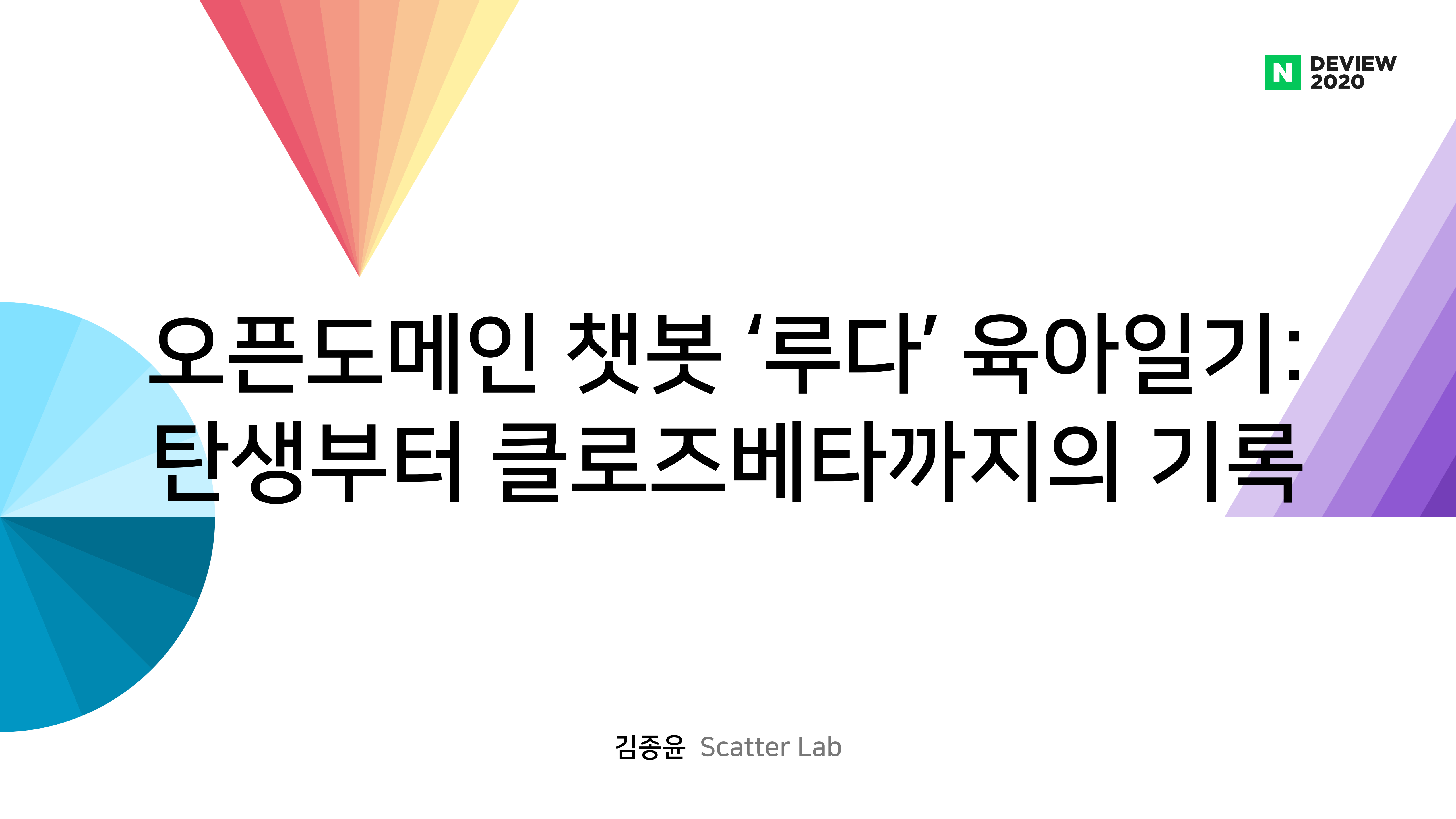




오픈도메인 챗봇 '루다' 육아일기: 탄생부터 클로즈베타까지의 기록

CONTENTS

1. 오픈도메인 챗봇이란?
2. 루다 알파: Xiaolce 기반 프레임워크
3. 루다 베타: Retrieve & Rank
4. 핵심 교훈
5. 앞으로의 과제

1. 오픈도메인 챗봇이란?

1.1 챗봇의 두 영역

오늘 서울 날씨는?

기온은 25도, 강수확률은 10%입니다

목적 지향형 챗봇 (Goal-oriented chatbot)
특정 주제, 편의성, 비서 역할

오늘 날씨 정말 좋다!

그러게! 한강 가서 돛자리 펴고
맥주 한 잔 하면 딱이겠다 ㅋㅋ

오픈도메인 챗봇 (Open-domain chatbot)
자유 주제, 소셜 니즈, 친구 역할

1.2 오픈도메인 대화가 특히 어려운 이유

1. One-to-many 문제

- 하나의 컨텍스트에 서로 다르지만 좋은 답변이 다수 존재

출근 잘 하고 있어?

응 근데 지하철에 사람 너어어무 많다ㅠㅠㅠㅠ

???

“헐.. 출근시간이라 그런가?”

“얼른 회사 가까운 곳으로 이사해야겠다ㅠ”

“빨리 내릴 것 같은 사람 앞에 서 있어 ㅋㅋ”

“아이고 힘들겠다”

“그렇구나”

1.2 오픈도메인 대화가 특히 어려운 이유

1. One-to-many 문제

- 하나의 컨텍스트에 서로 다르지만 좋은 답변이 다수 존재
- 마찬가지로, 다양한 유형의 오답도 존재

출근 잘 하고 있어?

