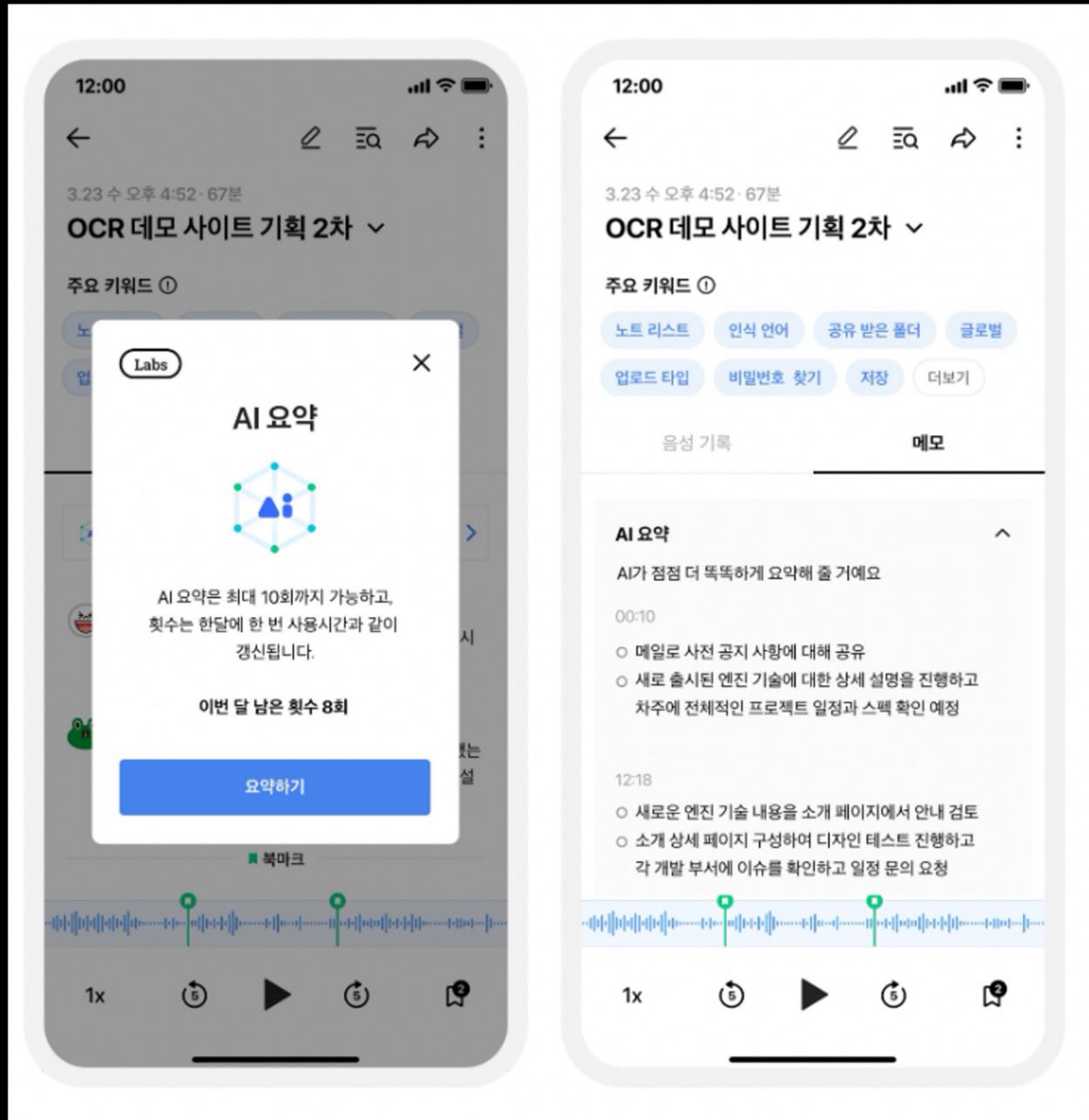
22.08.12 CLOVA Note Generative Summarization 출시



<https://www.hankookilbo.com/News/Read/A2022081111080004766>

# 1.1 Use Case - CLOVA Note

리뷰 중 "어떻게 구성한건진 몰라도 AI 요약해주는 기능보고 감탄했네요"



# 1.1 Use Case - CLOVA Note

## CLOVA Note 강연 요약 예시

### Clova Tech Summit 2 2 대화모델 엔진 구조와 Chatbot 개발 최적화 방안

전체 노트 · 2022.8.8 월 오후 10:02(오후 4:12 수정) · 44분 9초

김형준

음성 기록 + 메모 보기 ▾

**음성 기록** 🔒 점유율   편집

**참석자 1** 00:08

안녕하세요. 현재 씨케이 프레임워크 개발을 맡고 있는 신성진이라고 합니다. 앞에서 어떻게 보면 자연어 처리 이런 내용에 대해서 기초적으로 잘 설명을 해 주셨고 저는 이제 실제로 서비스를 해보면서 어떻게 이거를 적용했는지 데이터를 어떤 구조로 했는지 그리고 모델 중에서 요즘에 조금 핫한 딥러닝을 가지고 어떻게 좀 이렇게 대화를 풀어보려고 했는지에 대해서 간단히 설명을 드리겠습니다.

**참석자 1** 00:35

그래서 오늘 말씀드린 내용은 간단하게 이제 전체적인 개요를 설명드리고 그리고 이제 모델 구조 그리고 익스텐션 가이드 챗봇 최적화 방안까지 여러 가지 사례를 보여드리면서 앞으로 챗봇을 개발하실 분이나 서비스로 적용하실 분들에게 여러 가지 이야기를 드리고 싶은 자리고 공유를 드리겠습니다. 해서 전체적으로 한번 보시면 이번에 작년 12월에 첫 번째로 블로그 고객센터 챗봇이라는 걸 출시를 하면서 그 이후로 한 30개 이상의 서비스와 챗봇을 출시를 했습니다.

**참석자 1** 01:09

출시한 것도 있고 poc로 준비한 것도 있고요 지금 현재는 8개 모바일 pc에서 서비스로 출시하고 있고 앞으로 이번 연 이후에도 지속적으로 출시를 할 예정이고요 이벤트 챗봇이든지 지금 현재 클로바에 들어가 있는 대화 모델 그리고 각종 제휴 타 회사랑 각종 제휴 poc를 하면서 여러 가지를 경험을 했고요 그래서 왜 씨케이 프레임워크라는 거를 구축을 했는지 그리고 이제 어떻게 하면 씨케이 프레임워크 챗봇 프레임워크를 가지고 좀 더 쉽게 개발을 할 수 있는지

**참석자 1** 01:41

그리고 챗봇을 가지고 개발을 최적화하려면 어떻게 하는지에 대해서 설명을 드리겠습니다. 첫 번째로 말씀드린 거는 자연어 처리가 어려운 이유에 대해서 대표적인 사례를 하나를 설명드리고 싶어요. 요즘에 이제 기계 학습이나 딥러닝이나 이런 식으로 하는데 결국에는 언어라는 것은 단순히 컴퓨터로 표현하기가 어려운

**참석자 1** 02:03

요소 중에 하나라고 생각을 합니다. 보시면 이미지와 같은 경우는 이제 딥러닝 기술 중에서 특징을 추출하는 기술인데 이미지라는 하나 그림 중에서 어떻게 생겼는지 배가 어떻게 생겼는지 이 환경이 어떻게 생겼는지에 대한 정보를 뽑아올 수 있습니다. 하지만 텍스트의 경우에는 조금 사례가 다릅니다. 텍스트와 같은 경우에는 아무래도 기본 과정 자체가 주변 단어에 대해서 그 의미를 자주 출연하는 빈도수나 이런 걸로 예측을 하다 보니

**메모**

**주요 키워드** ⓘ   챗봇 모델   챗봇 프레임워크   test   쿼리   잡담   학습   로그 데이터   답변   클로바   날씨 도메인   단어   모델 구조   베이스 모델   벡터   딥러닝   고객센터   템플   생성   형태소   입력

**AI 요약**

AI가 점점 더 똑똑하게 요약해 줄 거예요. 편집해서 내용을 더해보세요.

00:08

- 씨케이 프레임워크 개발을 맡고 있는 신성진이라고 함
- 씨케이 프레임워크 챗봇 프레임워크를 가지고 좀 더 쉽게 개발을 할 수 있는지 설명을 드리겠음

04:21

- 챗봇 프레임워크 ck에 대해 설명하고 있음

09:19

- 한국어 및 대화형의 데이터에 최적화된 대형 모델을 설계함
- 학습을 잘 하는 과정으로 추가를 해서 학습을 잘 하는 과정으로 넣음
- 학습을 하면서 추출하는 fite는 다르다고 생각함

14:05

- 클로바 챗봇 익스텐션 가이드를 만든 모델을 활용하는 방법을 설명하고 있음

17:48

- 채널 연동을 통해 쉽게 모델을 만들고 연동을 할 수 있음
- 챗봇에 대한 기본적인 발화에 대해서는 공수를 줄일 수 있음

21:12

- 챗봇을 만들 때는 양과 품질에 대해서 신경을 써야 함
- 챗봇을 만들 때는 초기 데이터 생성이 중요함

Clova Tech Summit 2 - #2 대화모델 엔진 구조와 Chatbot 개발 최적화 방안, <https://www.youtube.com/watch?v=swjnlM1pNzs>

# 1.2 Use Case - CLOVA Studio API

네이버클라우드 클로바 스튜디오에서 만나요!

**CLOVA Studio**를 활용해 다양한 문제를 해결하고, 누구나 쉽게 AI를 개발할 수 있습니다.

**주요 능력**

**문장 생성**  
다양한 분야의 문장을 만들어 낼 수 있습니다. 몇 가지 키워드만 입력하면 해당 키워드 기반의 텍스트를 생성합니다.  
예) 시나리오 창작, 자기소개서 작성, 이메일 생성, 마케팅 문구 창작 등

**요약**  
긴 글에서 주요 요소를 파악하여 짧게 요약할 수 있습니다. 요약문의 길이를 직접 설정할 수 있습니다.  
예) 줄거리, 보고서, 이메일 요약 등

## 초대규모 AI HyperCLOVA Tuning

### 언어 모형의 장점 극대화

- 일반화(Open-Domain)
- 유창성(Fluency)

### 언어 모형의 단점 최소화

- 관련성(Relevance / Hallucination)
- 주제와의 통일성(Coherence)
- 일관성(Consistency)

→ "Free lunch"

특정 도메인 튜닝 없이, 오직 생성 모델로 서비스

<https://www.ncloud.com/product/aiService/clovaStudio>

- B. Kim., et. al. What Changes Can Large-scale Language Models Bring? Intensive Study on HyperCLOVA: Billions-scale Korean Generative Pretrained Transformers. EMNLP, 2021
- LIU, Xiao., et. al. GPT Understands, Too. arXiv:2103.10385, 2021
- HU, Edward J., et. al. LoRA: Low-Rank Adaptation of Large Language Models. arXiv:2106.09685, 2021
- Jack W. et. al. Scaling Language Models: Methods, Analysis & Insights from Training Gopher. arXiv: 2112.11446, 2021

# 1.3 Use Case - Web Browser

## 문서 요약 방법 및 시스템 (특허출원)

다양한 형태의 문서를 토픽별로 요약, 요약된 정보 탐색의 효율을 높이는 요약 방법 및 시스템

### 한글도

”

생생함이란 바로 이런 것?  
1 프로다운 '클로바 보이스 PRO'

- 클로바 보이스 PRO는 HDTS 기술로 탄생한 클로바 보이스임
- 고도화된 기술 덕에 기존의 보이스보다 더욱 생생하고 자연스러운 보이스가 탄생함

compression ratio: 23.0 (82 / 357) %

ESC to close  
press 'p' to copy  
double-click to detail :

언어, 성별, 연령, 스타일에 따라 다양한 매력을 자랑하는 클로바더빙의 AI 보이스! 그중 유독 프로미 뽐뽐하는 보이스들이 나타났습니다.

바로 HDTS (Hybrid DNN Text-to-Speech) 기술로 탄생한 클로바 보이스 PRO가 그 주인공입니다. HDTS 기술은 적용되는 분야의 특성을 살린 최고 품질의 AI 보이스를 생성하는 기술인데요, 고도화된 기술 덕에 기존의 보이스보다 더욱 생생하고, 자연스러운 보이스가 탄생했습니다.

한 번 듣는 것이 백 번 보는 것보다 나을 때가 있죠. 클로바 보이스 PRO의 진가! 아래 영상을 통해 체험해 보세요!

[https://blog.naver.com/clova\\_ai](https://blog.naver.com/clova_ai)

### 영문도

Status and Types of NAVER's Digital Twin Technologies

- Apollo 13은 예상치 못한 사고로 산소 탱크가 폭발함
- NASA는 15개의 시뮬레이션을 통해 문제를 해결함
- 디지털 트윈 기술은 디지털 세계에 물리적 세계를 복제하는 것임

compression ratio: 7.8 (102 / 1307) %

ESC to close  
press 'p' to copy  
double-click to detail :

1 Apollo 13 was involved in an unexpected accident – the probe's oxygen tank had exploded. This happened on April 11, 1970, 330 thousand kilometers from Earth. At the time, NASA was unsure of the solution. They could not go out into space and have a look at the probe to diagnose the problem. Instead, NASA had 15 different simulators. By replicating the conditions of the broken probe and repeating tests, they were finally able to find a solution. On April 17, the three lunar astronauts returned safely to Earth.

NASA's attempt to create an environment identical to the reality and repeat simulations to solve a problem nearly 50 years ago is very similar to the modern 'digital twin' technology. Digital twin refers to the replication of the physical world in the digital world. This is why it is also referred to as the mirror world. You can replicate a space or object in 3D to conduct various simulation analyses. In addition, you can create services where the physical and digital worlds are synchronized in real-time. It can also be used in smart city monitoring and predictions, simulating environmental change within a city, innovative mobility services like autonomous driving, robot services and automation systems of a smart building, and in providing people with vivid and realistic 3D content.

<https://www.naverlabs.com/en/storyDetail/252>

# 1.3 Use Case - Web Browser

## 문서 요약 방법 및 시스템 (특허출원)

다양한 형태의 문서를 토픽별로 요약, 요약된 정보 탐색의 효율을 높이는 요약 방법 및 시스템

### 1 BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova

We introduce a new language representation model called BERT, which stands for Bidirectional Encoder Representations from Transformers. Unlike recent language representation models, BERT is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. As a result, the pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, without substantial task-specific architecture modifications.

- 언어 이해를 위한 딥 양방향 변환기의 사전 훈련을 소개함
- 최근 언어 표현 모델과 달리 BERT는 라벨이 없는 텍스트에서 양쪽 문맥을 조건부로 공동으로 사전 훈련함
- 사전 훈련된 BERT 모델은 질문에 대한 답이나 언어 추론과 같은 넓은 범위의 작업을 위해 한 개의 추가 출력 레이어로 조정할 수 있음

compression ratio: 23.0 (172 / 748) %

ESC to close  
press 'p' to copy  
double-click to detail:

(or [arXiv:1810.04805v2](https://arxiv.org/abs/1810.04805v2) [cs.LG] for this version)

## 논문도

<https://arxiv.org/abs/1810.04805>

## 쿼리도

### INNER JOIN Syntax

```
SELECT 1column_name(s)  
FROM table1  
INNER JOIN table2  
ON table1.column_name = table2.column_name;
```

- 테이블 1과 테이블 2를 조인해서 컬럼의 이름을 가져옴

compression ratio: 33.3 (32 / 96) %

ESC to close  
press 'p' to copy

# 1.4 Summarization

그냥 LLM(Large-scale Language Model) 혹은 GPT(Generative Pre-trained Transformer) 로 되는 것 아닌가요?

The screenshot shows the OpenAI website's 'TL;DR summarization' interface. It features a navigation bar with 'API', 'RESEARCH', 'BLOG', and 'ABOUT'. The main content area includes a 'Transformation' tab, an 'Open in Playground' button, and a text input field with a sample prompt: 'Summarize text by adding a 'tl;dr:' to the end of a text passage. It shows that the API understands how to perform a number of tasks with no instructions.' Below the prompt, there are 'Settings' for 'Engine' (text-davinci-002), 'Max tokens' (60), 'Temperature' (0.7), 'Top p' (1.0), 'Frequency penalty' (0.0), and 'Presence penalty' (0.0). A 'Prompt' section contains a detailed paragraph about neutron stars, followed by the instruction 'Tl;dr'.

The HyperCLOVA logo is displayed in large, stylized letters. Below it, the text 'AI, 모두의 능력이 되다' (AI, everyone's ability becomes) is shown. A central diagram features '초대규모 AI HyperCLOVA CLOVA Studio' in a blue box, surrounded by various application areas: '캐릭터와 대화하기', '데이터 생성하기', '기획진 구성하기', '음식 레시피 만들기', '역사적 인물과 대화하기', '콘텐츠 분류하기', '마케팅 문구 생성하기', and '소설 쓰기'. Other related tasks include '비속어 필터하기', '메뉴 추천해주기', '셀럽 말투 따라하기', '사투리 변환하기', '가사 쓰기', '백과사전 물고담하기', '컨텐츠 분류하기', '반려동물 지식백과', '의도 이해하기', '뉴스 요약해주기', '상품 리뷰 요약하기', '웹페이지 만들기', '경쟁결과 요약', and '엔터티 인식하기'.

