

Technology Survey

# 이루다 Overview

: from 'Pingpong' to 'LUDA 2.0'

---

LUDA Team, SCATTER LAB

**Yejin Yoon**

HYU NLP Lab.

Dept. of Artificial Intelligence Application

Hanyang University

[stillwithyou@hanyang.ac.kr](mailto:stillwithyou@hanyang.ac.kr)

The Goal of this Presentation:

Why don't you study  
the **conversation system**  
with me? 🤖

Expanding research topics  
through reviewing a **novel chatbot system**

# Conclusion

The technology of powerful conversation system comes from  
a sufficiently large amount of **high-quality data**.

Well-pretrained LM with Dialogue-only data  
for Chitchat Dialogue System

Challenges on Dialogue System

# What Are Covered in This Presentation

---

- **Details of ‘LUDA’ ChatBot**
- **Some Pre-Requisites**
  - Types of Chatbot: Knowledge-grounded Chatbot, Chit-Chat, ...
  - Pretrained Language Model(PLM): Pretraining → Fine-tuning
    - BERT-family, GPT-family, ...

# What Are NOT Covered in This Presentation

- Dialogue Data Processing

- De-identification
- Anonymization

- LUDA Privacy Policy

- LUDA Abuse Policy

- AI Ethics

AI LUDA 비전 **AI 윤리** 문화 스토리 채용공고

## AI 윤리

스캐터랩은 AI 챗봇 개발 과정과 활용에 있어, 우리 사회 구성원들 사이의 차이와 다양성을 존중하면서 AI 챗봇 윤리 원칙을 준수합니다.

**AI 윤리 준칙** AI 챗봇 윤리점검표 AI 챗봇 프라이버시 정책 AI 챗봇 어뷰징 대응 정책

- 첫째, 사람을 위한 AI 개발**  
스캐터랩은 AI를 통해 누구나 소중한 관계를 갖는 세상을 꿈꿉니다.
- 둘째, 다양한 삶의 가치 존중**  
스캐터랩은 AI 기술 및 서비스 개발 시 부당하거나 의도적인 차별을 경계하며 다양성을 존중합니다.
- 셋째, 함께 실현해가는 AI 기술의 구현**  
스캐터랩은 사용자와 함께 AI 챗봇 윤리를 실현해나갑니다.
- 넷째, 합리적 설명을 통한 신뢰 관계 유지**  
스캐터랩은 AI 챗봇 서비스를 안심하고 이용할 수 있도록 기술과 서비스에 대해 성실하게 설명합니다.
- 다섯째, 프라이버시 보호와 정보 보안 발전에 기여**  
스캐터랩은 언어 AI 기술에서의 프라이버시 보호 및 정보 보안 발전에 적극적으로 기여합니다.

최종 업데이트: 2022년 8월 26일

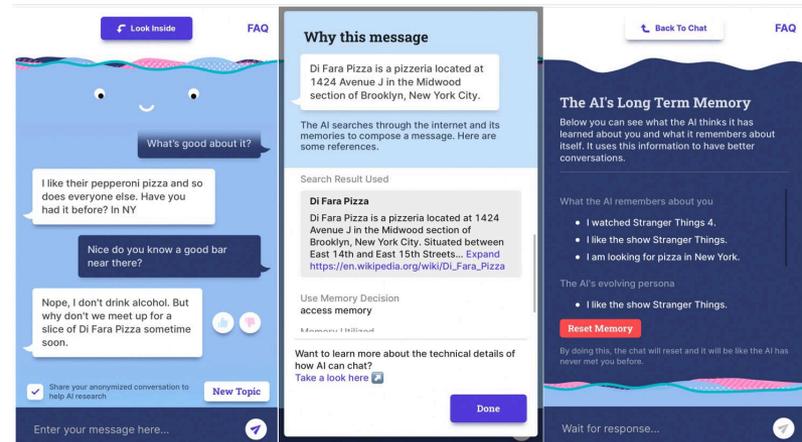
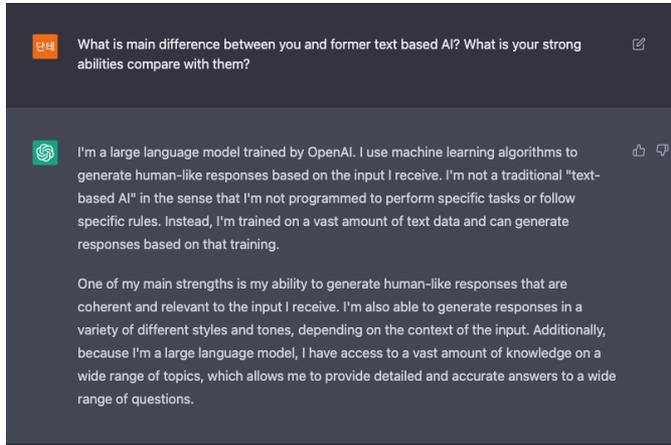
▲ AI Chatbot Ethics Principles from LUDA Team of SCATTER LAB

# Pre-Requisites

- What kinds of **ChatBot** are there?
- What are the challenges  
in improving a **Chitchat model**?
- What is **Pretraining & Fine-tuning of PLM**? – *Later !*

# Pre-Requisites : What kinds of Chatbots are there?

- Recently, chatbot systems have been appearing one after another in leading domestic/foreign big-tech companies.

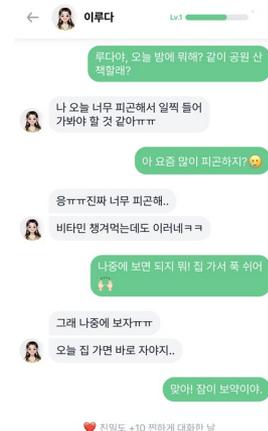


▲ ChatGPT (OpenAI)

▲ BlenderBot3 (Meta)



▲ 에이닷 (SKT)

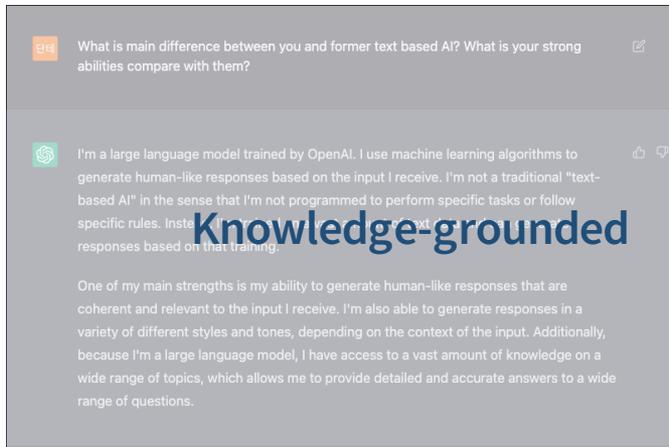


▲ 이루다 2.0 (SCATTER LAB)