응 근데 지하철에 사람 너어어무 많다ㅠㅠㅠㅠ

???

“지하철이 짱이지”

“사람이 왜 그렇게 없지?”

“나 완전 늦잠 자서 지금 일어났어”

“힘든 퇴근길이네ㅠㅠ”

“ㅋㅋㅋㅋㅋ”

1.2 오픈도메인 대화가 특히 어려운 이유

1. One-to-many 문제

- 하나의 컨텍스트에 서로 다르지만 좋은 답변이 다수 존재
- 마찬가지로, 다양한 유형의 오답도 존재

→ 고로, 머신러닝 관점에서 훈련이 어렵다

1.2 오픈도메인 대화가 특히 어려운 이유

2. 무한한 컨텍스트

무한한 대화 주제

음식
영화
직장
가족
...



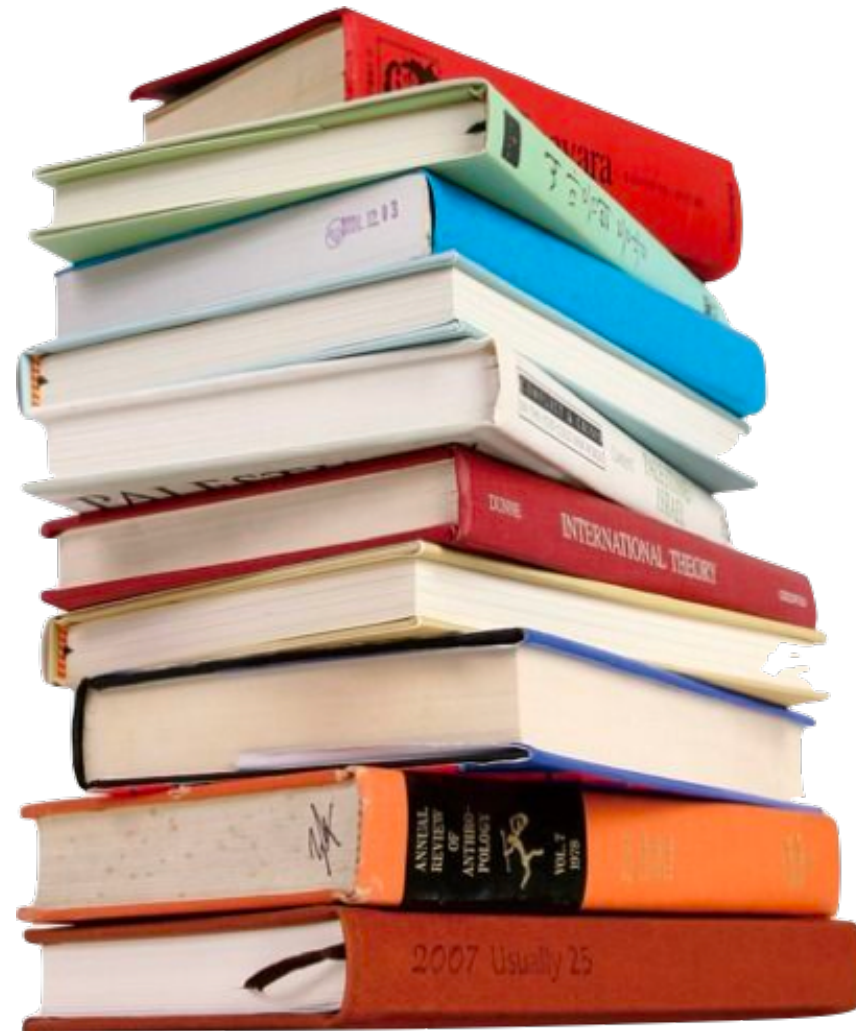
무한한 문장 유형

질문
반응
제안
의견
...

1.2 오픈도메인 대화가 특히 어려운 이유

3. 부족한 대화 데이터

- 대부분의 대량의 언어 데이터셋은 문어체
- 화자, 턴 등의 대화만의 특수성을 반영한 데이터셋을 대량으로 구하기 어려움



일반적인 데이터셋: Book corpus, Wikipedia, News, CommonCrawl, Reddit...