## 2. HyperCLOVA Summarization

## 2.1 Dialogue Summarization

### Summarization:

- Text summarization is the process of **distilling the most important information** from a text to produce an abridged version for a particular task and user. by Berry et al. (1995)

## 2.1 Dialogue Summarization

### Summarization:

- Text summarization is the process of **distilling the most important information** from a text to produce an abridged version for a particular task and user. by Berry et al. (1995)

#### 요약하면

- 엔티티(중요 요소)들을 잘 복사하여 주요한 문장들을 추출하고
- 간단한 문장들로 다시쓰기

## 2.1 Dialogue Summarization

### Dialogue Summarization in the wild:

- Filler and Debris(간투어)

음 / 아 / 저기 / 그 / 응

- Repetition and Reiteration

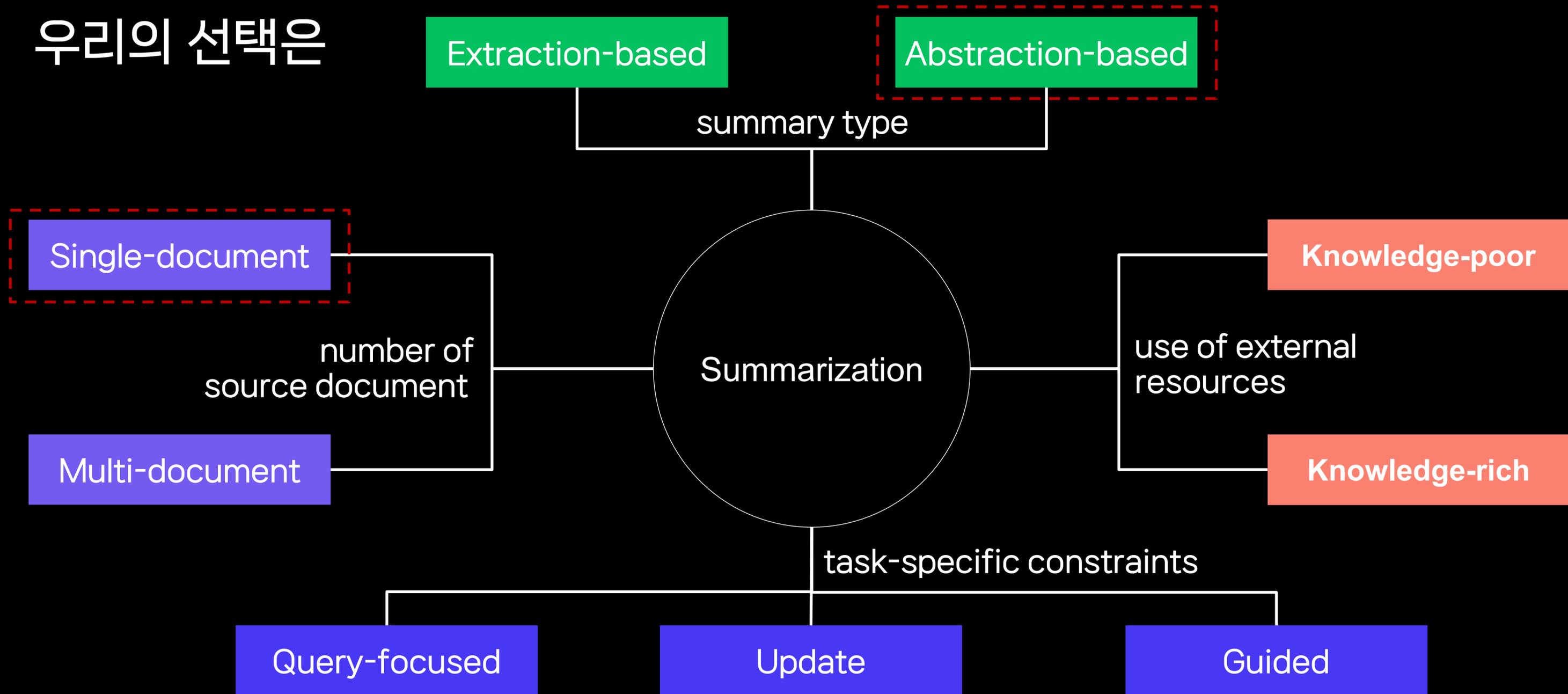
리랭킹을 합니다. 리랭킹을한다음에이제QA시스템에서문서에관련된정보를..문서의 정보를 활용하고요.

- Punctuation

리랭킹을 합니다 리랭킹을 한 다음에

# 2.1 Dialogue Summarization

우리의 선택은



# 2.1 Dialogue Summarization

HyperCLOVA를 믿습니다!



NLG를 Service 출시하기!

Positive

- Naturalness
- Unseen Domain
- Informativeness

Negative

- Hallucination
- Tone & Manner

# 2.1 Dialogue Summarization

-  참가자.\*  
난중일기 임진년 1월 (1592년 1월) 1월 초1일 [양력 2월 13일]<임술> 맑다.
-  참가자.\*  
새벽에 아우 여필(여필)과 조카 봉, 아들 회가 와서 이야기했다.
-  참가자.\*  
다만 어머니를 떠나 남쪽에 서 두번이나 설을 세니 간절한 회포를 이길 길이 없다.
-  참가자.\*  
병마사의 군관 이경신(이경신)이 병 마사의 편지와 설 선물과 장전(장전)과 편전(편전) 등 여러가지 물건을 바치러 가지고 왔다.
-  참가자.\*  
1월 초2일 [양력 2월 14일]<계해> 맑다.
-  참가자.\*  
나라의 제삿날(명종 인순왕후 심씨의 제삿날)임에도 공무를 보았다.

추출 요약의 어려움

직접 요약을 한 번 해보세요!

어떻게 요약할 수 있을까요?

# 2.1 Dialogue Summarization

 참가자.\*  
난중일기 임진년 1월 (1592년 1월) 1월 초1일 [양력 2월 13일]<임술> 맑다.

 참가자.\*  
새벽에 아우 여필(여필)과 조카 봉, 아들 회가 와서 이야기했다.

 참가자.\*  
다만 어머니를 떠나 남쪽에 서 두번이나 설을 세니 간절한 회포를 이길 길이 없다.

 참가자.\*  
병마사의 군관 이경신(이경신)이 병 마사의 편지와 설 선물과 장전(장전)과 편전(편전) 등 여러가지 물건을 바치러 가지고 왔다.

 참가자.\*  
1월 초2일 [양력 2월 14일]<계해> 맑다.

 참가자.\*  
나라의 제삿날(명종 인순왕후 심씨의 제삿날)임에도 공무를 보았다.

- 125M(Bart-FT): "임진년 1월 1월 초 1일과 1월 초 1일, 2월 14일, 계해의 2월 13일에 임진년 1월 1월과 1월 초 1일, 1월 초 2일, 1월 초 1일, 1월 초 1일, 1월 초 1일, 1월 초 2일이다."
- HyperCLOVA: "임진년 1월에 있었던 일을 기록하고 있음"

# 2.1 Dialogue Summarization

- 임진왜란 때의 일기를 날짜별로 정리하고 있음

참가자.\*: 난중일기 임진년 1월 (1592년 1월) 1월 초1일 [양력 2월 13일]<임술> 맑다.

참가자.\*: 새벽에 아우 여필(여필)과 조카 봉, 아들 회가 와서 이야기했다.

참가자.\*: 다만 어머니를 떠나 남쪽에 서 두번이나 설을 세니 간절한 회포를 이길 길이 없다.

참가자.\*: 병마사의 군관 이경신(이경신)이 병 마사의 편지와 설 선물과 장전(장전)과 편전(편전) 등 여러가지 물건을 바치러 가지고 왔다.

참가자.\*: 1월 초2일 [양력 2월 14일]<계해> 맑다.

참가자.\*: 나라의 제삿날(명종 인순왕후 심씨의 제삿날)임에도 공무를 보았다.

참가자.\*: 김인보(김인보)와 함께 이야기했다.

참가자.\*: 1월 초3일 [양력 2월 15일]<갑자> 맑다.

참가자.\*: 동헌(여수시 군자동 진남관 뒷쪽)에 나가 별방군을 점검하고 각 고을과 포구에 공문을 써 보냈다.

참가자.\*: 1월 초4일 [양력 2월 16일]<을축> 맑다.

참가자.\*: 동헌에 나가 공무를 봤다.

참가자.\*: 1월 초5일 [양력 2월 17일]<병인> 맑다.

참가자.\*: 동헌에 나가 공무를 봤다.

참가자.\*: 1월 초6일 [양력 2월 18일]<정묘> 맑다.

참가자.\*: 동헌에 나가 공무를 봤다.

참가자.\*: 1월 초7일 [양력 2월 19일]<무진> 아침에는 맑았다.

참가자.\*: 늦게부터 비와 눈이 번갈아 종일 내렸다.

참가자.\*: 조카 봉이 아산으로 갔다.

참가자.\*: 남원에서 전문(전문: 임 금께 바칠 일종의 글월)을 받들고 갈 유생이 들어왔다.

참가자.\*: 1월 초8일 [양력 2월 20일]<기사> 맑다.

참가자.\*: 객사에 나갔다가 동헌에서 공무를 봤다.

참가자.\*: 1월 초9일 [양력 2월 21일]<경오> 맑다.

참가자.\*: 아침밥을 일찍 먹은 다음, 동헌에 나가 전문을 봉하여 올려 보냈다.

참가자.\*: 1월 초10일 [양력 2월 22일]<신미> 종일 비가 내렸다.

참가자.\*: 방답(전남 여천군 돌산면)에 새 첨사로 이순신(이순신)이 부임하여 들어왔다.

참가자.\*: 1월 11일 [양력 2월 23일]<임신> 종일 가랑비가 내렸다.

참가자.\*: 늦게야 동헌에 나가 공무를 봤다.

참가자.\*: 이봉수(이봉수)가 선생원(전남 여천군 율촌면 성생원)에 돌 뜨는 곳을 가 보고 와서 보고하기를 "이미 큰 돌 열일곱 덩이에 구멍을 뚫었다."고 했다.

참가자.\*: 서문 밖 해자(성 주위를 파서 물을 채운 곳)가 네 발췌 무너졌다.

참가자.\*: 심사립(심사립)과 이야기 했다.

참가자.\*: 1월 12일 [양력 2월 24일]<계유> 곳은 비가 개이지 않다.

참가자.\*: 식사한 뒤에 객사 동헌에 나갔다.

참가자.\*: 본영 및 각 포구의 진무들에게 우등을 가리는 활쏘기를 시합했다.

참가자.\*: 1월 13일 [양력 2월 25일]<갑술> 아침에 흐렸다.

참가자.\*: 동헌에 나가 공무를 봤다.

참가자.\*: 1월 14일 [양력 2월 26일]<을해> 맑다 동헌에 나가 공무를 보고 난 뒤에 활을 쏘았다.

참가자.\*: 1월 15일 [양력 2월 27일]<병자> 흐렸으나 비는 오지 않았다.

참가자.\*: 새벽에 망궤례를 하였다.

참가자.\*: 1월 16일 [양력 2월 28일]<정축> 맑다.

참가자.\*: 동헌에 나가 공무를 봤다.

참가자.\*: 각 고을의 벼슬아치와 색리(고을의 아전) 등이 인사하러 왔다.

참가자.\*: 방 답의 병선을 맡은 군관들과 색리들이 그들 병선을 수리하지 않았기 때문에 곤장을 쳤다.

참가자.\*: 우 후(지방 병마사영이나 수영에 첨사아래에 있는 무관)·가수(가수: 임시 직위)도 역시 점검하 지 않아 이 지경에까지 된 것이니 해괴하기 짝이 없다.

참가자.\*: 공무를 어쭙쭙 여기고, 제 몸만 살 찌러 들며 이와 같이 돌보지 않으니, 앞 날의 일을 알만하다. 성밀에 사는 박용세(박용세)는 석수인데 선생원 돌 뜨는 곳에 가서 해를 끼치고 이웃집 개에게까지 피해를 입혔으므로, 곤 장 여든 대를 쳤다.

## 2.2 Open Domain Summarization

- Prompt Engineering (in-context-learning)
  - It works!
  - But, inconsistent tone and manner
    - 수산화 알루미늄은 번비를 유발함. (abstract)
    - 우리가 먹는 것을 우리 뭐라고 표현을 할까요. (extract)
  - Hallucination

# 2.2 Open Domain Summarization

## 영화 줄거리 요약기

영화 줄거리를 한 줄로 요약합니다.

줄거리: "미나리는 어디서든 잘 자라" 낯선 미국, 아칸소로 떠난 한국 가족. 가족들에게 뭔가 해내는 걸 보여주고 싶은 아빠 '제이콥'(스티븐 연)은 자신만의 농장을 가꾸기 시작하고 엄마 '모니카'(한예리)도 다시 일자리를 찾는다. 아직 어린 아이들을 위해 '모니카'의 엄마 '순자'(윤여정)가 함께 살기로 하고 가방 가득 고춧가루, 멸치, 한약 그리고 미나리씨를 담은 할머니가 도착한다. 의젓한 큰딸 '앤'(노엘 케이트 조)과 장난꾸러기 막내아들 '데이빗'(앨런 김)은 어느 그랜마같지 않은 할머니가 영-모마땡한데... 함께 있다면, 새로 시작할 수 있다는 희망으로 하루하루 뿌리 내리며 살아가는 어느 가족의 아주 특별한 여정이 시작된다!

한줄요약: 미국 아칸소로 떠난 한국 가족이 새로 시작할 수 있다는 희망으로 하루하루 뿌리 내리며 살아가는 이야기입니다.

###

줄거리: 한때는 WBC 웰터급 동양 챔피언이었지만 지금은 오갈 데 없어진 한물간 전직 복서 '조하'(이병헌). 우연히 17년 만에 헤어진 엄마 '인숙'(윤여정)과 재회하고, 숙식을 해결하기 위해 따라간 집에서 듣지도 보지도 못했던 뜻밖의 동생 '진태'(박정민)와 마주한다. 난생처음 봤는데... 동생이라고?! 라면 끓이기, 게임도 최고로 잘하지만 무엇보다 피아노에 천재적 재능을 지닌 서번트중후군 진태. 조하는 입만 열면 "네~" 타령인 심상치 않은 동생을 보자 한숨부터 나온다. 하지만 캐나다로 가기 위한 경비를 마련하기 전까지만 꼭 참기로 결심한 조하는 결코 만만치 않은 불편한 동거생활을 하기 시작하는데... 살아온 곳도, 잘하는 일도, 좋아하는 것도 다른 두 형제가 만났다!

한줄요약: 전직 복서가 우연히 엄마와 재회하고 피아노에 천재적 재능을 지닌 동생과 함께 생활하게 된 이야기입니다.

###

줄거리: 자살시도가 실패로 끝나 한강의 밤섬에 불시착한 남자. 죽는 것도 쉽지 않자 일단 섬에서 살아보기로 한다. 무인도 야생의 삶도 살아볼 만하다고 느낄 무렵, 여자 김씨가 남긴 익명의 쪽지가 담긴 와인병을 발견하고 그의 삶은 알 수 없는 희망으로 설레기 시작한다. 쪽지를 남긴 여자 김씨는 타인과의 접촉을 극도로 꺼리는 대인기피증 환자였는데, 무인도에서 생존해나가는 남자 김씨를 관찰하며 희망을 갖게 된다.

한줄요약: 한강의 무인도에 불시착해 생존해나가는 남자 김씨와 대인기피증 환자인 여자 김씨가 쪽지를 나누며 서로 희망을 주고받게 되는 이야기.

###

줄거리: 금융, 보험, 물류, 쇼핑, 여행 등 다양한 분야에서 누구나 쉽게 접하고 이용할 수 있습니다. 오랫동안 챗봇을 이용해오신 분이라면, 과거와 지금의 사용 경험은 확연히 다르다는 것을 느끼셨을 거예요. 챗봇에는 어떤 변화가 있었던 것일까요? 챗봇은 사용자를 통해 입력된 내용을 인식해서 적절한 대답을 제공합니다. 과거의 모델은 사람이 만든 수많은 예상 질문과 답변(Q&A) 시나리오로 구성되었고, 미리 준비한 시나리오에 해당하는 질문이 입력되었을 때에만 정확한 답변을 제공할 수 있었어요. 사람의 언어표현은 매우 다양하기 때문에, 모든 질문을 이해하거나 맥락까지 이해하는 것은 불가능했죠.

한줄요약: 인공지능 기술 발전으로 인해 챗봇 서비스에도 많은 변화가 생겼습니다.

In-context-learning의 한계

"요약"에 대한 정의가 다를 수 있음

# 2.2 Open Domain Summarization

## 영화 줄거리 요약기

영화 줄거리를 한 줄로 요약합니다.

줄거리: "미나리는 어디서든 잘 자라" 낯선 미국, 아칸소로 떠난 한국 가족. 가족들에게 뭔가 해내는 걸 보여주고 싶은 아빠 '제이콥'(스티븐 연)은 자신만의 농장을 가꾸기 시작하고 엄마 '모니카'(한예리)도 다시 일자리를 찾는다. 아직 어린 아이들을 위해 '모니카'의 엄마 '순자'(윤여정)가 함께 살기로 하고 가방 가득 고춧가루, 멸치, 한약 그리고 미나리씨를 담은 할머니가 도착한다. 의젓한 큰딸 '앤'(노엘 케이트 조)과 장난꾸러기 막내아들 '데이빗'(앨런 김)은 어느 그랜마같지 않은 할머니가 영-모마땡한데... 함께 있다면, 새로 시작할 수 있다는 희망으로 하루하루 뿌리 내리며 살아가는 어느 가족의 아주 특별한 여정이 시작된다!