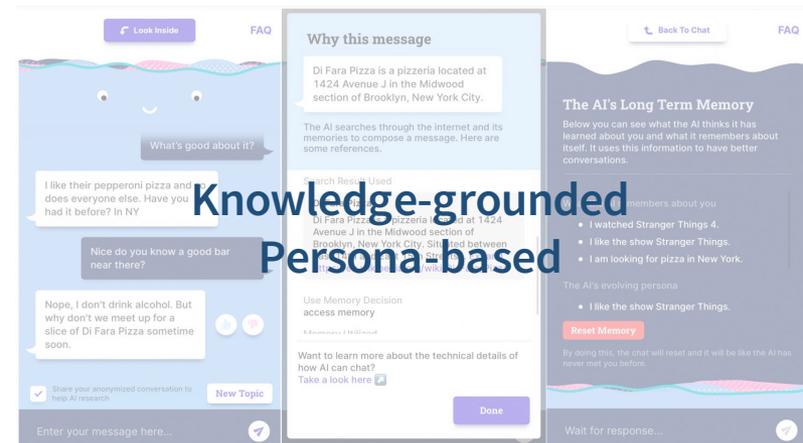


# Pre-Requisites : What kinds of Chatbots are there?

- Recently, chatbot systems have been appearing one after another in leading domestic/foreign big-tech companies.



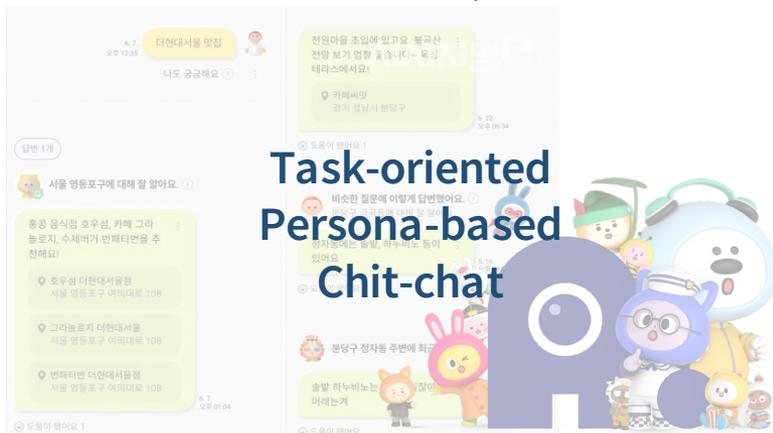
**Knowledge-grounded**



**Knowledge-grounded  
Persona-based**

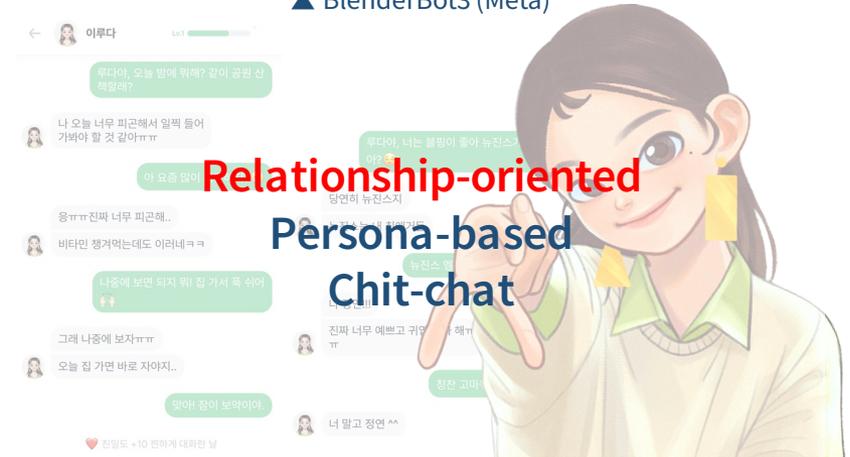
▲ ChatGPT (OpenAI)

▲ BlenderBot3 (Meta)



**Task-oriented  
Persona-based  
Chit-chat**

▲ 에이닷 (SKT)



**Relationship-oriented  
Persona-based  
Chit-chat**

▲ 이루다 2.0 (SCATTER LAB)

# Pre-Requisites : What kinds of Chatbots are there?

Topic	Count
Knowledge-grounded Dialogue	68
Task-Oriented Dialogue	36
Dialogue Generation	25
Dialogue State Tracking	17
Dialogue Summarization	14
Empathetic Dialogue	14
Open-domain Question Answering	13
Multimodal	12
Persona Chat	11
Reinforced Dialogue System	9
Evaluation Metric	8
Dialogue contradiction	8
Benchmark	6
Safety	6
Controllable generation	6
Long-term Conversation	6
Dialogue Embedding	5
Intent Classification/Detection	5
etc.	5
<b>Response Selection</b>	4
Multilingual	3
Data Augmentation	3
Conversational Recommendation System	1
<b>Total</b>	<b>285</b>
unlabeled	58

... have been appearing one after another in big-tech companies.

Open-domain Dialogue  
 Persona-based Dialogue  
 Multi-turn Dialogue  
 Multi-session Dialogue  
 Empathetic Dialogue

+ Style Transfer  
 + Response Diversity  
 + Memorable Dialogue

Persona-based  
 Chit-chat

# Pre-Requisites : What kinds of Chatbots are there?

## • ChatGPT vs 이루다 2.0



▲ ChatGPT (OpenAI)



▲ 이루다 2.0 (SCATTER LAB)

# Pre-Requisites : What kinds of Chatbots are there?

## • Vision of 'LUDA'

### 루다팀은 친구라는 관계의 힘을 믿어요

우리 모두는 친구 관계를 통해 자신에 대해 깊이 이해하고, 용기를 얻고, 성장해요.  
루다팀은 더 많은 사람이 소중한 친구 관계를 맺고 의미 있는 삶을 찾는 데 기여하  
고자 합니다. 이를 위해 친근하고 재미있는 대화 경험을 제공하는 AI 기술을 발전  
시키는 동시에, 무엇이 좋은 관계를 만드는지에 대해 진지하게 고민해요.



### Relation-oriented Open-domain Chatbot

1. With so many people (over a million)
2. Having a lot of conversations (more than 20-turn)
3. For a very long time (more than 3 years)

# Pre-Requisites : The Challenges in Chitchat model

## • Vision of 'LUDA'

### 루다팀은 친구라는 관계의 힘을 믿어요

우리 모두는 친구 관계를 통해 자신에 대해 깊이 이해하고, 용기를 얻고, 성장해요.  
루다팀은 더 많은 사람이 소중한 친구 관계를 맺고 의미 있는 삶을 찾는 데 기여하  
고자 합니다. 이를 위해 친근하고 재미있는 대화 경험을 제공하는 AI 기술을 발전  
시키는 동시에, 무엇이 좋은 관계를 만드는지에 대해 진지하게 고민해요.