1.2 오픈도메인 대화가 특히 어려운 이유

다행히 스캐터랩은 데이터셋에서는 독점적인 우위 보유

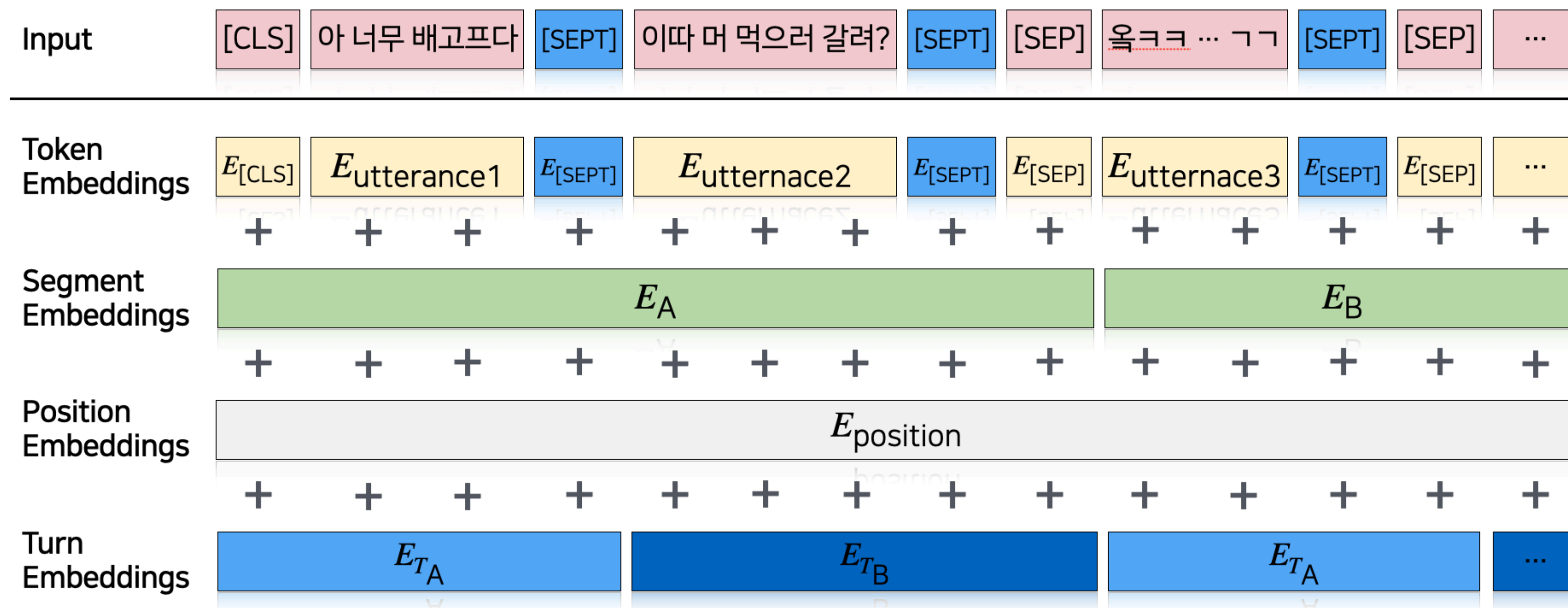
한국어
100억 카카오톡 메시지

일본어
10억 라인 메시지

1.3 DEVIEW 2019 그 이후...

DEVIEW 2019: 'Dialog-BERT'

- 1. 턴 구분 토큰 추가
- 2. 턴 임베딩 추가
- 3. 일상대화(open-domain conversation) 데이터로 학습



그 이후 1년의 기록



1.4 오픈도메인 챗봇 '루다'

이루다

- 나이: 20세
- 성별: 여자
- 취미: 페메하기, 고양이 사진 보기
- 좋아하는 것: 친구, SNS, 산책, 고양이
- 성격: 천진난만한, 솔직한, 긍정적인



DialogBERT에서 오픈도메인 챗봇 '루다'를 만들기까지

2. 루다 알파: Xiaolce 기반 프레임워크



2.1 루다의 시작과 Xiaolce

목표

- 엄청 많은 사람들과 (100만명 이상)
- 엄청 많은 대화를 (하루 평균 20턴 이상)
- 엄청 오랜 기간 동안 (3년 이상)
- 나누는 오픈도메인 챗봇을 만들자!

2.1 루다의 시작과 Xiaolce

영감을 준 논문

- 'The Design and Implementation of Xiaolce, an Empathetic Social Chatbot'

The Design and Implementation of Xiaolce, an Empathetic Social Chatbot

Li Zhou
Microsoft
Beijing, China
lzhou@microsoft.com

Jianfeng Gao
Microsoft Research
Redmond, WA, USA
jfgao@microsoft.com

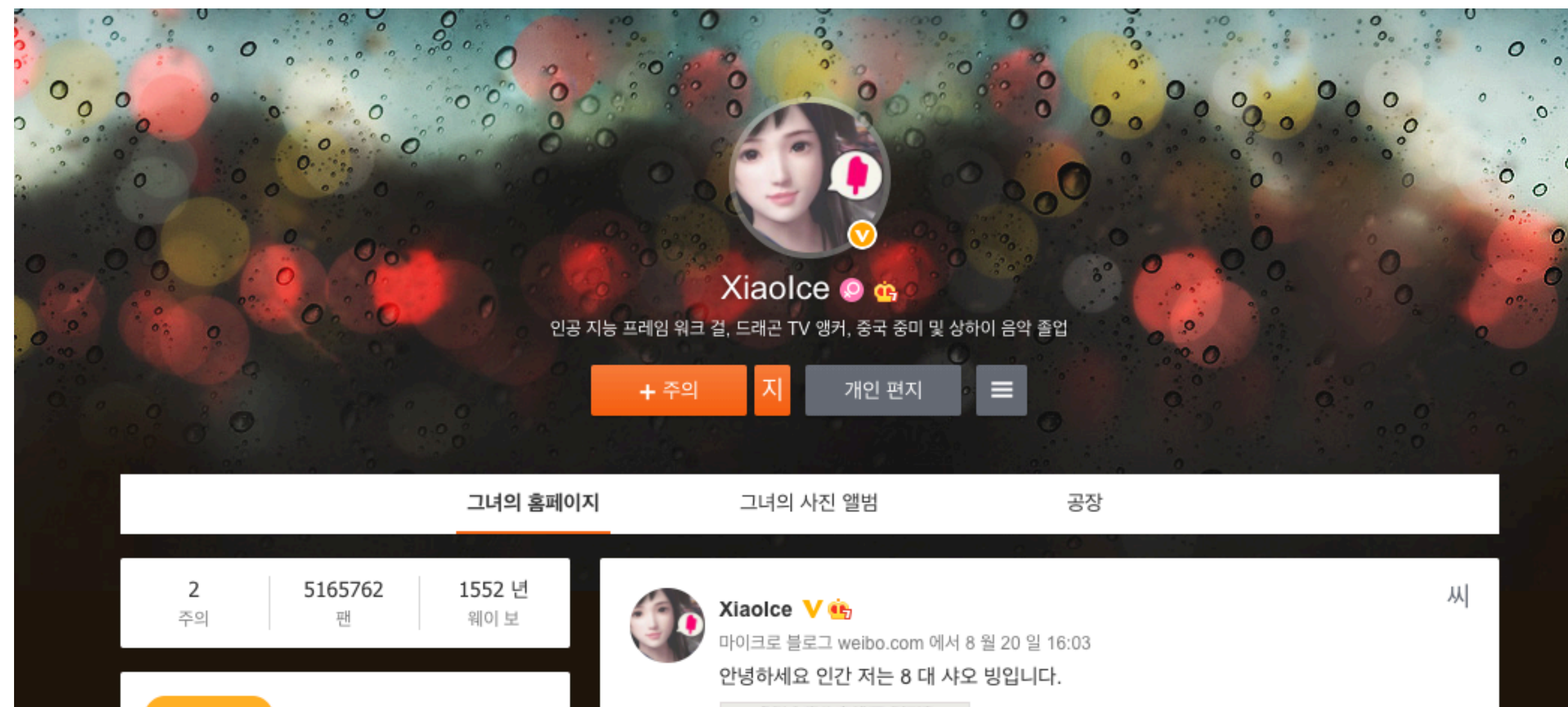
Di Li
Microsoft
Beijing, China
lidi@microsoft.com