한줄요약: 미국 아칸소로 떠난 한국 가족이 새로 시작할 수 있다는 희망으로 하루하루 뿌리 내리며 살아가는 이야기입니다.

###

줄거리: 한때는 WBC 웰터급 동양 챔피언이었지만 지금은 오갈 데 없어진 한물간 전직 복서 '조하'(이병헌). 우연히 17년 만에 헤어진 엄마 '인숙'(윤여정)과 재회하고, 숙식을 해결하기 위해 따라간 집에서 듣지도 보지도 못했던 뜻밖의 동생 '진태'(박정민)와 마주한다. 난생처음 봤는데... 동생이라고?! 라면 끓이기, 게임도 최고로 잘하지만 무엇보다 피아노에 천재적 재능을 지닌 서번트중후군 진태. 조하는 입만 열면 "네~" 타령인 심상치 않은 동생을 보자 한숨부터 나온다. 하지만 캐나다로 가기 위한 경비를 마련하기 전까지만 꼭 참기로 결심한 조하는 결코 만만치 않은 불편한 동거생활을 하기 시작하는데... 살아온 곳도, 잘하는 일도, 좋아하는 것도 다른 두 형제가 만났다!

한줄요약: 전직 복서가 우연히 엄마와 재회하고 피아노에 천재적 재능을 지닌 동생과 함께 생활하게 된 이야기입니다.

###

줄거리: 자살시도가 실패로 끝나 한강의 밤섬에 불시착한 남자. 죽는 것도 쉽지 않자 일단 섬에서 살아보기로 한다. 무인도 야생의 삶도 살아볼 만하다고 느낄 무렵, 여자 김씨가 남긴 익명의 쪽지가 담긴 와인병을 발견하고 그의 삶은 알 수 없는 희망으로 설레기 시작한다. 쪽지를 남긴 여자 김씨는 타인과의 접촉을 극도로 꺼리는 대인기피증 환자였는데, 무인도에서 생존해나가는 남자 김씨를 관찰하며 희망을 갖게 된다.

한줄요약: 한강의 무인도에 불시착해 생존해나가는 남자 김씨와 대인기피증 환자인 여자 김씨가 쪽지를 나누며 서로 희망을 주고받게 되는 이야기.

###

줄거리: 금융, 보험, 물류, 쇼핑, 여행 등 다양한 분야에서 누구나 쉽게 접하고 이용할 수 있습니다. 오랫동안 챗봇을 이용해오신 분이라면, 과거와 지금의 사용 경험은 확연히 다르다는 것을 느끼셨을 거예요. 챗봇에는 어떤 변화가 있었던 것일까요? 챗봇은 사용자를 통해 입력된 내용을 인식해서 적절한 대답을 제공합니다. 과거의 모델은 사람이 만든 수많은 예상 질문과 답변(Q&A) 시나리오로 구성되었고, 미리 준비한 시나리오에 해당하는 질문이 입력되었을 때에만 정확한 답변을 제공할 수 있었어요. 사람의 언어표현은 매우 다양하기 때문에, 모든 질문을 이해하거나 맥락까지 이해하는 것은 불가능했죠.

한줄요약: 인공지능 기술 발전으로 인해 챗봇 서비스에도 많은 변화가 생겼습니다.

튜닝을 통해 조금 더 우리가 원하는 방향으로 요약을 합시다!

- 챗봇은 사용자를 통해 입력된 내용을 인식해서 적절한 대답을 제공함
- 과거의 모델은 사람이 만든 수많은 예상 질문과 답변(Q&A) 시나리오로 구성되었음
- 미리 준비한 시나리오에 해당하는 질문이 입력되었을 때에만 정확한 답변을 제공할 수 있었음

compression ratio: 41.6 (138 / 332) %  
ESC to close  
press 'p' to copy  
double-click to detail :

1 금융, 보험, 물류, 쇼핑, 여행 등 다양한 분야에서 누구나 쉽게 접하고 이용할 수 있습니다. 오랫동안 챗봇을 이용해오신 분이라면, 과거와 지금의 사용 경험은 확연히 다르다는 것을 느끼셨을 거예요. 챗봇에는 어떤 변화가 있었던 것일까요?

1 챗봇은 사용자를 통해 입력된 내용을 인식해서 적절한 대답을 제공합니다. 과거의 모델은 사람이 만든 수많은 예상 질문과 답변(Q&A) 시나리오로 구성되었고, 미리 준비한 시나리오에 해당하는 질문이 입력되었을 때에만 정확한 답변을 제공할 수 있었어요. 사람의 언어표현은 매우 다양하기 때문에, 모든 질문을 이해하거나 맥락까지 이해하는 것은 불가능했죠.

## 2.2 Open Domain Summarization

- Tune with Suffix(template / instruction: 요약해봐!)  
"협회의 소의 이익은 행정소송에 요구되는 소송요건의 하나이기 때문에 처분의 취소를 구할 소의 이익이 없는 경우, 소송요건의 흠결로 소는 각하됩니다"
- w/o suffix(\n) => "질문:소이익과 관련된 소송은? 답변:행정소송"
- w/ suffix => "소의 이익이 없는 경우 각하됨" : Instruct(태스크 수행) 능력

# 2.3 Dataset

## AI HUB

Training	
book_text.jsonl:	160,002
dialogue_text.jsonl:	279,994
document_text.jsonl:	388,139
journal_text.jsonl:	679,160
Validation	
book_text.jsonl:	19,999
dialogue_text.jsonl:	35,003
document_text.jsonl:	47,142
journal_text.jsonl:	109,122

## 도메인에 따라 입력과 출력이 서로 상이함

0	dialogue	P01: 엄마 담배는 몸에 안좋지? P02: 글썄 P01: 절대 피우면 안되겠네 P02: 피우지마 P01: 알겠어 엄마도 피우지마 P02: 난 당연하얀피우지 P01: 우리가족 전부 담배는 멀리하자 P02: 그래 P01: 히히 잘됐다	담배는 몸에 안 좋아서 피우지 말고 가족 모두 멀리하자고 이야기한다.
1	document	수출입업체인 원고가 의류제품을 제조·수출함에 있어 같은 그룹내 종합무역상사인 소외 회사의 직수출실적을 지원하기 위하여 동 회사와 수출대행계약을 체결하고 동 회사를 수익자로 하여 신용장을 개설한 후 동 회사명의로 제품을 수출하면서 그 수출물품에 관하여 소외 회사는 원고로부터 세금계산서를 교부받지 않았을 뿐 아니라, 부가가치세신고시 이를 수출로 신고하지 않았고, 외화수입금도 원신용장상 수익자인 소외 회사에 입금되나 당해은행에서 매수 즉시, 원화로 원고의 당좌계정에 입금되고 소외 회사는 대행수수료만 받아왔다면 소외 회사는 수출에 실질적 참여한 바 없이 원고의 수출대행을 한 자로 명목상의 수출자에 불과할 뿐 원고가 이를 실질적으로 수출하여 그 대금에 해당하는 외화를 수입하였다 할 것이므로, 원고는 구 법인세법시행령(1983.12.31 대통령령 제10119호로 개정되기 전의 령) 제17조 제1항 제1호 소정의 수출사업을 영위하는 법인에 해당되어 그 수출금액에 대응한 해외시장개척준비금을 설정할 수 있다.	수출입업체인 원고가 의류제품을 제조·수출함에 있어 소외 회사의 직수출실적을 지원하기 위하여 동 회사와 수출대행계약을 체결하고, 소외 회사는 수출에 실질 참여한 바 없이 명목상의 수출자에 불과할 뿐 원고가 이를 실질적으로 수출하여 그 대금에 해당하는 외화를 수입하였다 할 것이므로, 원고는 구 법인세법시행령(1983.12.31 대통령령 제10119호로 개정되기 전의 령) 제17조 제1항 제1호 소정의 수출사업을 영위하는 법인에 해당되어 그 수출금액에 대응한 해외시장개척준비금을 설정할 수 있다.
0	journal	본 연구는 경제협력개발기구(Organization for Economic Cooperation and Development, OECD)가 실시한 교수·학습 국제조사(Teaching and Learning International Survey 2013, TALIS) 중 교직 선택에 대한 교사 후회에 관한 조사를 연구의 출발점으로 삼았다. 한국의 경우 중학교 교사가 참여한 이 조사에서 응답 교사의 20.1%가 교사가 된 것을 후회한다는 결과가 나오자 사회적 관심이 집중되었다. 교직 선택을 후회하는 교사 비율은 조사대상국 중 최고로 높은 수치로 이는 곧 교사의 낮은 사기와 만족도를 표현하는 것으로 해석되었고 교원 정책에 대한 비판의 근거로 제시되었다. 이 연구는 교사가 경험하는 후회의 감정은 매우 복잡한 현상이고 다양한 변인이 작용하는 심리적메커니즘을 가지고 있다는 선행연구에 기초하여 이러한 단순한 해석이 갖는 위험을 지적하고 탐색적인 차원에서 대안적인 해석의 가능성을 제시하였다. 기회(대안)와 후회, 정상(norm)과 후회, 후회의 기능이라는 3가지 영역을 검토하면서, 교사 후회 현상은 다양한 변인의 영향을 받을 수 있음을 지적하고 관련 정책적 시사점도 아울러 모색하였다. 특히, 교사 후회는 교직 선택동기에 관한 연구, 교직을 선택한 시기에 관한 조사, 교직 선택이 당연한 집단이 가진 특성, 후회를 감소시키기 위한 교사의 교직 행동 등과 관련지어 새롭게 해석될 수 있음을 보여주었다. 앞으로 후회 감정에 관련된 심리적 비교에서 고려되는 요소, 행위/비행위 차이, 후회에 관련된 심리적 처리의 휴리스틱과 편향 등 후속 연구가 적절히 이루어질 때 교사 후회 현상에 대한 제대로 된 해석이 가능하고, 교사 후회 현상을 바탕으로 한 정책 개발 및 평가도 타당하게 이루어질 수 있을 것이다.	본 논문은 경제협력개발기구가 실시한 교수·학습 국제조사 가운데 교직 선택에 대한 교사 후회에 대한 조사를 연구의 출발점으로 하였다. 한국의 경우 중학교 교사가 참여한 이 조사에서 응답 교사의 20.1%가 교사가 된 것을 후회한다는 결과가 나와 사회적 관심을 받았다.

## 2.3 Dataset

AI HUB (Labeled): 약 25만건

0	dialogue	<p>P01: 엄마 담배는 몸에 안좋지? P02: 글치 P01: 절대 피우면 안되겠네 P02: 피우지마 P01: 알겠어 엄마도 피우지마 P02: 난 당연히안피우지 P01: 우리가족 전부 담배는 멀리하자 P02: 그래 P01: 히히 잘됐다</p>	<p>담배는 몸에 안 좋아서 피우지 말고 가족 모두 멀리하자고 이야기한다.</p>
---	----------	--	---

## 2.3 Dataset

Note (Un-labeled): 약 3천건 => Annotation

- AI HUB

방송 / 회의 / 강의 등

이와 관련된 대책은 혹시 있는지 궁금한데요?

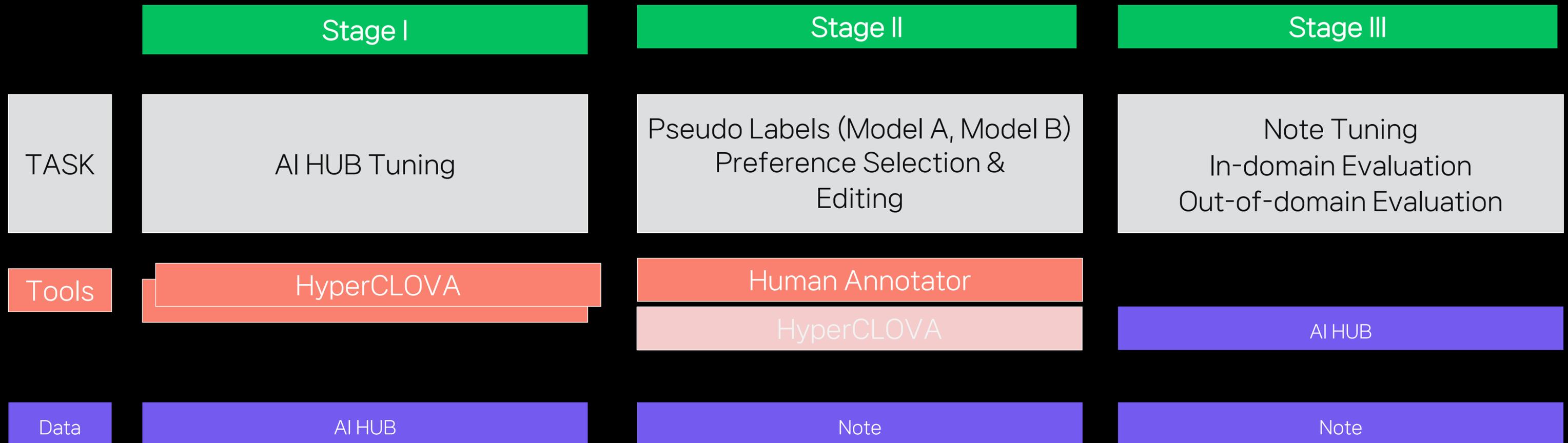
예 그래서 인제 지금 메르스 사태도 그렇고요 또 이제 태풍이라든가 이렇게 좀  
위급한 상황들이 재난적 상황에 있는 경우에 대비해서  
학교에 학생들이 안 나오더라도 돌봄을 하는 시스템을 지금 그래도 갖추어져 있습니다 유치원하고 예 초등학교에  
돌봄 교실은 평상시에 돌봄 교실을 운영하지 않는 학생까지도 이렇게  
받아들이는 방식으로 저희가 시스템이 돼 있으니까 너무 걱정하지 않으셔도 되고요 직장맘이라고  
보통 말씀을 하시지 않습니까 그리고 한 가지 더 이 자리를 빌어서  
일단은 입학할 3월 9일로 연기된 입학식도 안 한 상태지 않습니까 근데 그 학생도 학생입니다 그렇기 때문에  
그 학생도 돌봄에 바로 참여할 수가 있습니다 돌봄 교실에.  
그러니까 일 학년 학생들도 유치원이나 초등학교에 일 학년 들어가게 되어 있는 학생도 돌봄 교실을 바로 갈 수 있다.  
이렇게 좀 알아주시면 좋을 것 같고요 이제 저희가 학교에서 하는 돌봄이 있고 또  
서울시에서 하는 우리 동네 키움 센터 같은 데가 또 있습니다 한 100여 군데가 있고요.  
그다음에 또 또 여가부에서 하는 긴급 돌봄 서비스 같은 것도 있습니다 그래서  
좀 고런 것들을 이용하면  
그래도 뭐~ 이렇게 촘촘하게 생각보다는 시스템들이 갖춰져 있으니까  
활용을 하시면 좋겠습니다.