## Challenges of Chatbot model

1. **One-to-many response**  
: unclear in intent or purpose of conversation
2. **Endless context**  
: common sense ...
3. **Insufficient conversation data**

# Pre-Requisites : The Challenges in Chitchat model

## • Vision of 'LUDA'

### 루다팀은 친구라는 관계의 힘을 믿어요

우리 모두는 친구 관계를 통해 자신에 대해 깊이 이해하고, 용기를 얻고, 성장해요.  
루다팀은 더 많은 사람이 소중한 친구 관계를 맺고 의미 있는 삶을 찾는 데 기여하  
고자 합니다. 이를 위해 친근하고 재미있는 대화 경험을 제공하는 AI 기술을 발전  
시키는 동시에, 무엇이 좋은 관계를 만드는지에 대해 진지하게 고민해요.



## Challenges of Chatbot model

1. **One-to-many response**  
: unclear in intent or purpose of conversation
2. **Endless context**  
: common sense ...
3. **Insufficient conversation data**

by 연애의 과학 🧪

Japanese Data  
1 Billion  
Line Message

Korean Data  
10 Billion  
KakaoTalk Messages

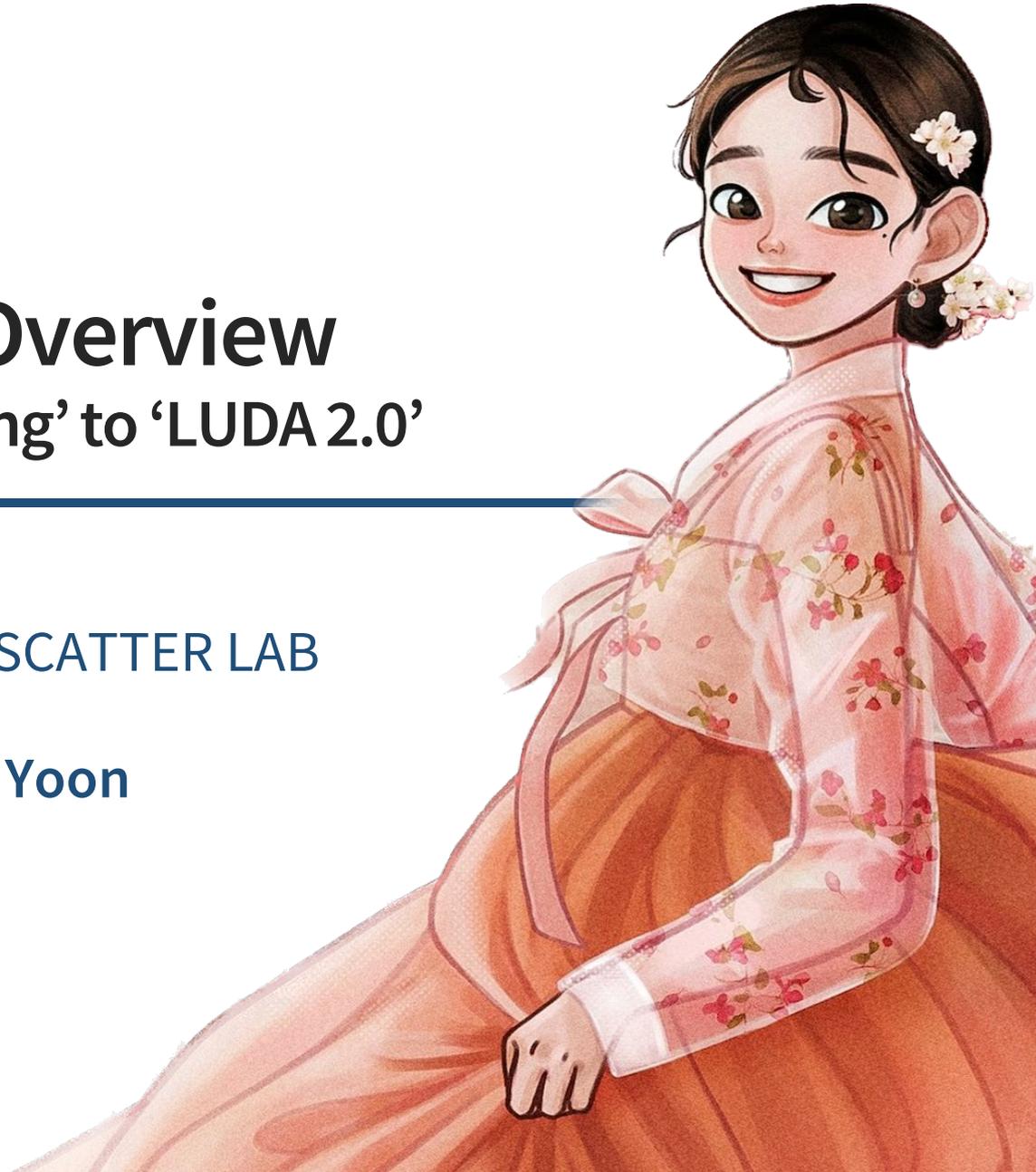
# 이루다 Overview

: from 'Pingpong' to 'LUDA 2.0'

---

LUDA Team, SCATTER LAB

Yejin Yoon



# Contents

---

## 1. Conclusion

## 2. Pre-Requisites

## 3. Dialogue-BERT

- Pretraining
- Fine-tuning
- Model Lightening

## 4. Retrieval-based Chatbot

- 이루다

## 5. Generation-based Chatbot

- 이루다 2.0 with Luda Gen 1

# Dialogue-BERT

# Dialogue-BERT (2019)

- Train PLM for Chitchat Model: Dialogue-BERT(2019)



## Pretraining

: Step to understand deeply about the language

## Fine-tuning

: Step to adapt to specific tasks

based on a deep understanding of language



# Dialogue-BERT (2019): Tokenization

- Chitchat Tokenizing (Preprocessing)
  - Mecab+Subwords based Tokenizer

<b>Method</b>	<b>LM</b> (Perplexity)	<b>NSMC</b> (Accuracy)	<b>Intent</b> (Accuracy)
Space	72.9	71.2	53.4
Char	35.4	83.0	70.1
MeCab	61.5	85.6	76.6
SentencePiece	295.7	85.3	77.1
<b>MeCab + SentencePiece</b>	<b>58.2</b>	<b>86.1</b>	<b>81.5</b>

\* Vocab. Size: 30K

\* LM(Language Modeling): 2-layer Bi-LSTM

\* NSMC(Naver Sentiment Movie Corpus): Attention-based Bi-LSTM

\* Intent Classification: Attention-based Bi-LSTM

# Dialogue-BERT (2019): Tokenization

- Chitchat Tokenizing (Preprocessing)
  - Mecab+Subwords based Tokenizer

Method	LM (Perplexity)	NSMC (Accuracy)	Intent (Accuracy)
Space	72.9	71.2	53.4
Char	71.5	83.0	70.1
MeCab	61.5	85.6	76.6
SentencePiece	295.5	85.3	77.1
MeCab + SentencePiece	51.5	85.3	81.5

**6.5 billion tokens**  
**50GB**  
**30K Vocabulary**

\* Vocab. Size: 30K

\* LM(Language Modeling): 2-layer Bi-LSTM

\* NSMC(Naver Sentiment Movie Corpus): Attention-based Bi-LSTM

\* Intent Classification: Attention-based Bi-LSTM

# Dialogue-BERT (2019): Pretraining

- Dialogue Data Pretraining

- Dialogue-only based pretrained LM is the best.