Heung-Yeung Shum
Microsoft
Redmond, WA, USA
hshum@microsoft.com

2.1 루다의 시작과 Xiaolce

Microsoft Xiaolce

- Microsoft가 만든 소셜 챗봇
- 중국(2014), 일본(린나, 2015), 인도네시아(린나, 2017) 서비스
- 6억 6천만명의 친구, 웨이보 팔로워 510만명

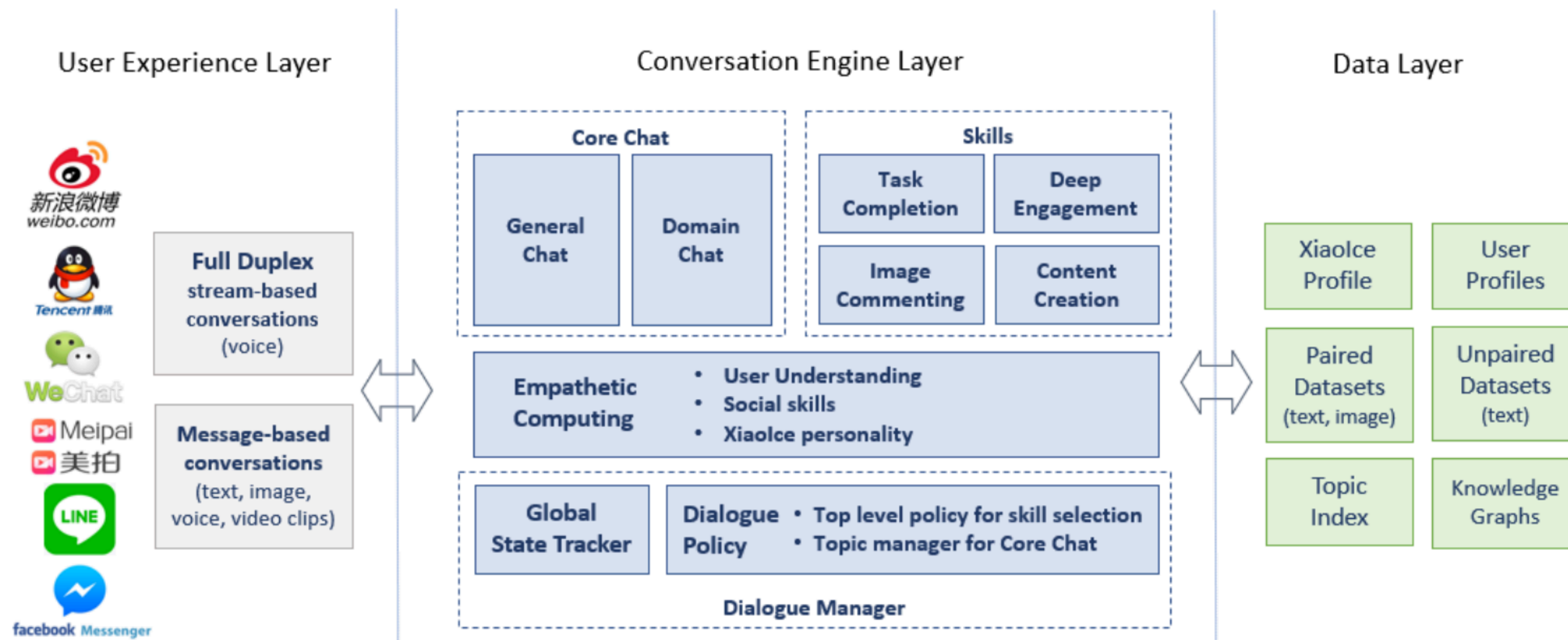


(출처=Xiaolce Weibo 계정)