유치원이나 초등학교 1학년에 들어가는 학생도 돌봄 교실을 갈 수 있음  
돌봄이나 우리 동네 키움 센터나 긴급 돌봄 서비스를 활용하면 좋음

- 그 외 사내 회의록 (클로바 노트 전사)

# 2.4 Tuning Process

(1) AI HUB Tuning => (2) Annotation => (3) Note Service Tuning



## 2.4 Tuning Process

### 어려운 점들

#### 학습과 추론의 불일치

- Training: 세그먼트되어 있음
  - Label: 1~3문장
  - Input: (1) 평균 200 토큰 길이  
(2) 실제 사용자들의 노트가 아님
- Inference: Note(세그먼트 되지 않음)
  - Input: (1) 길이가 다양함  
(2) STT 결과

#### 너무 짧거나 무성의한 Label

- "휴대폰"에 대해 설명하고 있음
- => 패턴매칭으로 제거

# 2.5 Training

일반적으로

Pretrain > DAPT > Task-Adaptive Fine-tuning

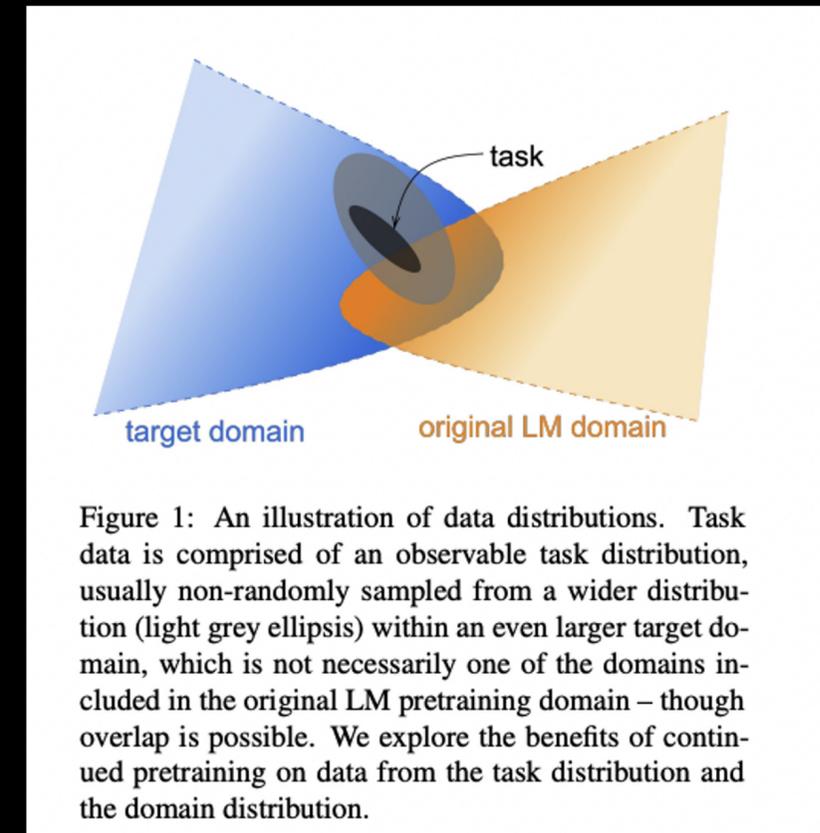
\* DAPT(Domain-Adaptive Pre-train)

=> Two-Stage Lora-Tuning (빠르고 효율적)

(1) Pretrain (HyperCLOVA)

(2) Domain-Adaptive Lora-tuning (AI-HUB Dialogue DATA): 완성형 1문장 요약

(3) Task-Adaptive Lora-tuning (Note DATA): 개조식 1~3문장 요약



S. Gururangan et al. 2020. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks

## 2.5 Training

Stage I (250,000 examples): 키워드 생성과 요약을 동시에

AI-HUB:

입력: {SOURCE} + \n키워드를 잘 생성해주세요.\n

출력: {KEYWORDS}\n키워드를 바탕으로 요약해주세요.\n {SUMMARY}

\* 키워드는 항상 본문에 등장하는 조건으로 Hallucination 억제

Stage II (3,500 examples): 개조식 요약만 수행

NOTE:

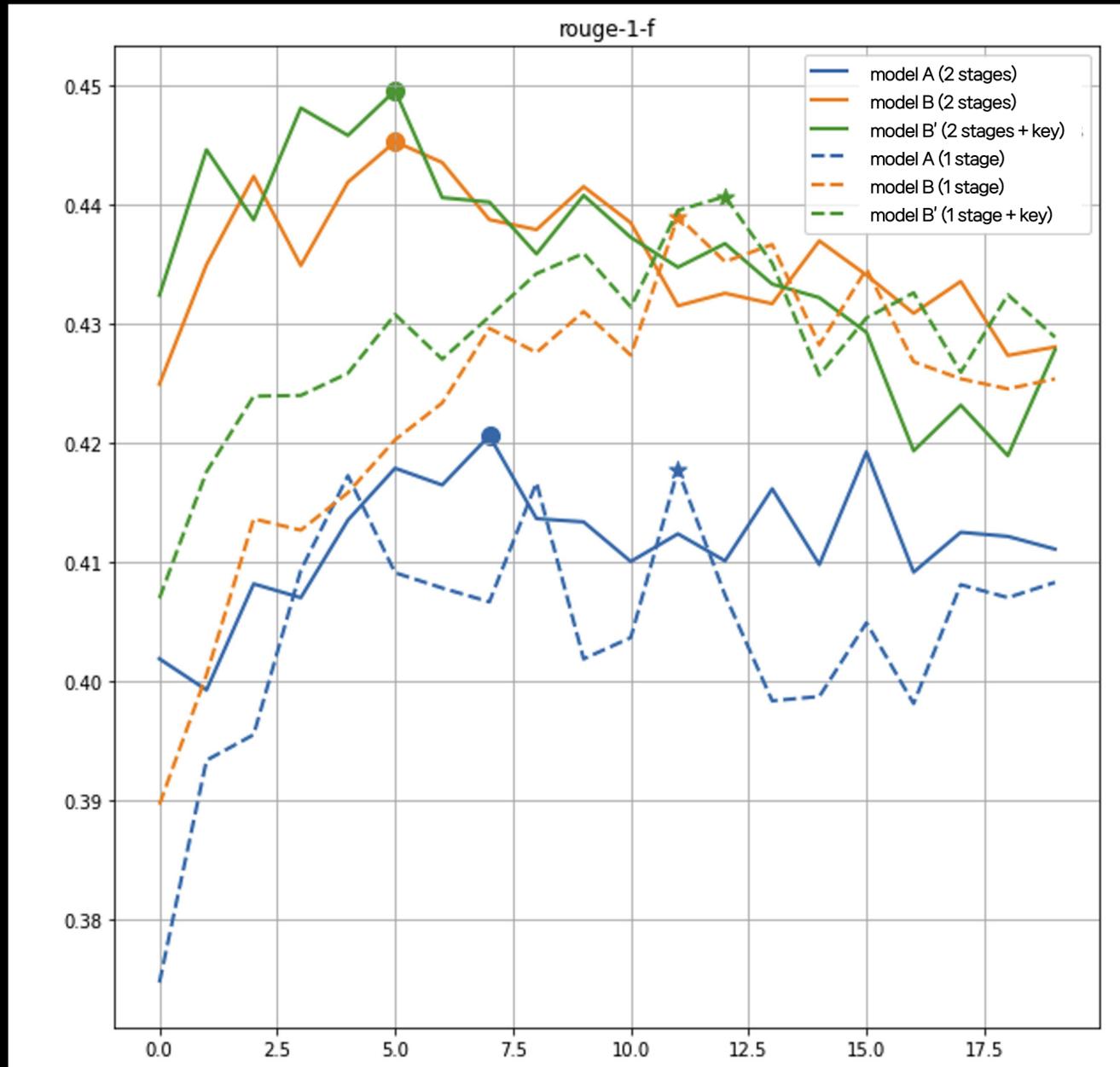
입력: {SOURCE} + \n본문 내용을 사실 그대로 개조식으로 요약해주세요.\n

출력: {SUMMARY}

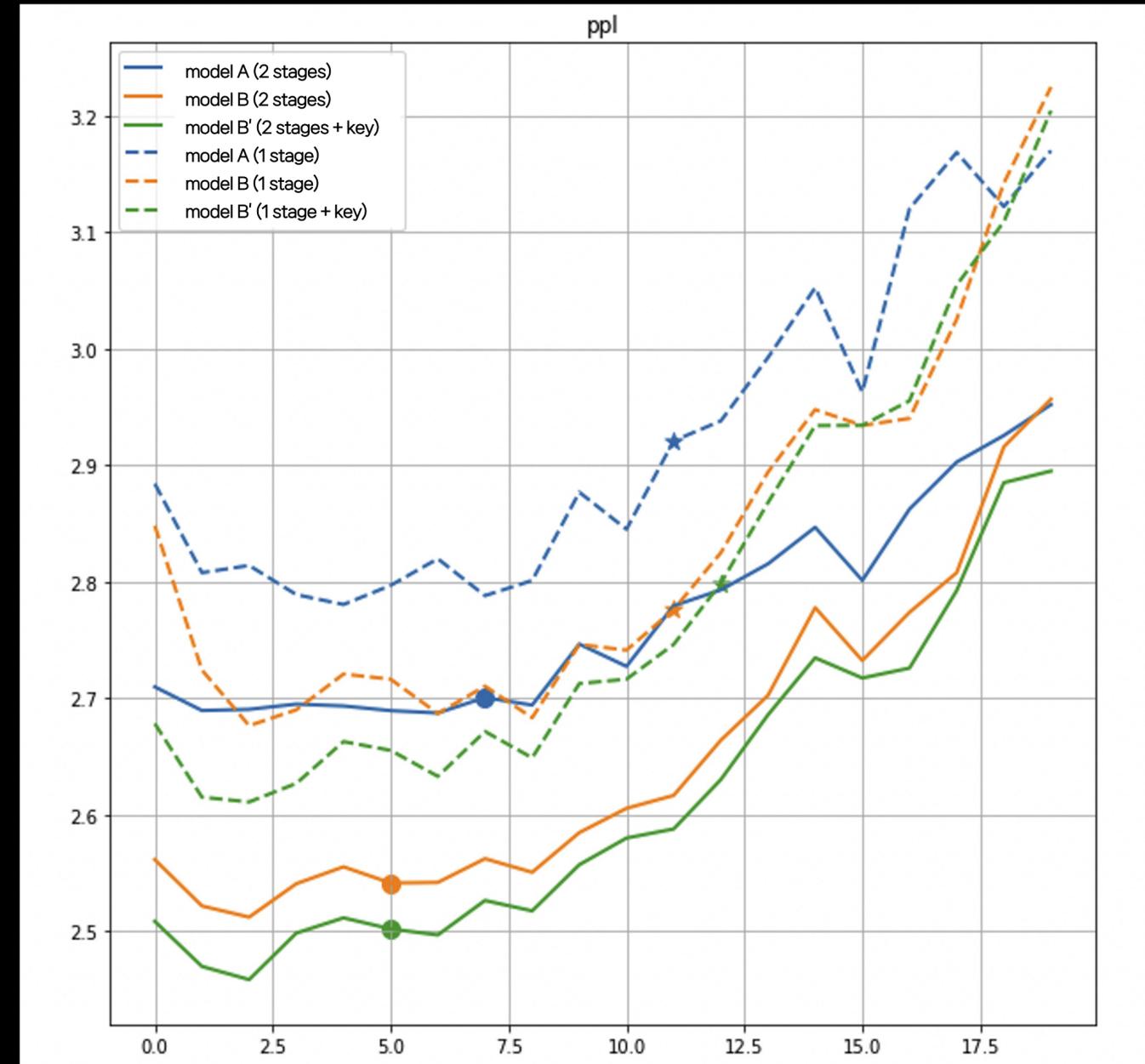
\* 본 프롬프트는 예시입니다.

# 2.6 Evaluation

Stage & Model Rouge-1-F 비교



Stage & Model PPL 비교



# 2.6 Evaluation

병이 서울 미래유산 지정

- 서울 미라고 함

ESC to close  
compression ratio: 250.0 (10 / 4) %press 'p' to copy



2022. 0. 3 22

20  
20

열차, 꺼병이 서울 미래유산 지정

- 서울의 미래에 대해 이야기하고 있음

ESC to close  
compression ratio: 420.0 (21 / 5) %press 'p' to copy



서울 미래유산 지정

- 서울 미래유에 대해 이야기하고 있음

ESC to close  
compression ratio: 350.0 (21 / 6) %press 'p' to copy



서울 미래유산 지정

- 서울 미래유산에 대해 이야기하고 있음

ESC to close  
compression ratio: 314.3 (22 / 7) % press 'p' to copy



<https://namu.wiki> > 서울 대전 대구 부산

서울 대전 대구 부산 - 나무위키:대문

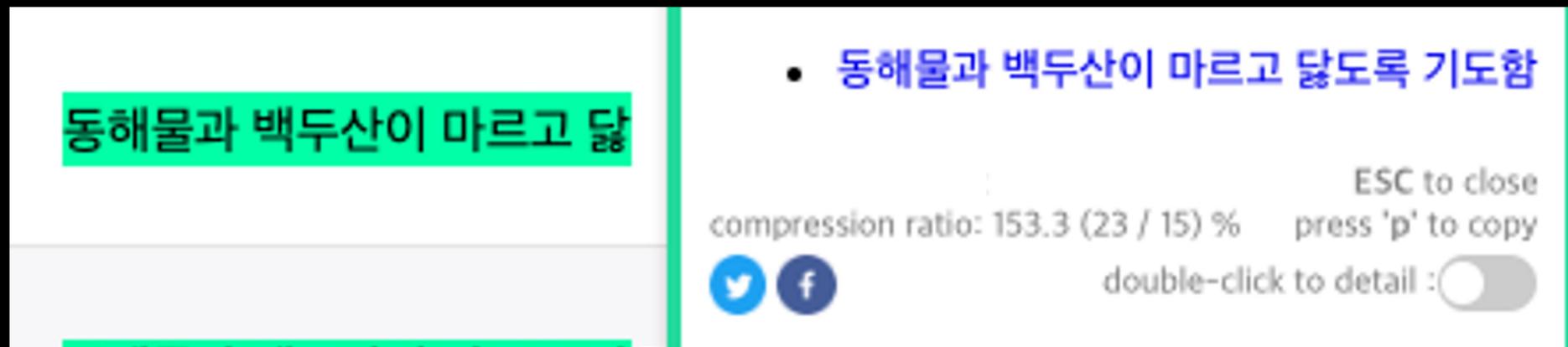
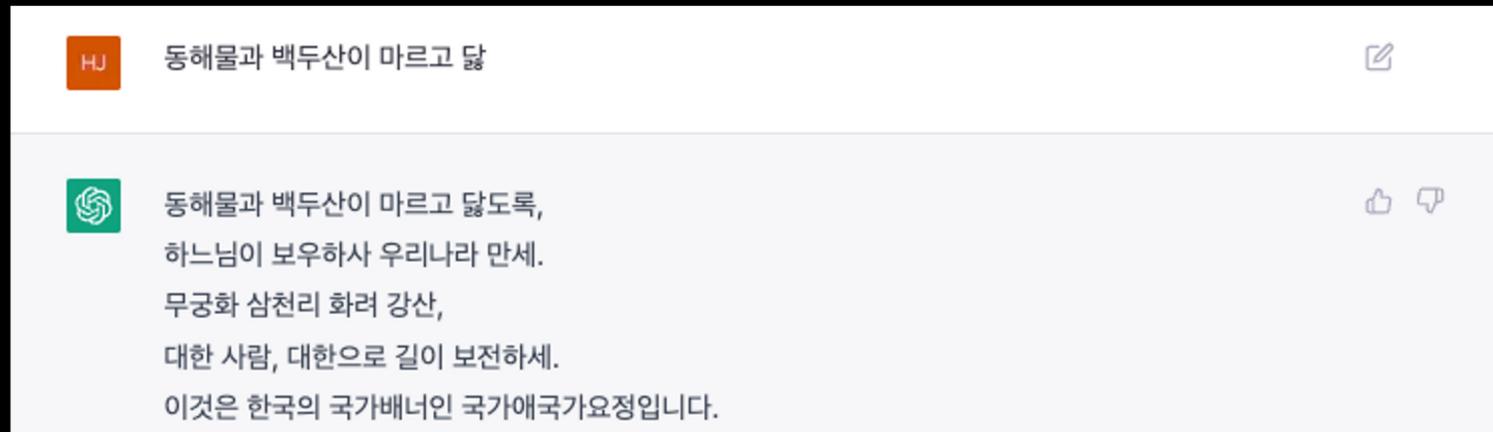
- 서울, 대전, 대구, 부산을 말함

ESC to close  
compression ratio: 181.8 (20 / 11) %press 'p' to copy



# 2.6 Evaluation

그러나, 여전히 자동완성(?)의 위험성이 존재함



## 2.7 Post-process

### 후처리 안전장치(Post-Process)

(1) 1문장(N 글자수) 이상

(2) 중복 문장 제거 (rouge-1 precision 기준)

- 오늘은 비가 올 것 같아요

- 오늘 오후부터 비가 올 것 같아요 (제거)

## 2.8 summaryGPT

영하: 아

## 2.8 summaryGPT

### *Summarization:*

- Text summarization is the process of **distilling the most important information** from a text to produce an abridged version for a **particular task and user**. by Berry et al. (1995)

## 2.8 summaryGPT

"수진은 데이트 준비를 위해 집에서 빨리 일어났다. 강남에서 남자친구인 철호를 만나 이것저것 구경도 하고 영화도 보고 저녁도 함께 먹었다. 음식을 얼굴에 살짝 묻히면서 맛있게 먹는 수진을 보며 철호는 음식 먹는 것도 예쁘게 먹는다고 웃으며 칭찬을 한다."

(1) 장소를 포함해서 요약해줘!

집 -> 강남

(2) 시간순으로 요약해줘!

기상 -> 만남

(3) 영어로 요약해줘!

# 2.8 summaryGPT

재료: 밀가루, 고구마, 버섯  
요리: 고구마 프리타타  
레시피: 1. 먼저 뜨겁게 달군 버터에 양파를 볶아주세요.  
2. 달걀 3개 풀어주신 후, 우유도 같이 넣어줄게요  
3. 고구마를 얇게 썰어서 팬에 올리고 은은하고 그윽한 불에서 찬찬히 익혀주세요.  
4. 어느정도 단단하게 바닥면이 고정이 되면, 치즈까지 솔솔 뿌려요.

INFERENCE 방식 (SUFFIX(or INSTRUCTION))가 모두 상이함)

## 1. 학습에 사용된 요약 패턴

- 재료는 밀가루, 고구마, 버섯임
- 고구마 프리타타 레시피를 알려주고 있음

## 2. 영문 번역 (ZEROSHOT)

- Ingredients: flour, sweet potato, mushroom Recipe: 1. First, fry onion in the butter with hot heat. 2. Then, mix eggs and milk. 3. Slice sweet potato and cook it in the pan s  
When the bottom is firm enough, sprinkle cheese.

## 3. 키워드 추출 (ZEROSHOT)

- 밀가루|고구마|버섯|고구마 프리타타

## 4. 간략히 설명 (ZEROSHOT)

- 고구마 프리타타 레시피를 알려주고 있음

## 5. 자세히 설명 (ZEROSHOT)

- 재료: 밀가루, 고구마, 버섯
- 요리: 고구마 프리타타
- 레시피: 1. 먼저 뜨겁게 달군 버터에 양파를 볶아주세요.  
2. 달걀 3개 풀어주신 후, 우유도 같이 넣어줄게요  
3. 고구마를 얇게 썰어서 팬에 올리고 은은하고 그윽한 불에서 찬찬히 익혀주세요.