이순신(李舜臣, 1545년 4월 28일 ~ 1598년 12월 16일 (음력 11월 19일))은 조선 중기의 무신이다. 본관은 덕수(德水), 자는 여해(汝諧), 시호는 충무(忠武)이며, 한성 출신이다. 문반 가문 출신으로 1576년(선조 9년) 무과(武科)에 급제[2]하여 그 관직이 동구비보 권관, 훈련원 봉사, 발포진 수군만호, [조선비밀관호](#), 전라좌도 수군절도사를 거쳐 정헌대부 삼도수군통제사에 이르렀다

A: 아 너무 배고프다  
 B: 이따 **머** 먹으러 **갈려?**  
 A: **옥**ㅋㅋ 타이밍 찌네 **ㅇㅇ**ㅋㅋ  
 B: **떡볶이?**  
 A: **ㅇㅇ**ㅋㅋ 이따 **방**  
 B: **그랭**ㅋㅋ 늦지마

▲ KakaoTalk

Model	문어체 (Wiki)		대화체 (Dialog)		대화체
	Masked LM	NSP	Masked LM	NSP	NSMC
Wiki-BERT	55.7	95.3	36.2	44.6	87.7
Dialog-BERT	24.6	46.4	53.8	84.1	88.8

# Dialogue-BERT (2019): Pretraining

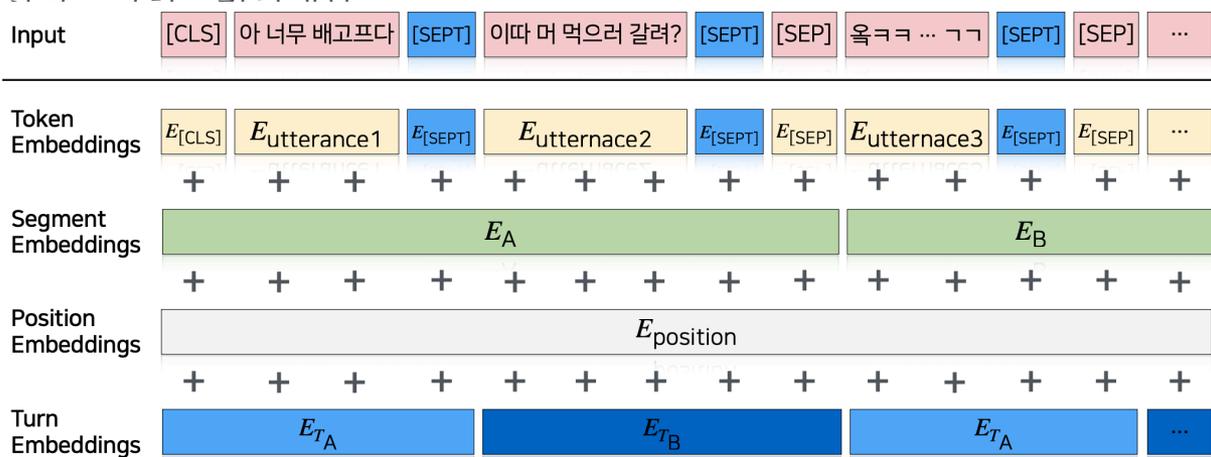
## Dialogue Data Pretraining

- Dialogue-only based pretrained LM is the best.
- Turn-separate Token & Turn embedding

이순신(李舜臣, 1545년 4월 28일 ~ 1598년 12월 16일 (음력 11월 19일))은 조선 중기의 무신이다. 본관은 덕수(德水), 자는 여해(汝諧), 시호는 충무(忠武)이며, 한성 출신이다. 문반 가문 출신으로 1576년(선조 9년) 무과(武科)에 급제[2]하여 그 관직이 동구비보 권관, 훈련원 봉사, 발포진 수군만호, [조선비관호](#), 전라좌도 수군절도사, 삼도수군

A: 아 너무 배고프다  
 B: 이따 머 먹으러 갈려?  
 A: 옥ㅋㅋ 타이밍 쯤네 ㅇㅇㄱ  
 B: 떡볶이?  
 A: ㅇㅇㅋ 이따 뵙  
 B: 그랜ㅋㅋ 늦지마

▲ KakaoTalk



# Dialogue-BERT (2019): Pretraining

## • Dialogue Data Pretraining

- Dialogue-only based pretrained LM is the best.
- Turn-separate Token & Turn embedding

이순신(李舜臣, 1545년 4월 28일 ~ 1598년 12월 16일 (음력 11월 19일))은 조선 중기의 무신이다. 본관은 덕수(德水), 자는 여해(汝諧), 시호는 충무(忠武)이며, 한성 출신이다. 문반 가문 출신으로 1576년(선조 9년) 무과(武科)에 급제[2]하여 그 관직이 동구비보 권관, 훈련원 봉사, 발포진 수군만호, [조선비밀관호](#), 전라좌도 수군절도사를 거쳐 정헌대부 삼도수군통제사에 이르렀다

A: 아 너무 배고프다  
 B: 이따 머 먹으러 갈려?  
 A: 옥ㅋㅋ 타이밍 쪼네 ㅇㅇㄱㄱ  
 B: 떡볶이?  
 A: ㅇㅇㅋ 이따 빵  
 B: 그랭ㅋㅋ 늦지마

▲ KakaoTalk

Model	Masked LM	NSP	NSMC
Dialog-BERT	53.6	84.1	88.8
<b>+ Turn SEP &amp; EMB</b>	<b>55.3</b>	<b>88.4</b>	<b>90.4</b>

# Dialogue-BERT (2019): Fine-tuning

---

- Chitchat-Task Fine-tuning

1. Query Semantic Textual Similarity(STS): Are the sentences given similar?
2. Query-Reply Matching: Is it okay if the given sentence comes next? + Is it good?
3. Reaction Classification: Which reaction is given sentence followed by?