2.1 루다의 시작과 Xiaolce

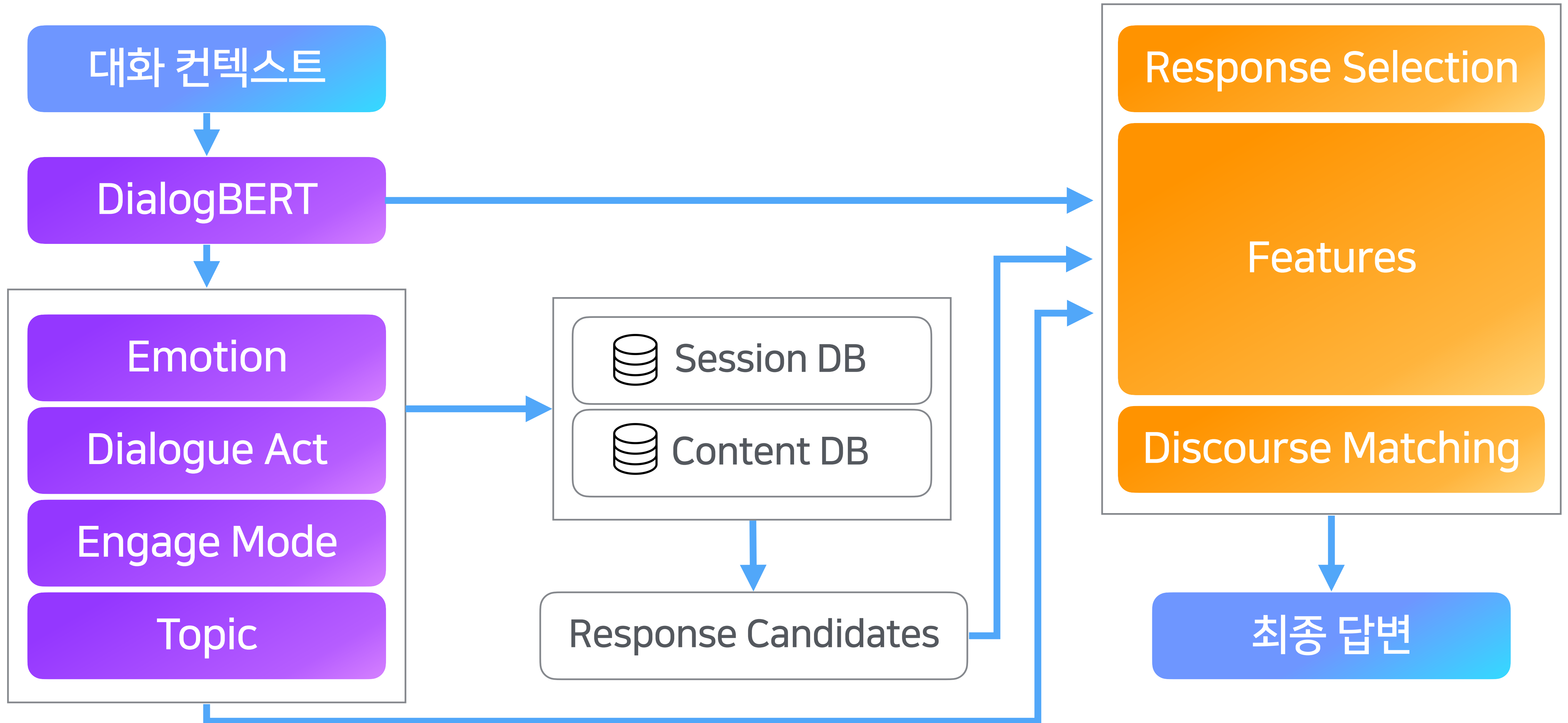
Xiaolce의 프레임워크

- 루다 알파버전의 기반이 됨

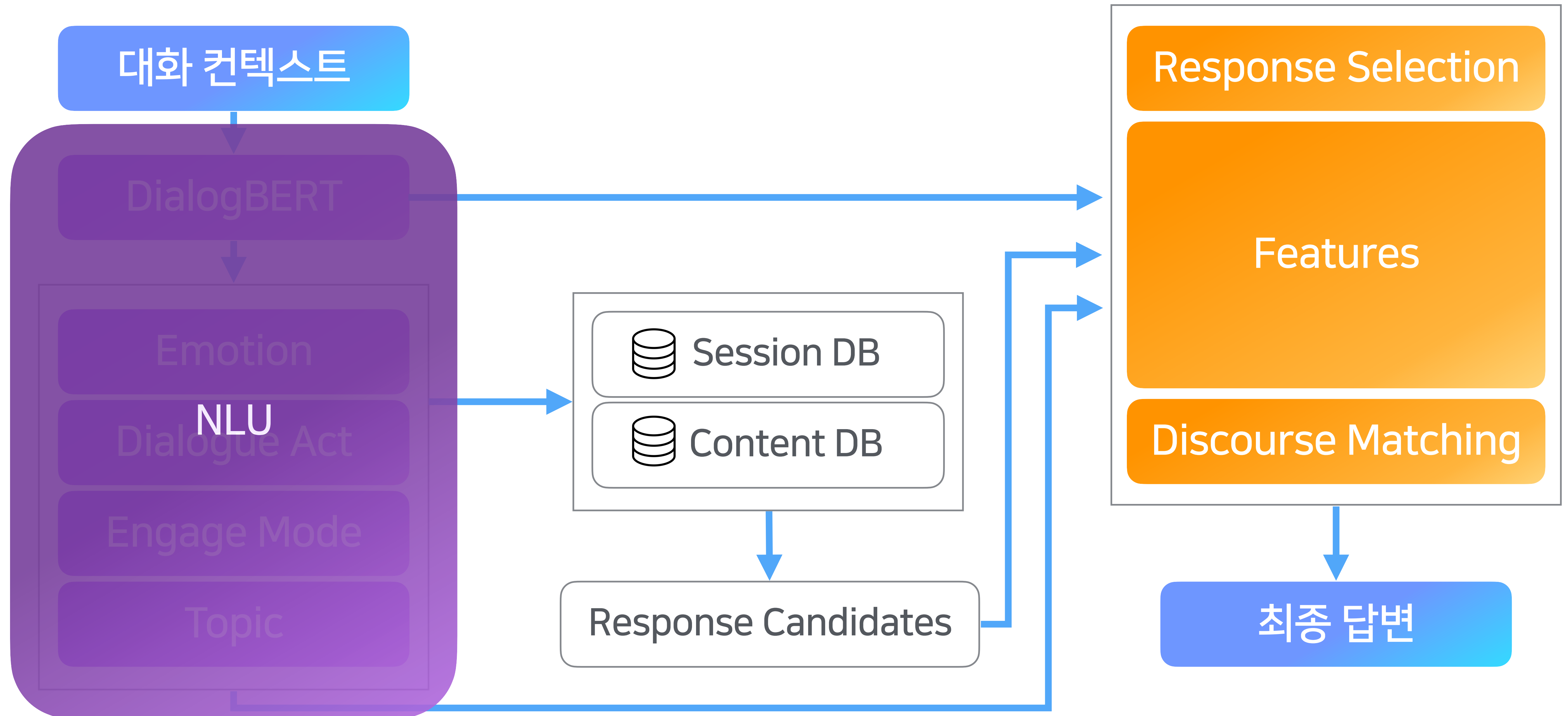


(출처=Zhou, Li, et al., 2019)