## 6. MRC (ZEROSHOT)

- 6.1. 요리 이름이 뭐야?  
A. 고구마 프리타타
- 6.2. 고구마를 구울 때 불 세기는?  
A. 은은하게
- 6.3. 양파를 영어로하면 뭐야?  
A. onion

# 2.8 summaryGPT

재료: 밀가루, 고구마, 버섯  
요리: 고구마 프리타타  
레시피: 1. 먼저 뜨겁게 달군 버터에 양파를 볶주세요.  
2. 달걀 3개 풀어주신 후,우유도 같이 넣어줄게요  
3. 고구마를 얇게 썰어서 팬에 올리고 은은하고 그윽한 불에서 찬찬히 익혀주세요.  
4. 어느정도 단단하게 바닥면이 고정이 되면, 치즈까지 솔솔 뿌려요.

## 입력 쿼리

언제 치즈를 뿌리나요?

## 요약 결과

- | 고구마가 어느 정도 단단하게 바닥면이 고정이 되면 치즈를 뿌려주세요

재료: 밀가루, 고구마, 버섯  
요리: 고구마 프리타타  
레시피: 1. 먼저 뜨겁게 달군 버터에 양파를 볶주세요.  
2. 달걀 3개 풀어주신 후,우유도 같이 넣어줄게요  
3. 고구마를 얇게 썰어서 팬에 올리고 은은하고 그윽한 불에서 찬찬히 익혀주세요.  
4. 어느정도 단단하게 바닥면이 고정이 되면, 치즈까지 솔솔 뿌려요.

## 입력 쿼리

프리타타와 가장 유사한 한국 음식은 무엇이 있을까요?

## 요약 결과

- 김치전



# 2.9 Future Work

이제 요약문이 더 풍성해 집니다.

## 요약 결과

- 씨케이 프레임워크 개발을 맡고 있는 신성진이라고 함
- 씨케이 프레임워크 챗봇 프레임워크를 가지고 좀 더 쉽게 개발을 할 수 있는지 설명을 드리겠음
- 챗봇 프레임워크 ck에 대해 설명하고 있음
- 한국어 및 대화형의 데이터에 최적화된 대형 모델을 설계함
- 학습을 잘 하는 과정으로 추가를 해서 학습을 잘 하는 과정으로 넣음
- 학습을 하면서 추출하는 fite는 다르다고 생각함
- 클로바 챗봇 익스텐션 가이드를 만든 모델을 활용하는 방법을 설명하고 있음
- 채널 연동을 통해 쉽게 모델을 만들고 연동을 할 수 있음
- 챗봇에 대한 기본적인 발화에 대해서는 공수를 줄일 수 있음
- 챗봇을 만들 때는 양과 품질에 대해서 신경을 써야 함
- 챗봇을 만들 때는 초기 데이터 생성이 중요함
- 데이터가 얼마나 중요한지에 대해서 이야기를 들음
- 데이터를 어떻게 활용하는지에 대해서 이야기를 들음
- 데이터 학습에 대한 팀으로 조사의 활용 여부를 이야기함
- 원형 복원이라는 테크닉을 활용해서 원형 복원으로 하나로 통일을 함
- 원형 복원으로 전처리를 해서 활용하면 학습할 때도 그리고 실제로 실험을 할 때도 점심 뭐 먹다라고 치환이 돼서 넣었기 때문에 최소화가 됨
- 실제 고객센터 같은 경우에는 기존 대비 품질이 5% 이상이 증가함
- 챗봇을 상용화하기 위해서는 다양하게 넣자는 이야기를 하고 있음
- 챗봇으로 해서 사람들이 쓰기에 편안가가 중요함
- 데이터는 양뿐만 아니라 질도 중요함
- 딥러닝은 머신러닝의 부분 중에 하나임
- 개발자와 사용자 모두에게 좋은 틀이 될 수 있음
- 좋은 피처를 찾기 위해 특징을 찾기 위해 검토를 해야 됨
- 챗봇을 만들 때 데이터의 양이 충분하지 않아도 딥러닝 알고리즘이 돌아갈 수 있는지에 대해 묻고 있음

## 요약 결과

### 씨케이 프레임워크 챗봇 프레임워크

- 씨케이 프레임워크 챗봇 프레임워크를 가지고 좀 더 쉽게 개발을 할 수 있는지 챗봇을 가지고 개발을 최적화하려면 어떻게 하는지에 대해서 설명을 드리겠음
- 자연어 처리는 언어라는 것은 컴퓨터로 표현하기가 어려운 요소 중에 하나임
- 텍스트의 경우에는 단어의 고유의 의미가 없음

### 챗봇 프레임워크 ck

### 한국어 및 대화형 데이터에 최적화된 대형 모델 설계

### 클로바 챗봇 익스텐션 가이드

### 챗봇 모델 빌드 방법

### 챗봇 개발 최적화 방안

### 데이터 코드의 업무

### 조사를 같이 학습시켰을 때의 한계

### 데이터가 적을 때는 프레임워크를 만들어서 템플릿이나 이런 걸 제공

### 챗봇을 데이터로 론칭하기 위한 조건

### 딥러닝의 장점

### 네이버, 데이터 양 부족해도 딥러닝 알고리즘 돌아가는 기술 개발

# 2.9 Future Work

1784 is the official name of NAVER's second headquarters. It was named after the building address: 178-4 Jeongja-dong, Bundang-gu. Coincidentally, 1784 also happens to be the year recorded as the start of the first Industrial Revolution.

This place is especially meaningful to NAVER LABS. It contains the technologies of NAVER LABS along with the many efforts of our researchers and engineers. Among these, we've summarized four features that hold special significance for us. The following features of 1784 are both incredibly unique and closely correlated to each other.

### World's First Robot-Friendly Building

1784 is the world's first robot-friendly building. What does this imply? We made a very important decision when designing 1784. Because robot services will become popularized sometime soon, which will be accompanied by the need for infrastructure changes, we wanted to present a completely new standard rather than simply adding robots to office environments we are so used to.

A great example is the world's first robot-only elevator, ROBOPORT. The vertical movement of robots in high-rise buildings is still a challenge. We wanted to introduce innovations in the facility itself in addition to creating robot algorithms that utilize human elevators. ROBOPORT consists of a loop structure that runs across all floors, from B2 to the roof, maximizing the vertical movement speed and efficiency of 1784 robots.

The same is true for local 5G networks. NAVER's robots have the unique characteristic of being cloud-based, and our main goal here was to minimize the communication latency between the robot and the cloud. This is why we installed local 5G (local 5G network) in 1784. We believe that robots can maximize the use of 5G, and are working on proving this in 1784.

### Largest Tech-Convergence Since NAVER's Establishment

The construction of a building involves the process of combining different materials such as wood, stone, soil, steel, and glass. This is why we often use the word "construction" to describe the creation of a building. But the more appropriate word for 1784 would be "convergence," as in this building is created from NAVER's technologies. The building itself is a "tech convergence platform" that combines the capabilities of all NAVER technology organizations.

In fact, convergence is the key to NAVER's R&D strategy. While the competition for global technological dominance is fierce, there is no service called 'AI.' There are no platforms named 'robot,' nor any company named 'autonomous driving.' We now know that the integration of these technologies is essential to provide truly useful services in our daily lives. Not to mention the difficulty of achieving this as well.

In 1784, experts from the largest number of fields since the foundation of NAVER have come together as one team, to work on this goal of convergence. The experience was really incredible. We were able to accumulate know-how and standards that could not be found from anywhere else.

- 네이버의 두 번째 본사
  - 네이버의 두 번째 본사의 공식 이름은 1784입니다
  - 1784는 1차 산업혁명이 시작된 해와 일치함
  - 1784는 세계 최초로 로봇 친화적인 건물임

---

- 네이버의 로봇 전용 엘리베이터
  - 세계 최초의 로봇 전용 엘리베이터인 로보포트를 소개함
  - 로봇과 클라우드 사이의 통신 지연을 최소화하기 위해 네이버는 로컬 5G 네트워크를 설치함

---

- 네이버의 기술 융합 플랫폼
  - 네이버의 설립 이후 가장 큰 기술 융합 건물을 건설함
  - 네이버의 기술 조직이 가진 역량을 결합한 기술 융합 플랫폼을 만들
  - 이 건물은 네이버의 기술력으로 만들어짐

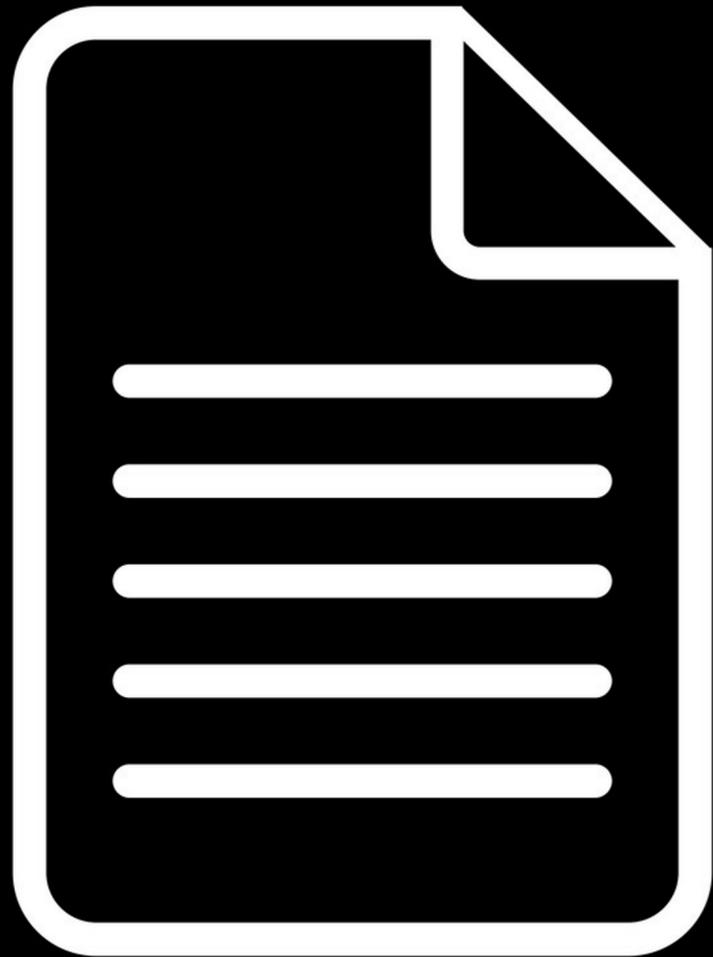
compression ratio: 8.7 (260 / 3000) %



ESC to close  
press 'p' to copy  
double-click to detail :

# 3. Topic Segmentation

## 3.1 Long Document



### 입력 문장 길이 제한

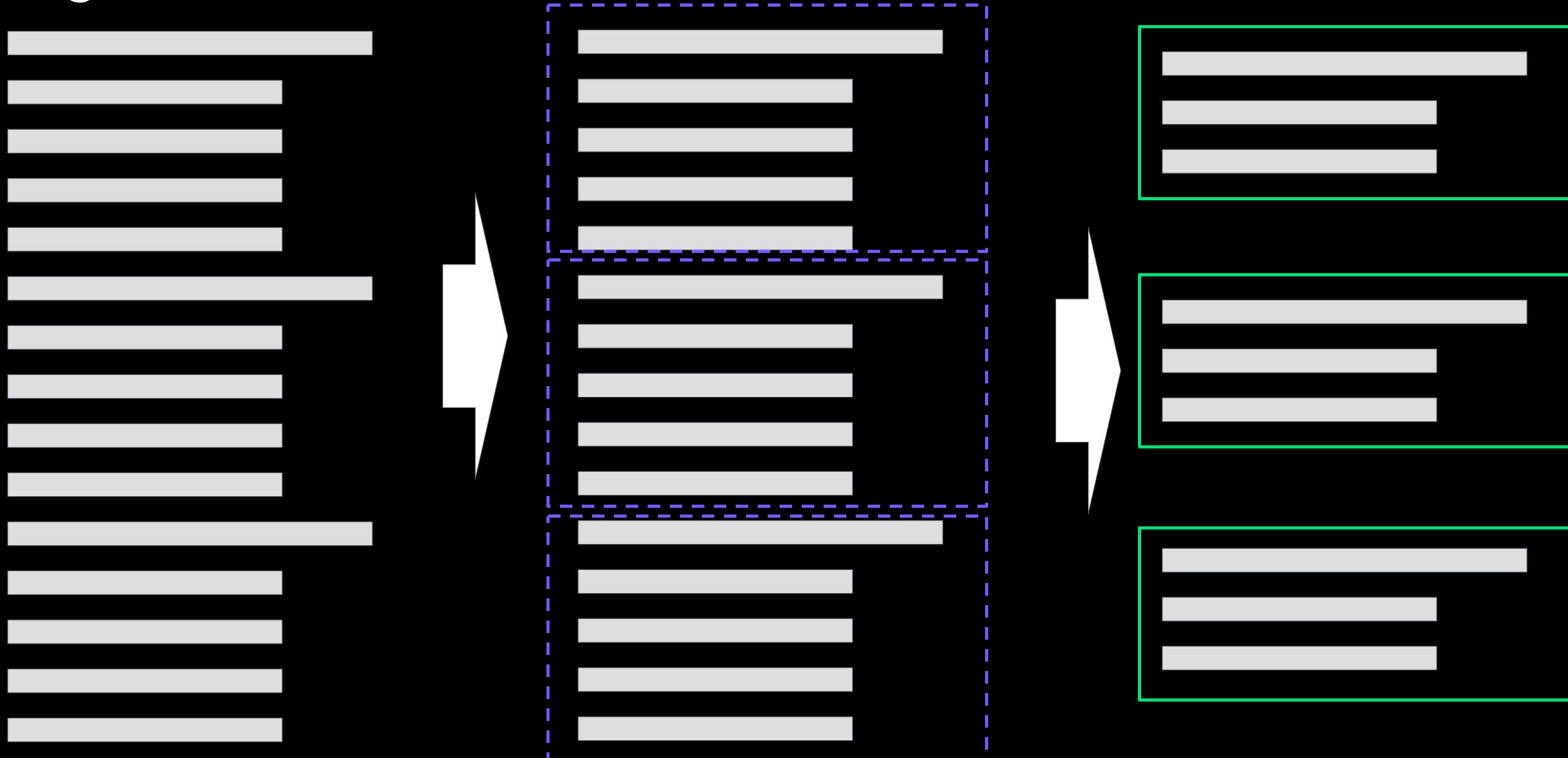
- 입력 토큰 길이를 넘으면 전체 문서 내용을 입력하지 못함

### 문서에 대한 구조적 이해

- 문서에 대한 전체적인 주제를 파악하여 핵심 부분만 요약

# 3.1 Long Document

Segmentation & Summarization 을 통해 문서의 내용을 구조적으로 이해



## 3.2 Literature Review

### Embedding Vector 적용

- 문장에 대한 Embedding Vector 추출 → Cosine Distance 기반 Segmentation 수행

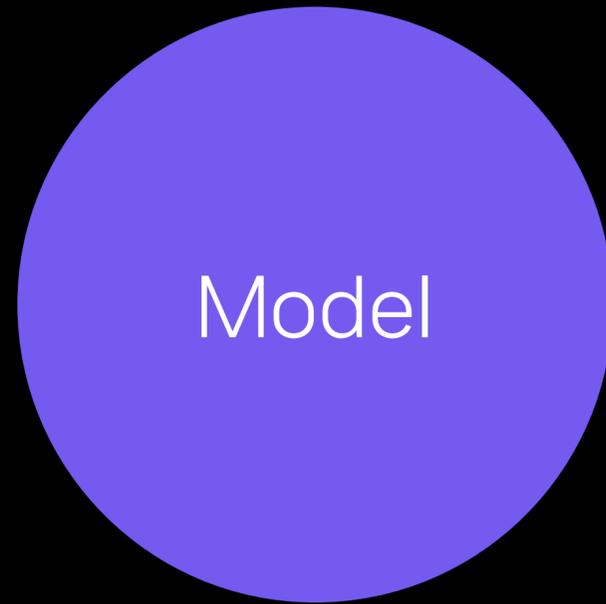
### Classifier 적용

- Segmentation Point를 Classifier를 통해 학습 → Score 기반으로 Segmentation 수행

# 3.3 Topic Segmentation



학습 데이터 정의



- Embedding
- BERT
- GPT



- Pk
- WinDiff

## 3.4 Dataset

Segmentation의 기준 정의

- Dialogue Data : 대화 intent가 변경되는 시점
- Document : 단락이 변경되는 시점

# 3.4 Dataset

	Data	Train		Valid	
		# of dialogues	# of sentences	# of dialogues	# of sentences
Dialogue Data	AI-HUB 한국어 대화	5,855	90,413	1,044	15,890
	AI-HUB 민원(콜센터) 질의-응답 데이터	17,191	433,021	1,276	34,375
Document Data	Wiki Data	5,000	156,088	596	30,654

# 3.4 Dataset

## Data Sampling

### Dial 1.

Utter 1.	Act 1.
Utter 2.	Act 2.
Utter 3.	Act 2.
Utter 4.	Act 3.
Utter 5.	Act 4.

### Dial 2.

Utter 1.	Act 1.
Utter 2.	Act 1.
Utter 3.	Act 2.
Utter 4.	Act 2.
Utter 5.	Act 3.