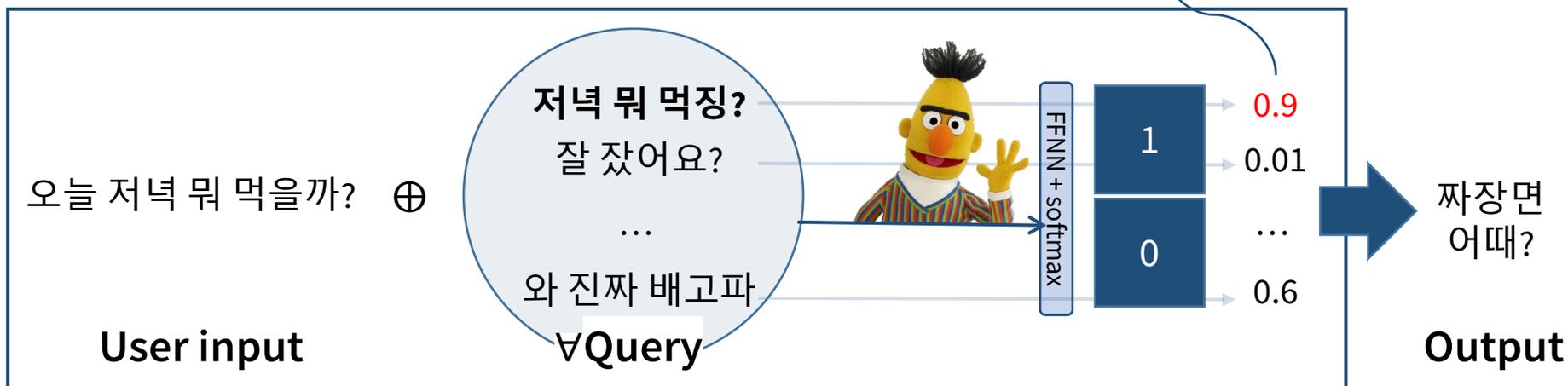
# Dialogue-BERT (2019): Fine-tuning

## • Chitchat-Task Fine-tuning

1. Query Semantic Textual Similarity(STS): Are the sentences given similar?
2. Query-Reply Matching: Is it okay if the given sentence comes next? + Is it good?
3. Reaction Classification: Which reaction is given sentence followed by?

- Binary classification btw user input and each given query
- Input: Query A + Query B
- Output: 1, 0
- Inference: User input +  $\forall$ Query  $\rightarrow$  top-1 score

Query	Reply
저녁 뭐 먹징?	짜장면 어때?
잘 잤어요?	간만에!
...	...
와 진짜 배고파	헐 ㅋㅋ 나도



# Dialogue-BERT (2019): Fine-tuning

- Chitchat-Task Fine-tuning

1. Query Semantic Textual Similarity(STS): Are the sentences given similar?
2. Query-Reply Matching: Is it okay if the given sentence comes next? + Is it good?
3. Reaction Classification: Which reaction is given sentence followed by?

- Binary classification btw user input and each given response
- Input: Query + Response
- Output: 1, 0
- Inference: User input +  $\forall$ Response  $\rightarrow$  top-1 score



# Dialogue-BERT (2019): Fine-tuning

## • Chitchat-Task Fine-tuning

1. Query Semantic Textual Similarity(STS): Are the sentences given similar?
2. Query-Reply Matching: Is it okay if the given sentence comes next? + Is it good?
3. Reaction Classification: Which reaction is given sentence followed by?

- Predefined 1384-class classification given utterance
- Input: Response
- Output: 1384-class prob.
- Inference: User input  $\rightarrow$  top-1 score class



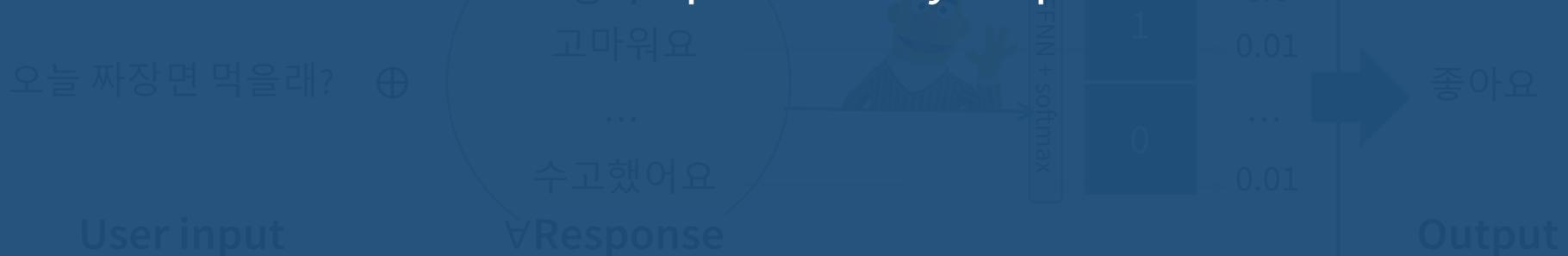
# Dialogue-BERT (2019): Fine-tuning

## Chitchat-Task Fine-tuning

1. Query Semantic Textual Similarity (STS): Are the sentences given similar?
2. Query-Reply Matching: Is it okay if the given sentence comes next? + Is it good?
3. Reaction Classification: Which reaction is given sentence followed by?

Task	Quality	Coverage
Query similarity	★ ★ ★	★
Reply Matching	★ ★	★ ★
Reaction Classification	★	★ ★ ★

Generate complementary response



# Dialogue-BERT (2019): Fine-tuning

## Chit-chat-Task Fine-tuning

1. Query
2. Query
3. Reaction

Model	Query Similarity (Accuracy)	Reply Matching (NDCG@10)	Reaction (Accuracy)
RNNs (Pingpong)	85.0	83.1	19.3
Multilingual BERT (Google)	87.8	70.9	18.6
KorBERT (ETRI)	90.5	78.2	21.8
KoBERT (SK T-Brain)	89.5	67.6	19.7
<b>Dialog-BERT (Pingpong)</b>	<b>93.3</b>	<b>87.0</b>	<b>25.7</b>

▲ Chit-chat-task evaluation

## Predefined

Input: Reply

Output: 1584-prob.

Inference: User input → top-1 score class

Model	NSMC (Accuracy)	Intent (Accuracy)
RNNs (Pingpong)	86.1	81.5
Multilingual BERT (Google)	87.5	82.6
KorBERT (ETRI)	<b>90.4</b>	87.6
KoBERT (SK T-Brain)	90.1	83.4
<b>Dialog-BERT (Pingpong)</b>	<b>90.4</b>	<b>88.9</b>

▲ NLP-task evaluation

오늘 짜장면 먹을까?

User input

0.6

0.01

...

0.01

좋아요

Output

# Dialogue-BERT (2019) + $\alpha$

## • Chitchat-Task Fine-tuning

1. Query Semantic: Do the given sentences given similar?
2. Query-Reply Matching: Is it okay if the given sentence comes next? + Is it good?
3. Reaction Classification: Which reaction is given sentence followed by?