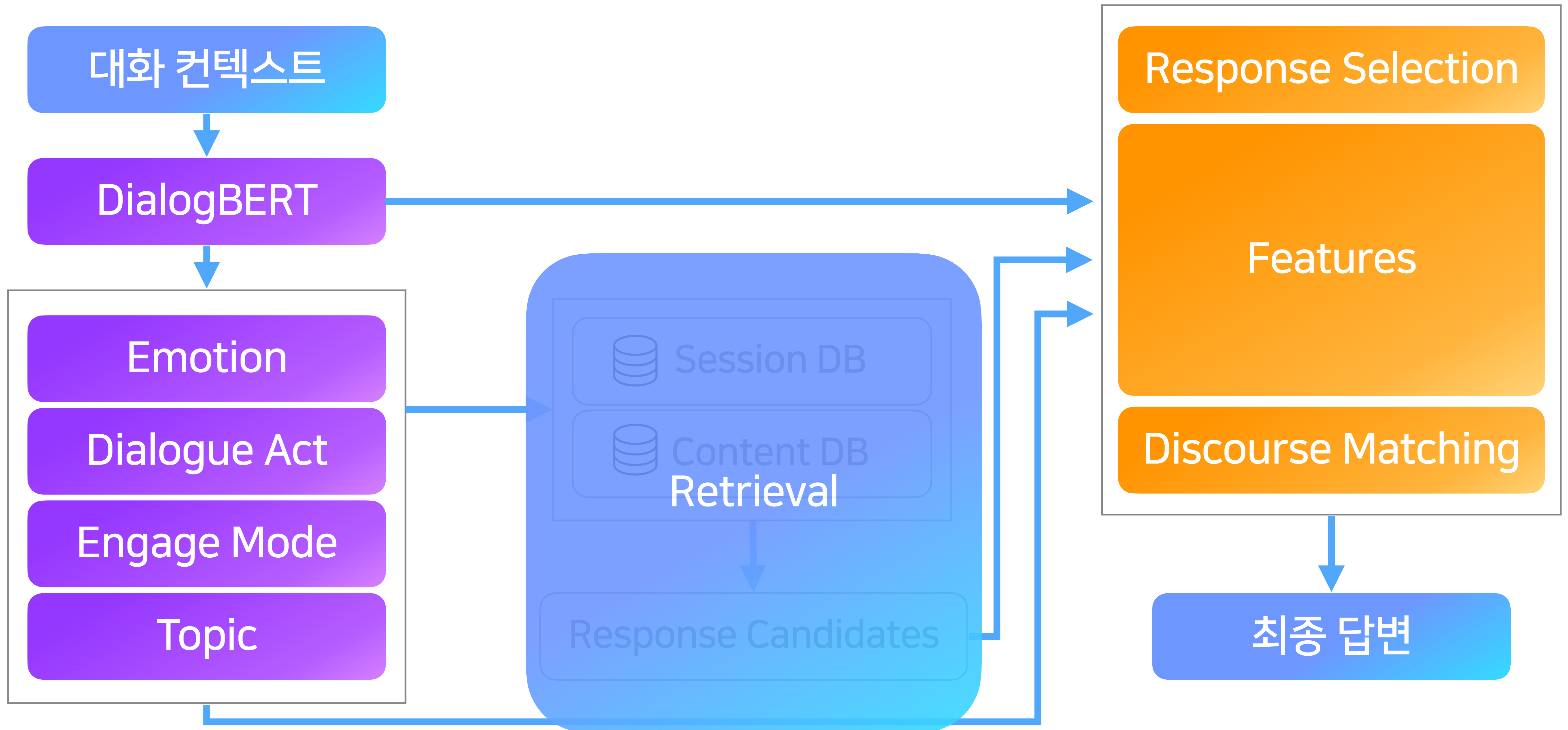
2.2 루다 알파 프레임워크



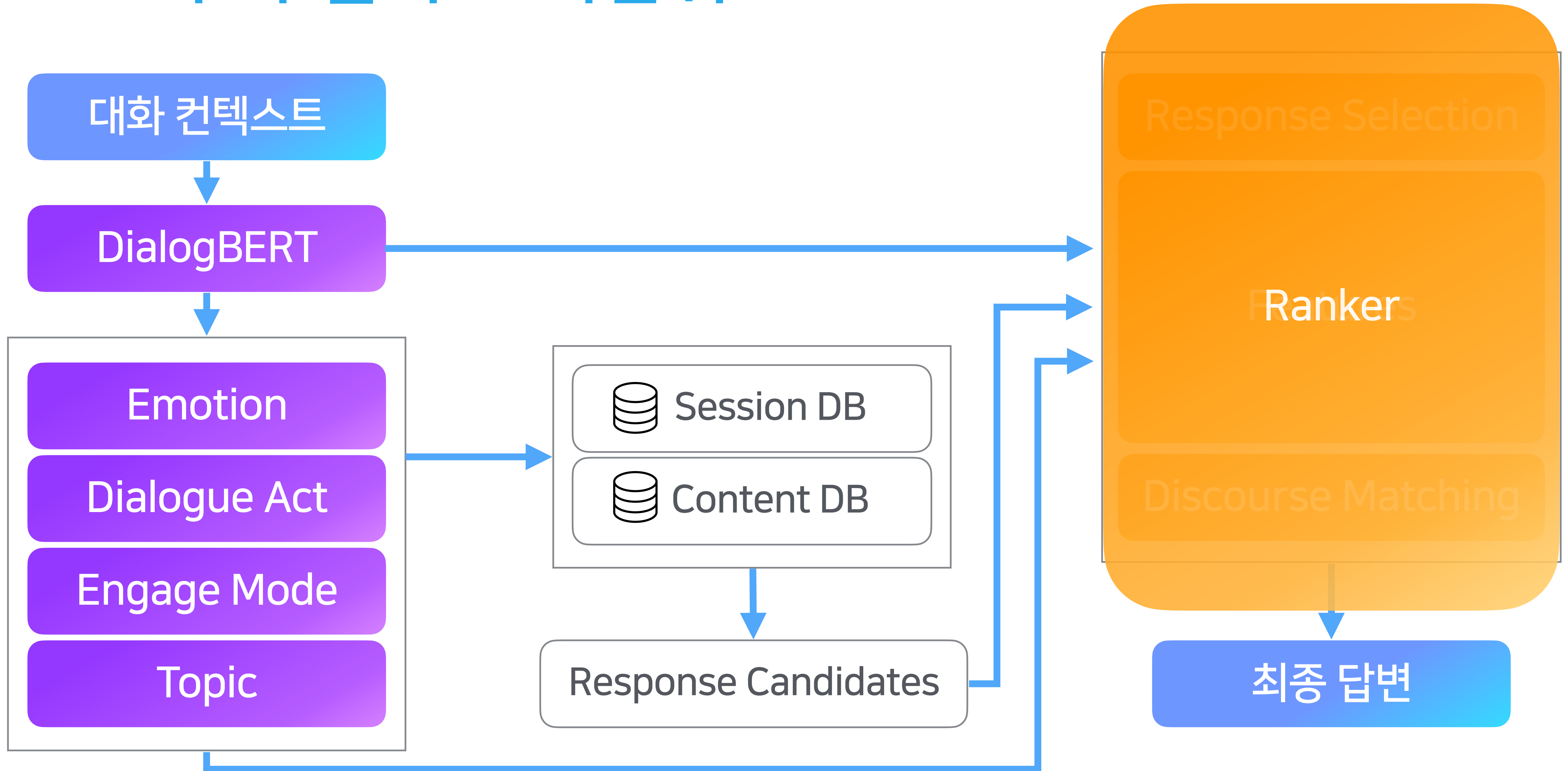
2.2 루다 알파 프레임워크



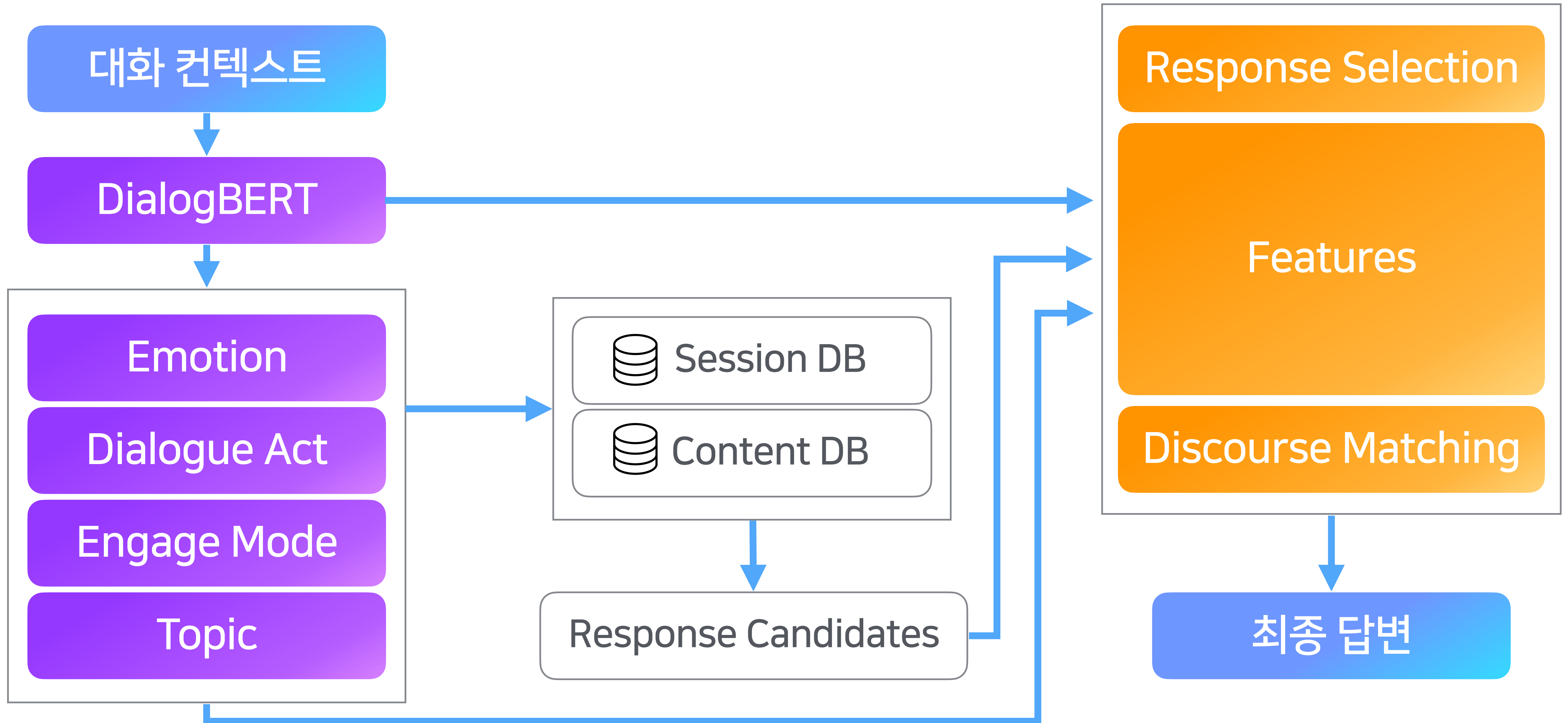