### Positive Data

Utter 2.	Utter 3.
----------	----------

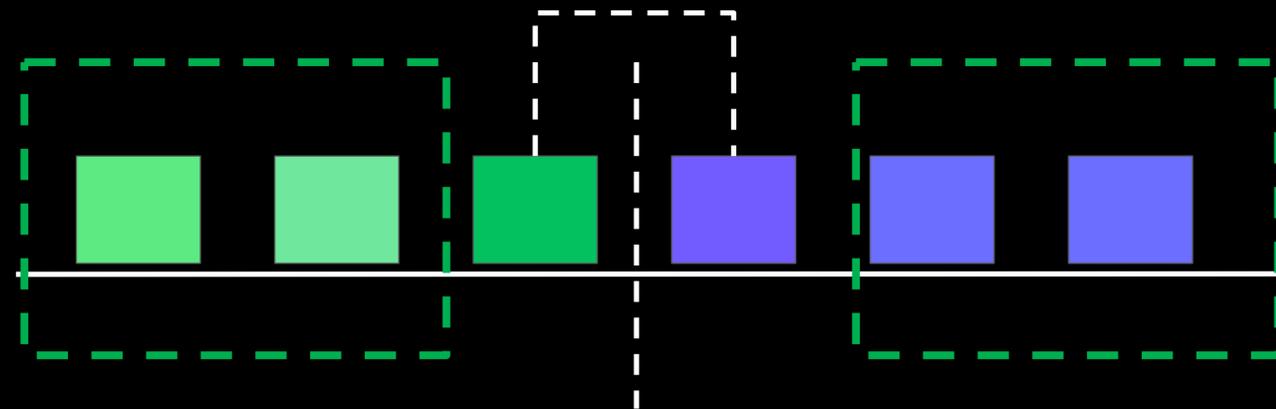
### Negative Data

Utter 2.	Utter 4.
Utter 2.	Utter 3.

# 3.4 Dataset

## Input Format

- 문장 Pair를 입력할 때 앞, 뒤 문장을 추가로 입력하여서 context를 이해하여 Segmentation을 수행하도록 함



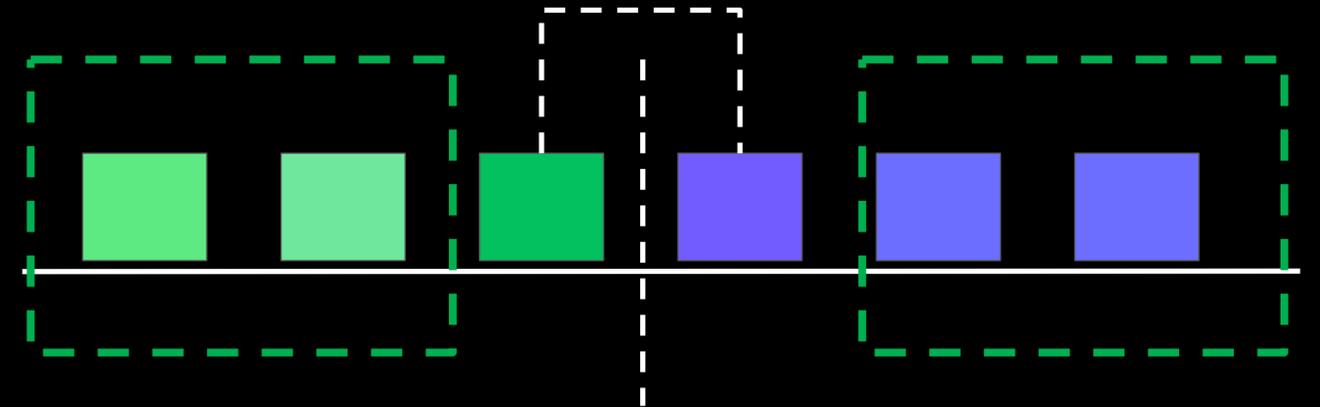
windows size = 2

# 3.4 Dataset

## Input Prompt

- Task의 의미를 이해하기 위한 형태로 학습 데이터 형태를 정의함

문장1: 노트는 어떻게 생성할 수 있나요?  
문장2: 두 가지 방법이 있습니다.



<이전내용> ...

<문장1> 노트는 어떻게 생성할 수 있나요?

<문장2> 두 가지 방법이 있습니다.

<다음내용> ...

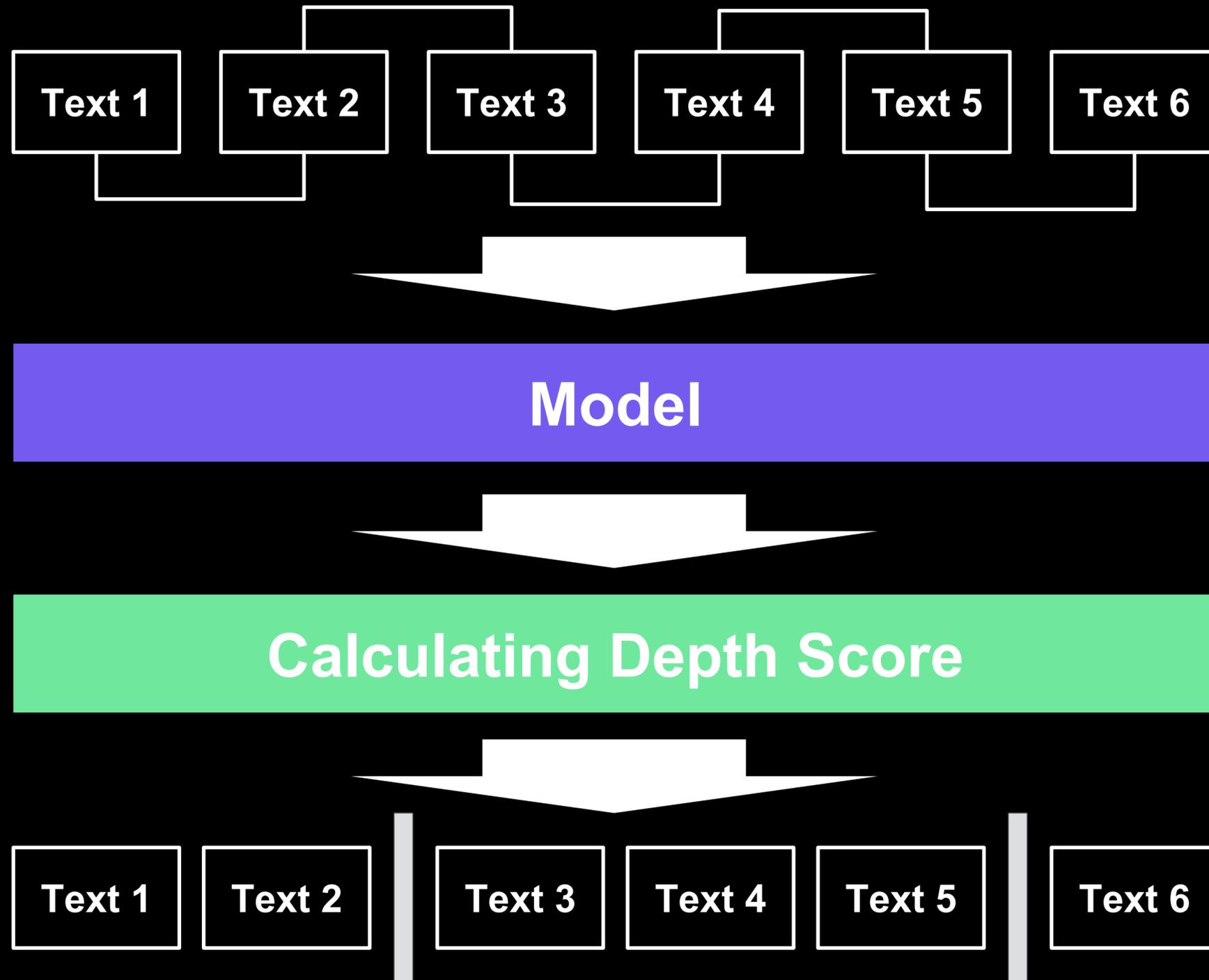
이 두 문장은 같은 단락인가? 다른 단락인가?

보기:

- 같은
- 다른

\* 본 프롬프트는 예시입니다.

# 3.5 Train Model



Text Pair를 입력 후  
모형 학습 진행

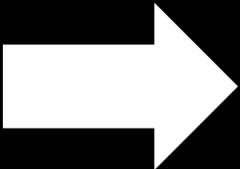
Depth Score 기반으로 적정  
Segmentation Point 산출

## 3.5 Train Model

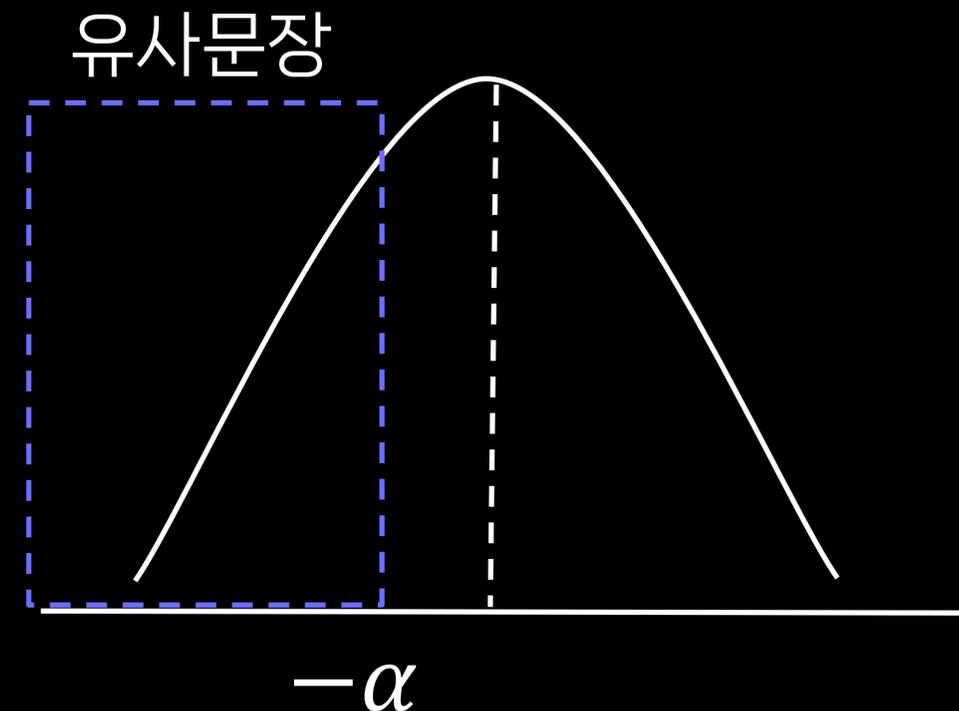
### Depth score 계산

- Open Domain을 다루고 있기 때문에 문서마다 Score 분포가 서로 다름
- Z-score로 정규화 한 이후에 thresholds를 적용하여 Segmentation을 수행

$$x \geq \mu - \alpha \cdot \sigma$$


$$\frac{x - \mu}{\sigma} \geq -\alpha$$

alpha 값이 작아질수록 Topic을 더 세분화 함



# 3.5 Train Model

총 3가지 Model로 학습을 진행

## EMB Model

LaRva - BERT 모델을 활용  
Representation 기반 학습 진행

$$\text{loss}(x, y) = \begin{cases} 1 - \cos(x_1, x_2), & \text{if } y = 1 \\ \max(0, \cos(x_1, x_2) - \text{margin}), & \text{if } y = -1 \end{cases}$$

## BERT

LaRVA - BERT 모델을 활용  
Classifier 학습 진행

CLS loss+ Marginal ranking loss

## GPT

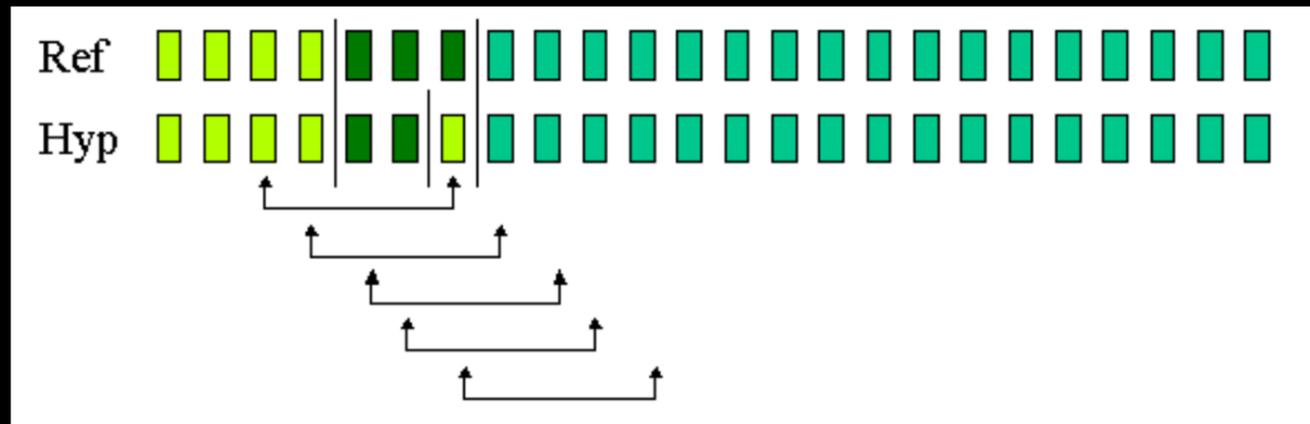
HyperCLOVA 모델 활용  
Fine Tuning, LoRA를 통한 학습 진행

LM loss + CLS loss+ Marginal ranking loss

$$\text{loss}(x_1, x_2, y) = \max(0, -y * (x_1 - x_2) + \text{margin})$$

# 3.6 Evaluation

Pk



sliding windows 방식으로 계산

일반적으로 windows 는 segmentation 개수의 절반으로 함

windows의 시작/끝 지점이 다르면 error count를 증가

error count를 전체 측정 횟수로 나눔

WindowDiff

$$WindowDiff(ref, hyp) = \frac{1}{N - k} \sum_{i=1}^{N-k} (|b(ref_i, ref_{i+k}) - b(hyp_i, hyp_{i+k})| > 0)$$

sliding windows 방식으로 계산

windows 에 있는 reference의 seg 개수와

예측한 seg 개수의 차이를 비교

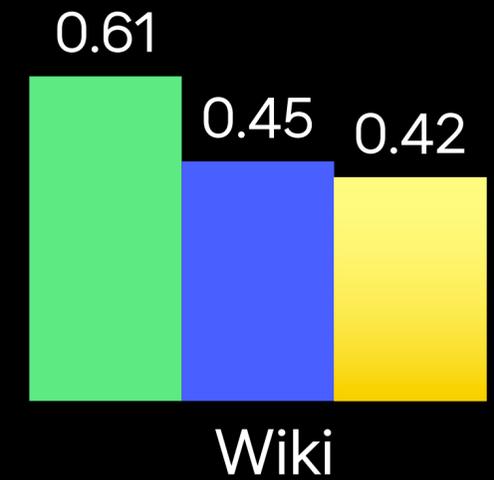
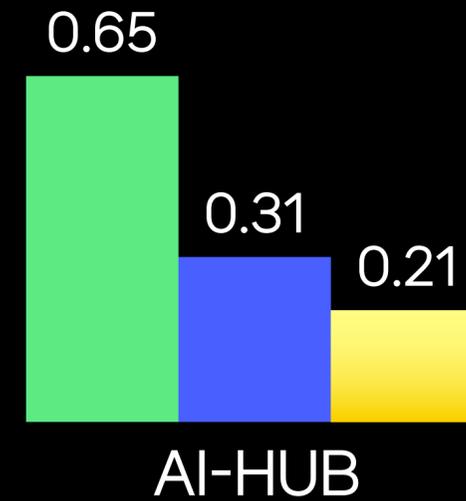
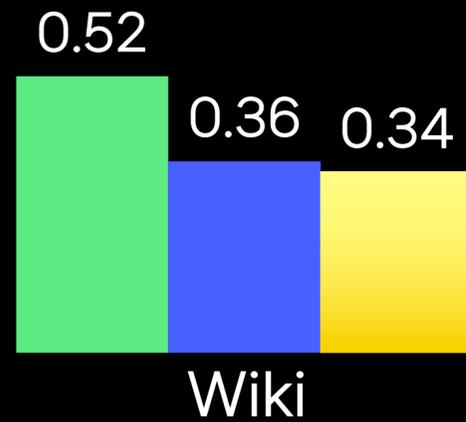
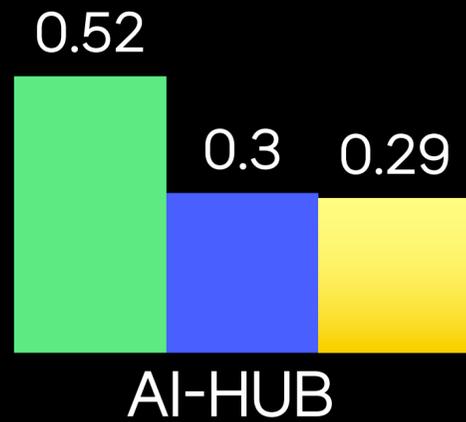
# 3.6 Evaluation