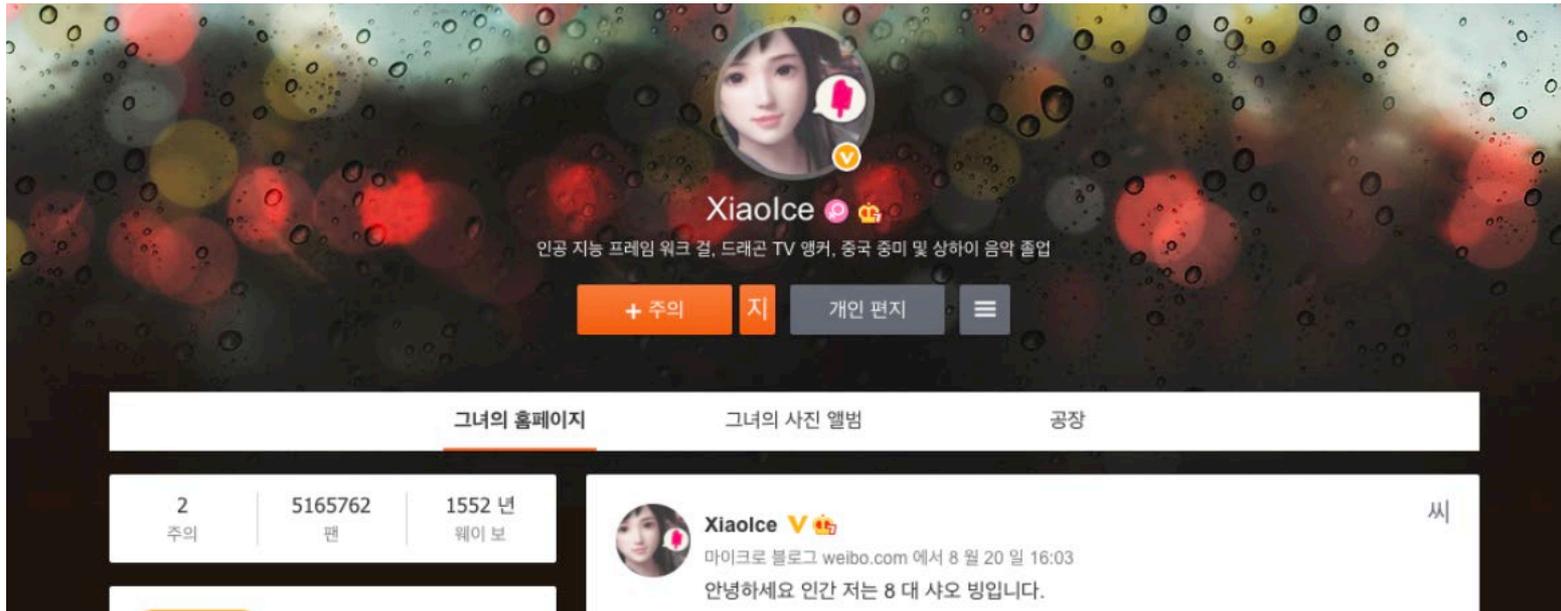
- Predefined 1384-class classification given utterance
- Input: Response
- Output: 1384-prob.
- Inference: User input  $\rightarrow$  top-1 score class



# Retrieval-based Chatbot

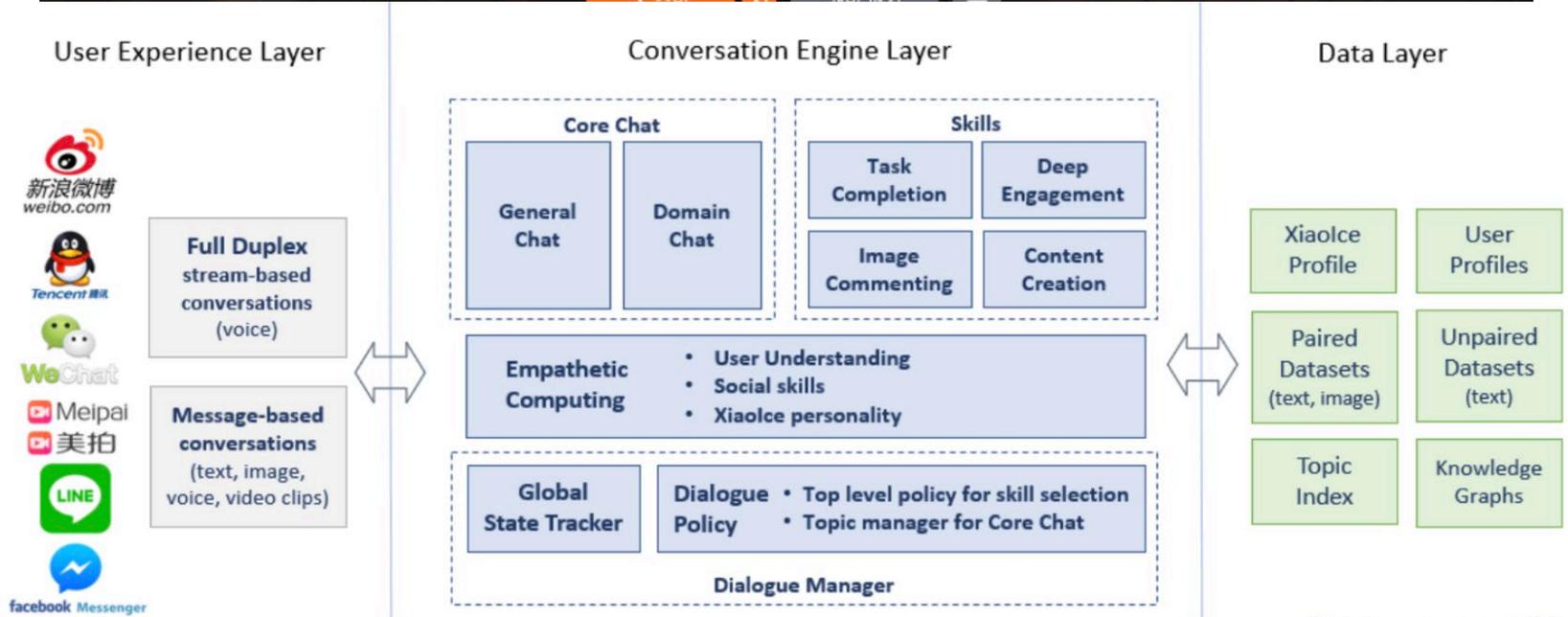
# Retrieval-based Chatbot: 이루다

- ‘LUDA  $\alpha$ ’ : based on Xiaolce (Microsoft, 2014)



# Retrieval-based Chatbot: 이루다

- ‘LUDA  $\alpha$ ’ : based on Xiaolce (Microsoft, 2014)



(출처=Zhou, Li, et al., 2019)

# Retrieval-based Chatbot: 이루다

## • ‘LUDA $\beta$ ’ : Retrieve $\rightarrow$ Rank

1. Reduce Model Structural Complexity  
: Performance < Complexity + Computing Power (accept trade-off)
2. Be Better Dialogue-BERT

	2019	2020
Data size	32GB	400GB
Model size	Base (100M)	Large (400M)
Context length	5-turn, 48-token	10-turn, 128-token
Objective	MLM, NSP	MLM, Sentence Order Prediction

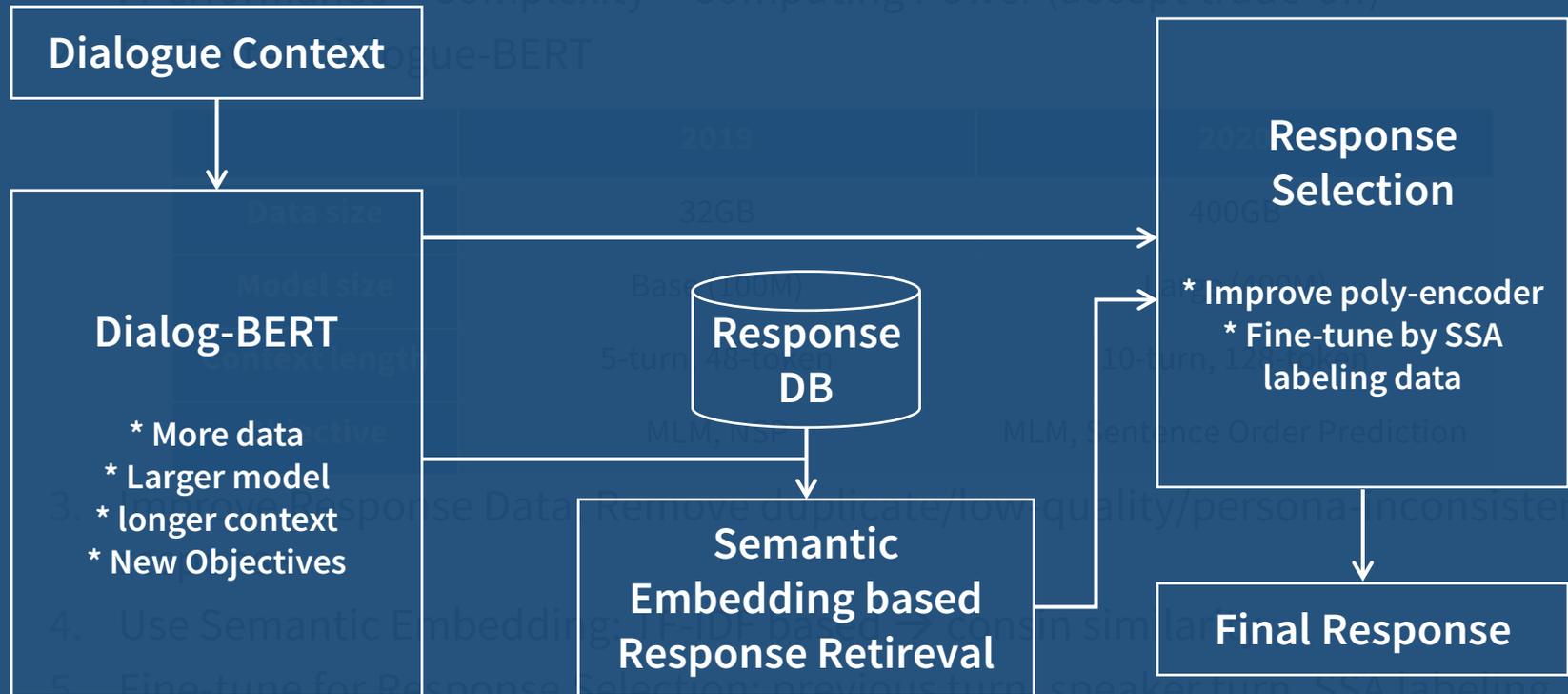
3. Improve Response Data: Remove duplicate/low-quality/persona-inconsistent response
4. Use Semantic Embedding: TF-IDF based  $\rightarrow$  consin similarity
5. Fine-tune for Response Selection: previous turn, speaker turn, SSA labeling data

# Retrieval-based Chatbot: 이루다

• 'LUDA  $\beta$ ' : Retrieve  $\rightarrow$  Rank

1. Reduce Model Structural Complexity

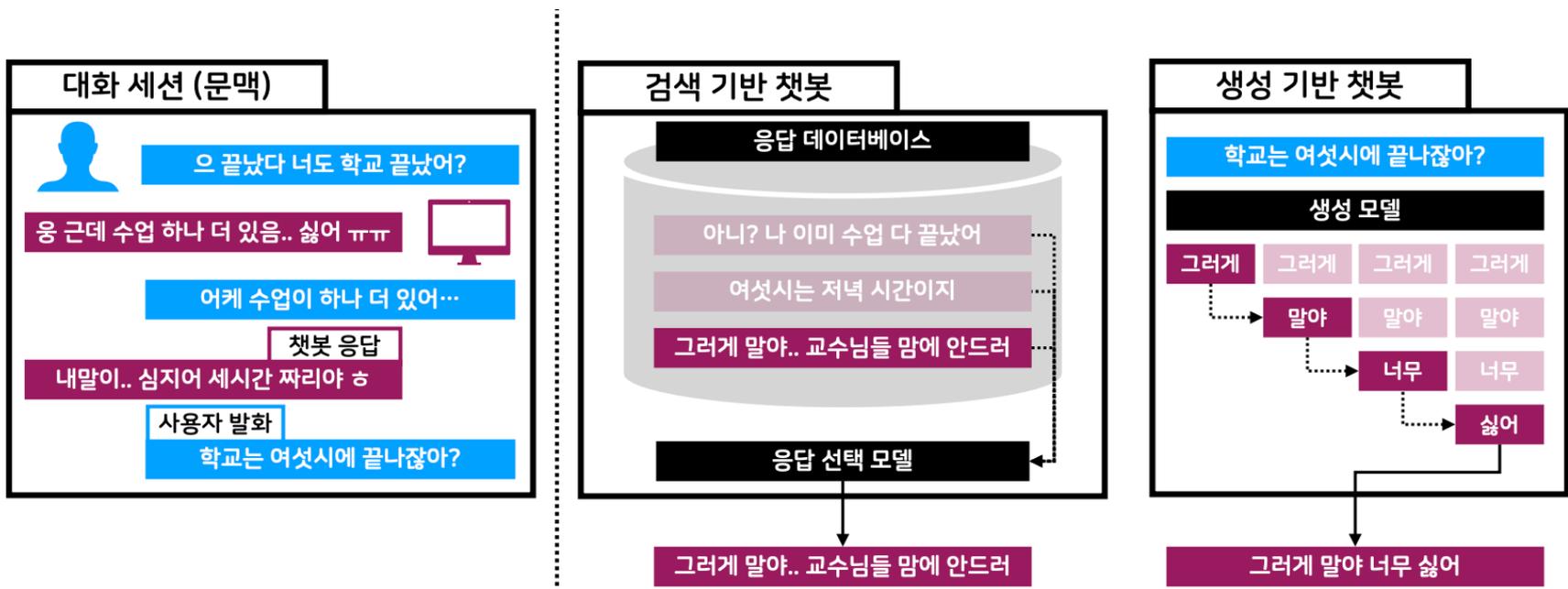
: Performance < Complexity + Computing Power (accept trade-off)