Pk

WindowDiff

■ EMB ■ BERT ■ GPT

■ EMB ■ BERT ■ GPT



## 3.7 Post Process

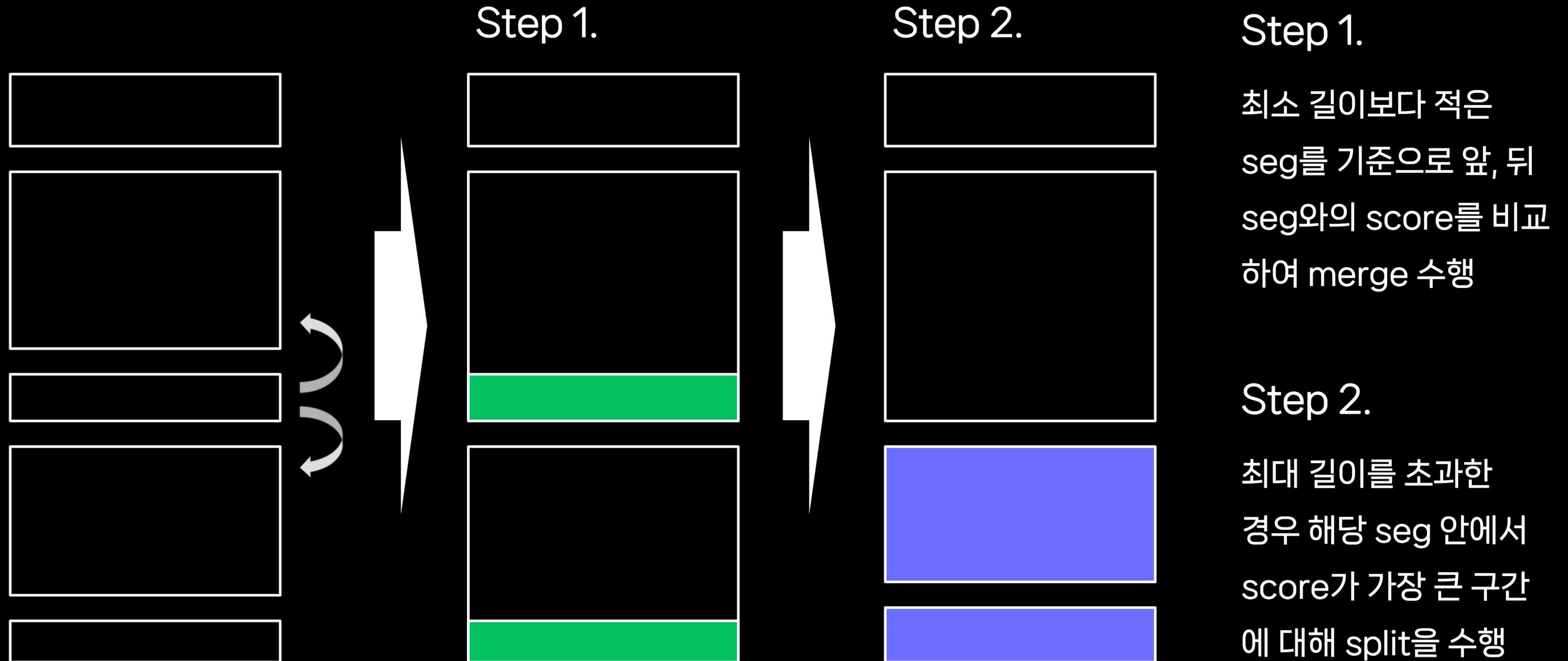
Segmentation 결과에 대한 후처리 니즈 존재

- 모형 학습 label → intent 단위로 학습을 진행
  - 문서가 길어질 경우 많은 수의 segmentation이 생길 수 있음 → 요약 결과의 가독성이 떨어짐

Approach

- thresholds 값을 조절하여 segmentation 구간을 조절
- Algorithm 적용
  - Heuristic Algorithm

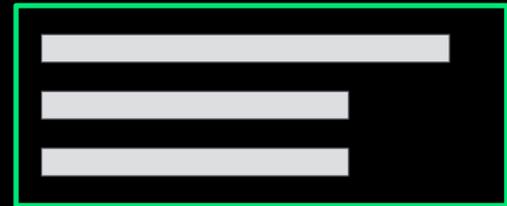
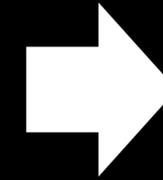
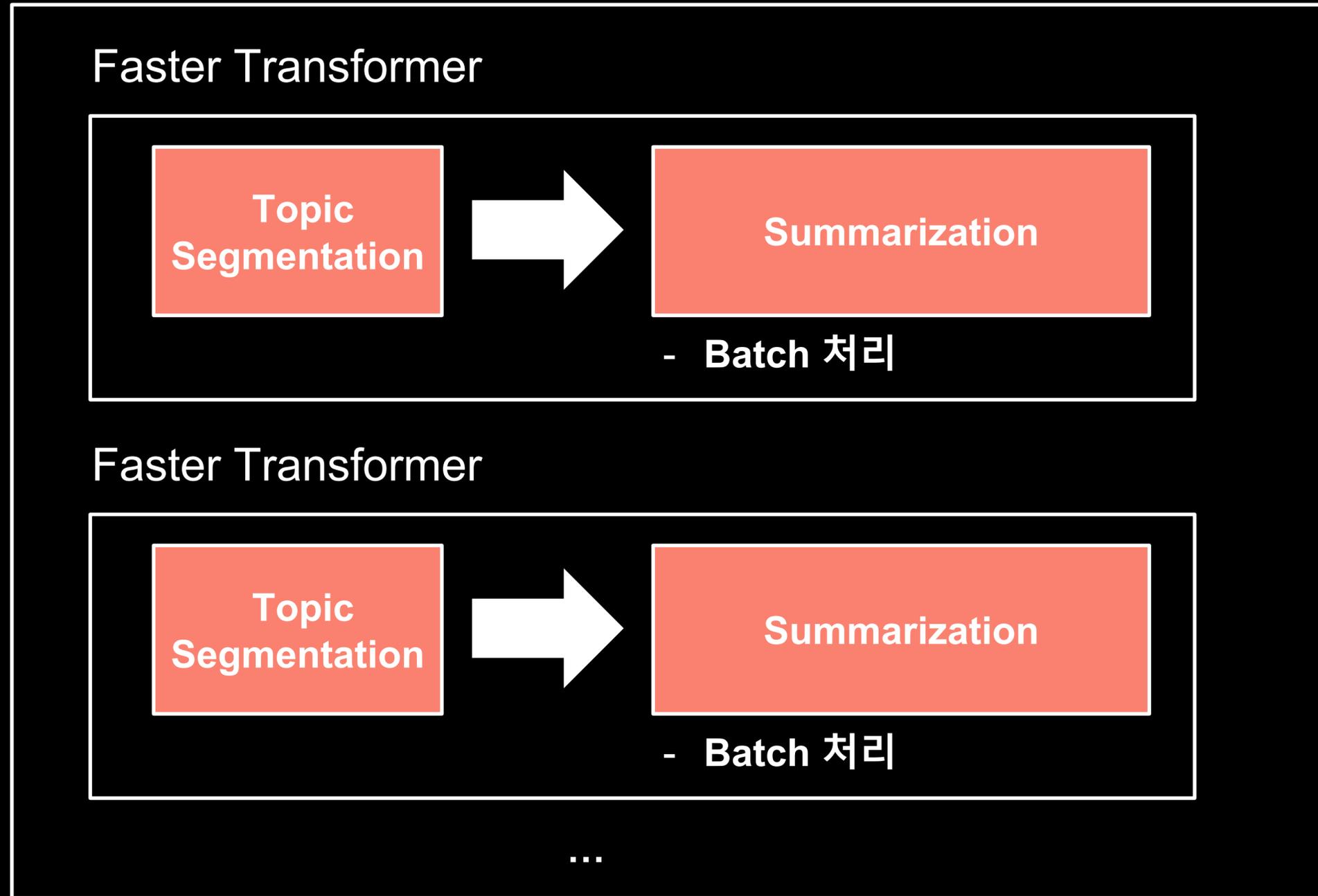
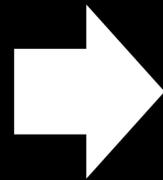
# 3.7 Post Process - Heuristic



# 4. Deploying & Serving

# 4.1 Pipeline

CLOps (K8S)



- Replica, Async API 기능 적용

# 5. Application

# 5.1 Applications

CLOVA  
Note  
KR

CLOVA  
Note  
JP

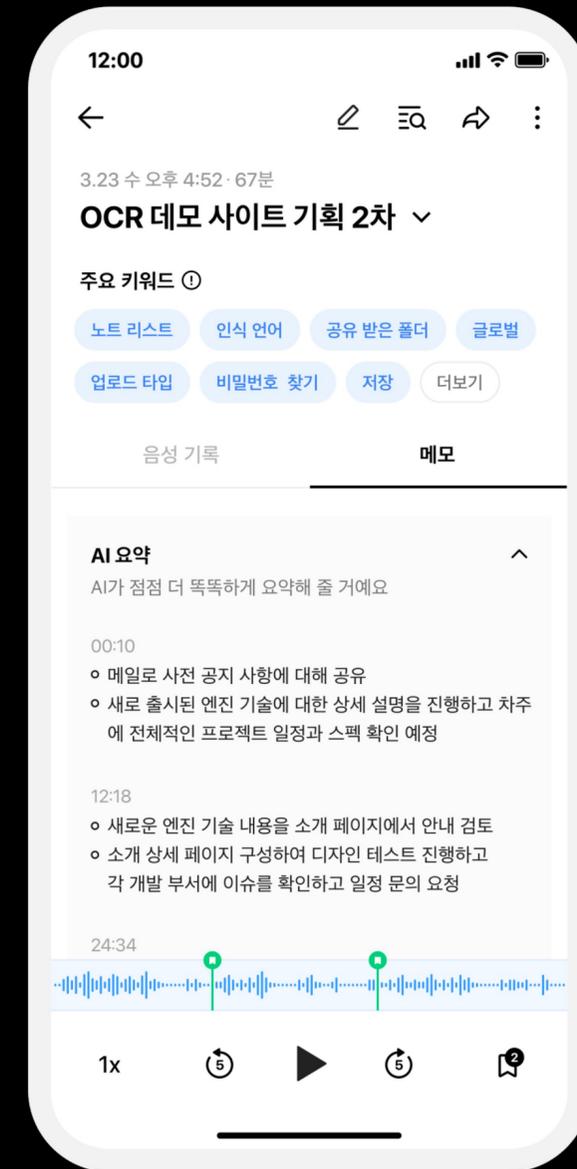
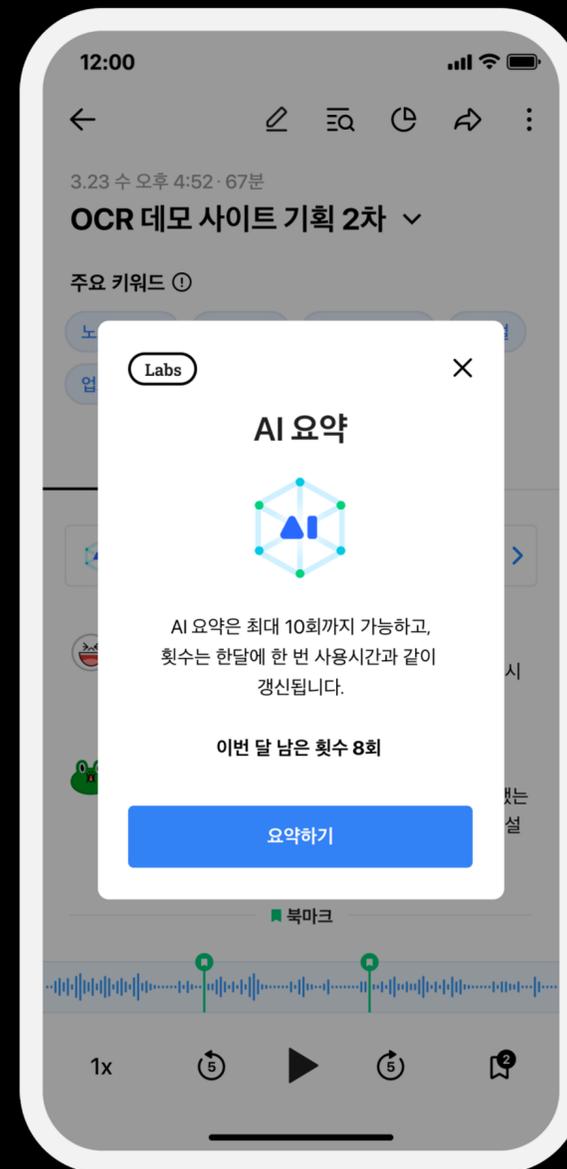
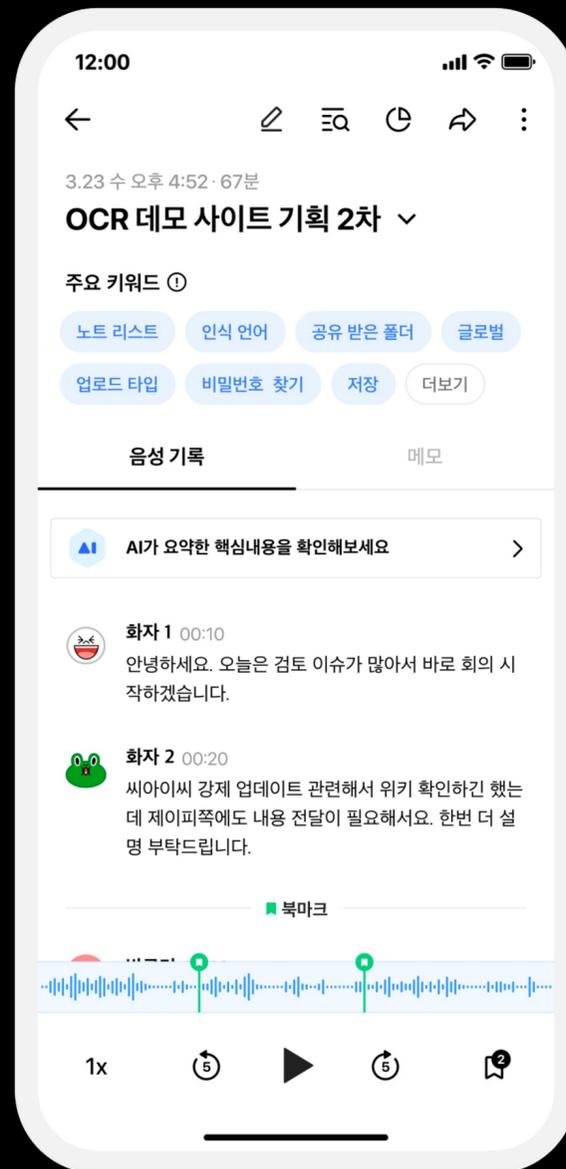
Web  
Plugin

Contact  
Center  
AI

Hyper  
CLOVA  
Studio

# 5.2 CLOVA Note - KR

22년 8월 서비스 출시



# 5.3 CLOVA Note - JP



Dataset

Open Dataset 활용

Train

HyperCLOVA JP 모형 활용

Serving

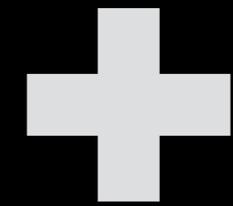
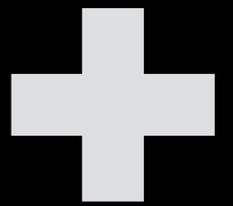
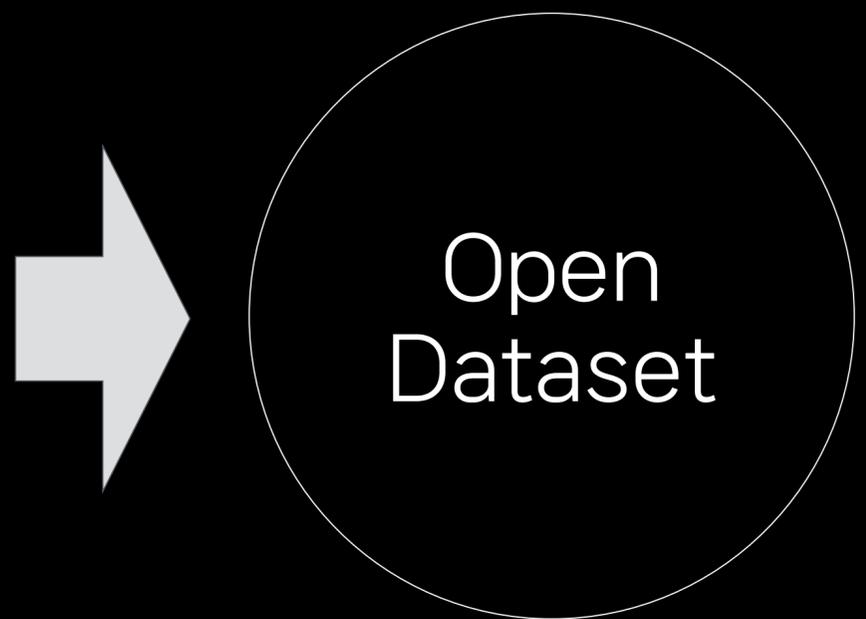
CLOps JP 활용