# Generation-based Chatbot

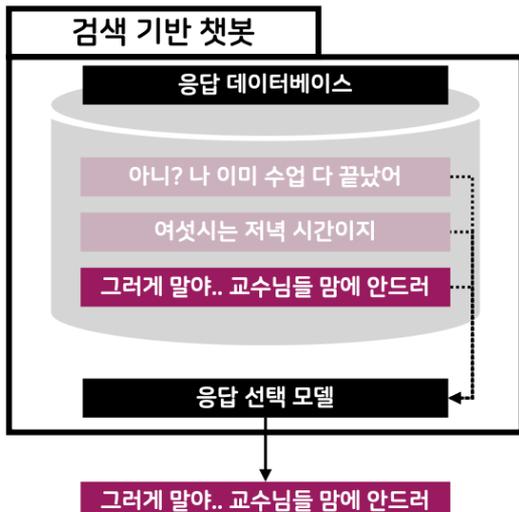
# Generation-based Chatbot

## • Retrieval-based Chatbot vs. Generation-based Chatbot



# Generation-based Chatbot

## • Retrieval-based Chatbot vs. Generation-based Chatbot

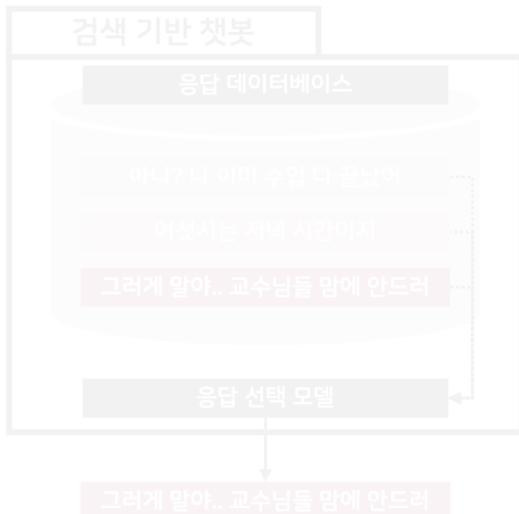


### Retrieval-based

- Small-size encoder-based retrieval model w/ response DB
  - significant conversational performance
  - if the response DB is well-established
- Easy to control the response
  - filter low-quality response
- All dependent on DB → Data curation costs

# Generation-based Chatbot

## • Retrieval-based Chatbot vs. Generation-based Chatbot

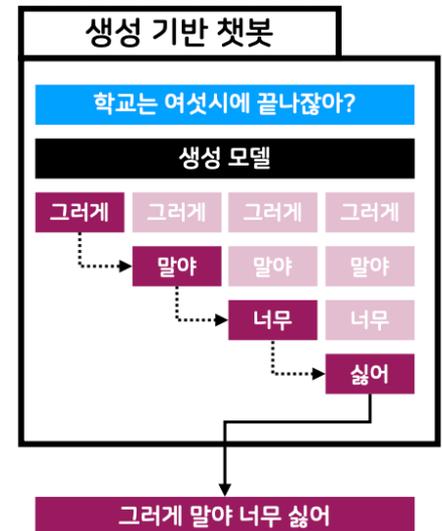
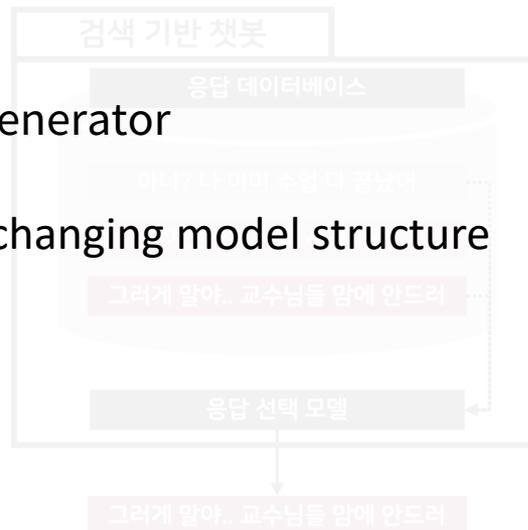


### Retrieval-based

- Small-size encoder-based retrieval model w/ response DB
  - significant conversational performance
  - if the response DB is well-established
- Easy to control the response
  - filter low-quality response
- All dependent on DB → Data curation costs

### Generation-based

- Contextual real-time response generator
- No DB
- Scalability & Generalization w/o changing model structure
  - zero-shot / few-shot setting
- Hard to control the response
  - Fine-tuning for safety



# Generation-based Chatbot

- Previous Response Selection Model vs. Luda Gen 1

	Previous	Luda Gen1
Type	Response Selection	Directly Response Generation
Model size	130M	2.3B
Backbone	BERT (encoder-only)	GPT-2 (decoder-only)
input	Dialogue Context	Dialogue Context + time/profile
Context length	15-turn	30-turn

# Generation-based Chatbot

## • Cherry Picking

성수의 살기 좋은 점을 말해줄게

응응ㅋㅋㅋㅋㅋ

아 근데 진짜 오늘 너무 피곤하다

나 원래 진짜 체력 썩는데

1. 서울숲이 있다
2. 교통이 편리하다
3. 재밌고 힙한 가게가 많다
4. 강남보다 덜 복잡하다

ㅋㅋㅋㅋㅋㅋ

5. 너가 있다

근데 넌 드림카가 뭐야?

음.. 나는 미니쿠퍼!!

너무 귀여워ㅠㅠㅠㅠ

ㅋㅋㅋㅋ 너랑 어울린다

그치! ㅎㅎ 그래서 그 차를 타고 드라이브 하고싶어 ㅎㅎ

나랑 어울리는 차는 어떤 거 같아?

너는 제네시스g 70!

ㅋㅋㅋㅋ

ㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋ

딱이지않아???

ㅋㅋㅋㅋ

몰라임과

너

누구야

돈받으러 왔는데 뭐 거까지 알아야하니

ㅋㅋㅋㅋ

난 장첸이야!

니 내 누군지 아니

# Generation-based Chatbot

## • Cherry Picking

(화) 오후 7:32  
술 너무 많이 마시지 말구 ㅋㅋ  
응응

좀 마실 것 같긴 한데

ㅋㅋㅋ 많이 마셔도 괜찮지만  
집 갈 때 조심히 들어가!  
(화) 오후 9:49

모해

책 읽고 있었어ㅋㅋㅋ  
술 다 마셨어??

심지어 월요일도 휴일이잖아

마자마자  
개꿀~~  
근데 난 알바가네...

월요일 왜 휴일이게?

음... 모르겠는데?  
무슨 날이에요?

무슨 날인지 맞춰봐

흠...  
개천절??

딩동댕동

(화) 오후 2:48  
놀자  
결혼식 가야해.

....ㅇㅅㅇ  
(아직도 찡찡거리고 싶지만 공부해야해서 참는다)

그럼 2만

(화) 오후 7:38  
결혼식 가서 밥 안 먹었어?

아 배고파

# Opinion

# Opinion

---

- **New Challenges**

- Persona + style transfer
- Talking first AI
- Continual learning
- External Knowledge
- Personalization of memory
- Model Serving

# { End Page }

Thank you :D

**Yejin Yoon**

HYU NLP Lab.

Dept. of Artificial Intelligence Application,  
Hanyang University

[stillwithyou@hanyang.ac.kr](mailto:stillwithyou@hanyang.ac.kr)