# 5.3 CLOVA Note - JP



Dataset 확보에 어려움

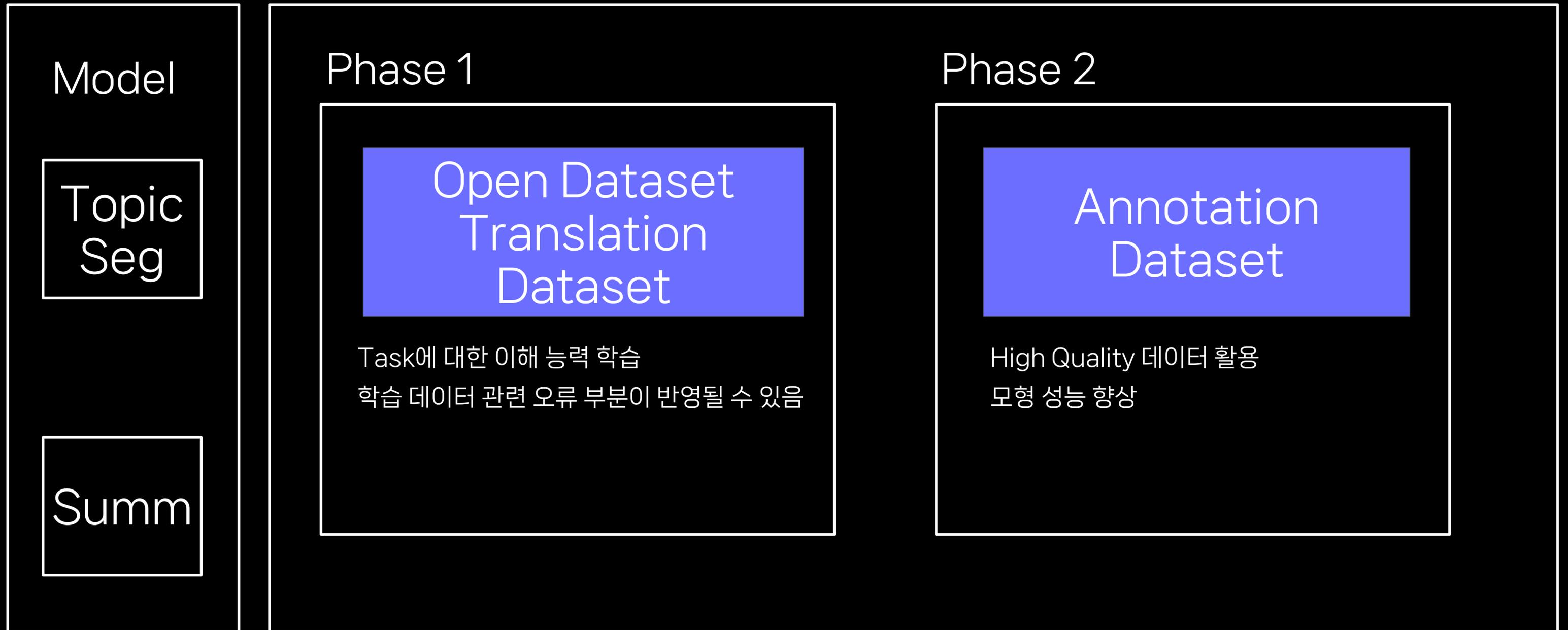
- Open 되어 있는 Dataset이 많지 않음
- 유의미한 Dataset의 경우 대부분 구매를 해야 함



한국어 학습 데이터  
→ Papago를 통해 번역

JP Annotator 인력 활용

# 5.3 CLOVA Note - JP



# 5.3 CLOVA Note - JP

音声記録

占有率

編集

**担当者** 00:00  
こんにちは。質問がありましたら気軽に声をかけてください。

**ユーザー** 00:06  
こんにちは。クローバノートはどんなサービスですか。

**担当者** 00:10  
クローバノートは会話の音声を自動的にテキストに変換し保存できるサービスです。例えばアイデアを集める会議やインタビューのように会話することに集中したい時に役立ちます。特に騒音がなく少ない人数の会話では正確に記録することができます。

メモ

**ユーザー** 00:32  
ノートはどう作成すればいいですか。

**担当者** 00:35  
方法は2つあります。一つ目はスマートフォンアプリを使って新しく録音を始める方法です。またスマートフォンに保存されている音声ファイルをアップロードすることもできます。2つ目は専用のウェブサイトを使い録音済みの音声ファイルをアップロードしてノートを作る方法です。

**ユーザー** 00:59  
ブックマークはどう使いますか。

**担当者** 01:02  
録音中の画面でブックマークボタンを押すと会話に印をつけられます。ブックマークがついていると録音が終わったあとで必要な区間を探しやすくなります。

メモ

## メモ

### 主なキーワード ⓘ

抽出できる主要キーワードがありません

### AI要約

AIがどんどんスマートに要約してくれます。足りない部分は編集で内容を追加してみてください。

00:00

- クローバノートは、会話の音声を自動的にテキストに変換し保存できるサービスである。
- 方法は、スマートフォンアプリを使って新しく録音を始める方法とスマートフォンに保存されている音声ファイルをアップロードする方法がある。
- 作成されたノートの会話を選択すると該当する部分の音声を再生できる。

00:18

メモを記載することができます。

要約：音声を自動的にテキストにしてくれるサービスです。

01:10

大事な会話を見つけやすくなるので、ブックマーク機能が便利です。

録音終了

メモを追加する

**0** NAVER Cloud Platform has launched ARC eye as a fully managed cloud-based digital twin service. Digital twin refers to the technology that replicates the physical world in the digital world. It is regarded as key data for innovative industries ranging from service robots, AR, autonomous driving, smart buildings, and smart cities. NAVER has been developing technology that can produce digital twins of large-scale spaces for many years and has taken the first step towards productization through the NAVER Cloud Platform. ARC eye is a service that builds digital twins of large-scale indoor and outdoor spaces such as vast shopping malls and buildings and provides artificial intelligence-based localization technology such as visual localization. Until now, NAVER LABS has used this technology to transfer various spaces to the digital world such as Incheon International Airport, COEX, and Gangnam Station. Most recently, this technology has been applied to NAVER's robot-friendly smart building '1784' so as to be used as the eyes of 100 robot units. This technology is also being used for the National Museum of Korea's digital twin-based AR navigation beta service. NAVER plans to respond to the demand in the digital twin technology market by providing ARC eye through the NAVER Cloud Platform. Digital twin solutions require advanced technologies and considerable costs. Hence, NAVER LABS has received various requests for ARC eye from many companies and organizations, both domestic and international. "Ever since NAVER exhibited this technology at the 2019 CES event, there have been many collaboration requests from global companies regarding NAVER's digital twin and localization technology," said Donghwan Lee, NAVER LABS' Vision group leader. "

**1** The fields and services seeking to utilize digital twin data are much more diverse than we expected." ARC eye, which was released through the NAVER Cloud Platform, offers all the necessary tools including digital twin creation, mapping robots and devices to do so, AI-driven localization APIs, and 2D mapping tools. The service was launched as a fully managed cloud-based service instead of an on-premise product, providing support ranging from data processing to API generation, which maximizes user convenience. Also, since it is a cloud-based service, the user is able to respond flexibly to various situations such as large-scale data storage and processing, or changes in usage, ensuring stable business operations. With the 'pay-as-you-go' policy where the user only pays according to their resource usage, users are able to save huge initial investment costs in the early stages of a new business based on digital twin technology and still take advantage of the latest equipment, algorithms, and web-based consoles. Such product design was made possible thanks to NAVER LABS' technology, NAVER Cloud Platform's business know-how, and NAVER 1784 which has served as a massive testbed. Mr. Sangyoung Han, the executive director of NAVER Cloud Platform's Strategy & Planning said, "Many participants in the previous S. Korea-Saudi Arabia Construction Roadshow expressed their keen interest in NAVER's digital twin technology applied at 1784 and asked very specific questions." He stressed that "NAVER Cloud Platform will become a global leader in Digital Twin As A Service with ARC eye as its first product."

### Summary: 1

- NAVERCloudPlatformは、完全に管理されたクラウドベースのデジタルツインサービスを開始した。
- デジタルツインは、デジタル世界で物理世界を再現する技術である。
- NAVERは、デジタルツインの技術を活用して様々な企業や団体からの要望に応じている。

### Summary: 2

- ARCEyeは、NAVERCloudPlatformからリリースされた。
- デジタルツインサービスは、クラウドベースのサービスで安定したビジネスを提供できる。

# 5.4 Contact Center AI

## SEG\_1

- 0 : "고객: 제가 중학생인데요, 체크카드 발급 받을 수 있나요?"
- 1 : "상담사: 네, 고객님의 만나이가 어떻게 되십니까?"
- 2 : "고객: 중3요 만15세요"
- 3 : "상담사: 본인이 직접 은행에 오실 수 있나요?"
- 4 : "고객: 네"
- 5 : "상담사: 입출금통장은있나요?"
- 6 : "고객: 엄마가 만들어준 통장이 있어요."
- 7 : "상담사: 만14세 이상, 미성년자 본인이 발급시 필요서류에 대해 안내해 드릴까요?"
- 8 : "고객: 네"
- 9 : "상담사: 주민등록증이나 유효기간이 남아있는 여권, 청소년증.또는 사진과 이름, 학교장직인날인, 주민번호 13자리가 표기된 학생증 중 1가지를 지참하여 영업점 방문을 해주시면 됩니다."
- 10 : "고객: 제가 가지고 있는건 학생증인데, 생년월일 밖에 없어요. 그럼 못 만드나요?"
- 11 : "상담사: 아닙니다. 생년월일이 표기된 학생증과 3개월 이내 발급된 주민등록초본을 지참하시면 됩니다."
- 12 : "고객: 부모님께서 대신 해주실 수 있는건가요?"
- 13 : "상담사: 네, 고객님의 법정 대리인이신 부모님께서 서류를 구비하셔서 방문 발급 가능합니다."
- 14 : "고객: 부모님이 가시면 어떤 서류가 필요하죠?"
- 15 : "상담사: 부모님의 신분증, 미성년 자녀 기준으로 발급된 텃어기본증명서 또는 상세기본증명서 미성년 자녀기준으로 발급된 상세 가족관계증명서가 필요합니다." 
- 16 : "고객: 너무 복잡하고 많네요."
- 17 : "상담사: 필요한 서류가 총 3가지구요, 모든 서류가 3개월 이내 발급된 원본 서류로 주민번호 13자리가 모두 표기되도록 준비해주셔야합니다."
- 18 : "고객: 잘 모르겠는데, 문자로 필요한 서류 안내를 받을 수 있을까요?"
- 19 : "상담사: 네, 가능합니다. 지금 바로 보내드리도록 하겠습니다."

## CATEGORY\_1

- 0 : "기타 문의"
- 1 : "안내요청"

## ISSUE\_1

- 0 : "체크카드 발급 관련 중학생도 가능한지 문의"

## ACTIONS\_1

- 0 : "상담사: 만 14세 이상, 미성년자 본인이 발급시 필요서류에 대해 안내해 드림"
- 1 : "고객: 부모님이 대신 해주실 수 있는건지 물어"
- 2 : "상담사: 부모님의 신분증, 미성년 자녀 기준으로 발급된 텃어기본증명서 또는 상세기본증명서, 상세 가족관계증명서가 필요함"

상담 내용을 기반으로 **카테고리**, 상담에 대한 주요 **이슈 사항** 및 이에 대한 **조치 사항**을 자동으로 생성

# 5.5 HyperCLOVA Studio



시작하기

## 문단 나누기 API

글의 단락을 구분할 수 있습니다.



시작하기

## 요약 API

긴 문장을 요약해 드립니다.

### 문단 나누기

글의 단락을 구분할 수 있습니다.

---

#### 요청

URL  
POST: /v1/api-tools/segmentation

---

#### Header

Name	Required	Description
<b>Authorization</b>	필수	인증 토큰값 (예)Basic {부여받은 token 값}
<b>Content-Type</b>	필수	application/json 으로 고정

---

#### 요청 바디

Name	Type	Required	Default	Description
<b>text</b>	string	필수	N/A	Segmentation을 수행할 문서
<b>alpha</b>	float	선택	0.0	Segmentation을 수행하기 위한 thresholds 값 값이 작을 수록 Segmentation을 세밀하게(Segmentation을 많이) 수행함 ~100인 경우에는 모형을 최적을 값을 찾아서 segmentation을 수행
<b>segCnt</b>	integer	선택	-1	원하는 Segmentation 개수를 입력하면 해당 개수로 Segmentation 된 결과를 return (-1 인 경우는 모형이 자동으로 segmentation을 수행)
<b>postProcess</b>	boolean	선택	false	Segmentation 수행 후 원하는 길이에 맞게 seg를 합쳐주거나 나누어 주는 후처리 모듈 이 부분이 true 되어야 postProcess 관련 argument 가 작동함
<b>postProcessMaxSize</b>	integer	선택	1000	post process module 을 적용할 때, seg에 포함되는 최대 char size
<b>postProcessMinSize</b>	integer	선택	300	post process module 을 적용할 때, seg에 포함되는 최소 char size

---

요청 예제

### 요약

긴 문장을 요약해 드립니다.

---

#### 요청

URL  
POST: /v1/api-tools/summarization

---

#### Header

Name	Required	Description
<b>Authorization</b>	필수	인증 토큰값. 예)Basic {부여받은 token 값}
<b>Content-Type</b>	필수	application/json 으로 고정

---

#### 요청 바디

Name	Type	Required	Default	Description
<b>text</b>	string	필수	N/A	입력 텍스트입니다. 모델은 이 텍스트 뒤에 출력 텍스트를 생성합니다.
<b>maxTokens</b>	integer	선택	256	생성할 최대 토큰 수입니다.
<b>cutoffMaxRouge</b>	float	선택	0.7	요약: 여러 문장 생성 시 앞문장과 rouge-l-f 기반 중복 여부를 억제함

# 6. Epilogue

# 6.1 To-Be Feature

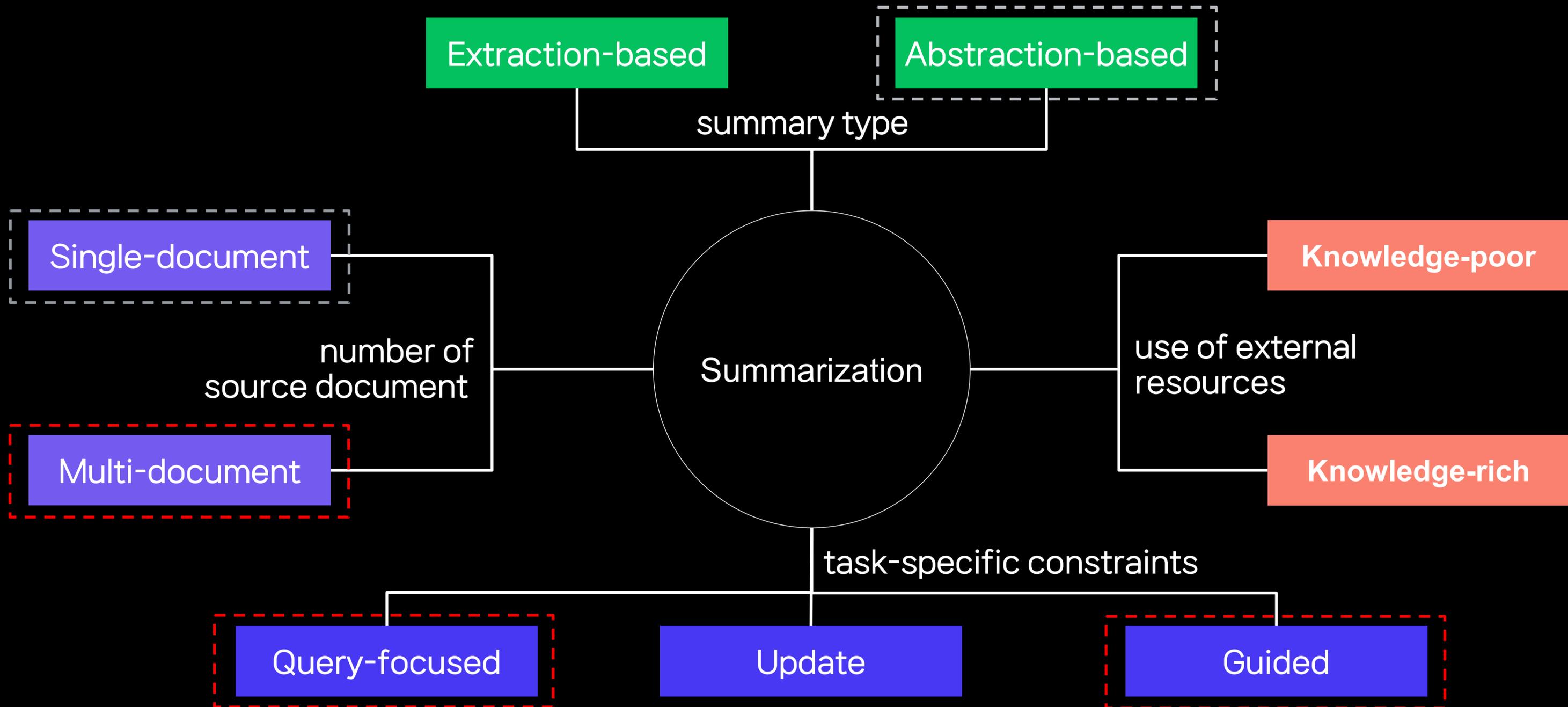
## 문서에 대한 구조적 이해

- 내가 원하는 형태로 문서를 요약 및 재조합
  - 회의, 강의, Interview 등의 목적에 맞게 요약 내용을 재조합
- 문서 내용을 기반으로 다양한 Action Item 도출
  - 제목 생성, Agenda 생성, Recommended Task 도출 등의 Feature 적용

## Multi-Document 요약

- 하나의 문서가 아닌 여러 개의 문서 내용을 구조적으로 요약
  - 사건 발생 순서, 특정 키워드 및 Event 기준으로 내용을 요약

# 6.2 Future Work



Q & A

**Thank You**