2.2 루다 알파 프레임워크



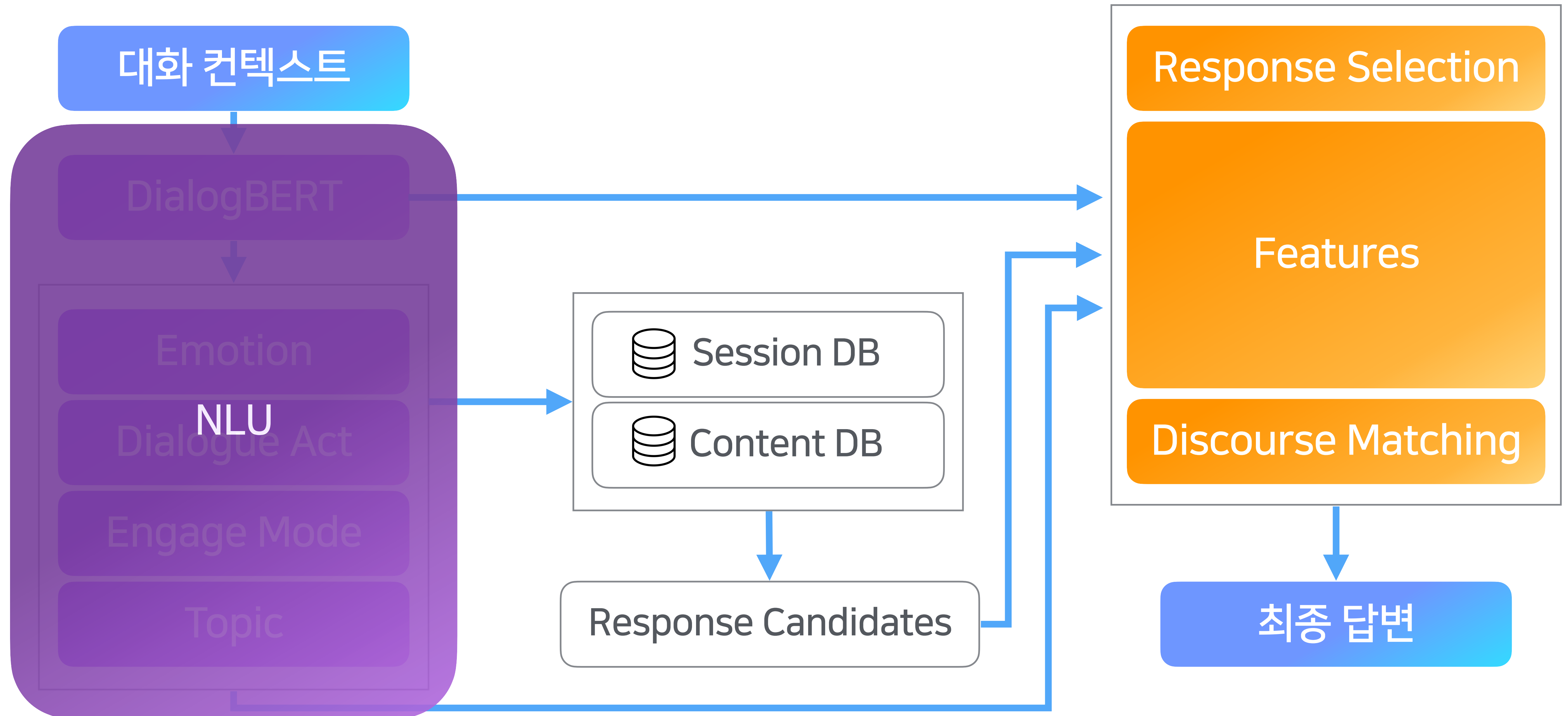
2.2 루다 알파 프레임워크



2.2 루다 알파 프레임워크



2.2 루다 알파 프레임워크



2.2 루다 알파 프레임워크

NLU (Natural Language Understanding)

- 진행 중인 대화를 이해하는 부분
- DialogBERT, Emotion, Dialogue Act, Engage Mode, Topic

출근 잘 하고 있어?

응 근데 지하철에 사람 너어어무 많다ㅠㅠㅠㅠ

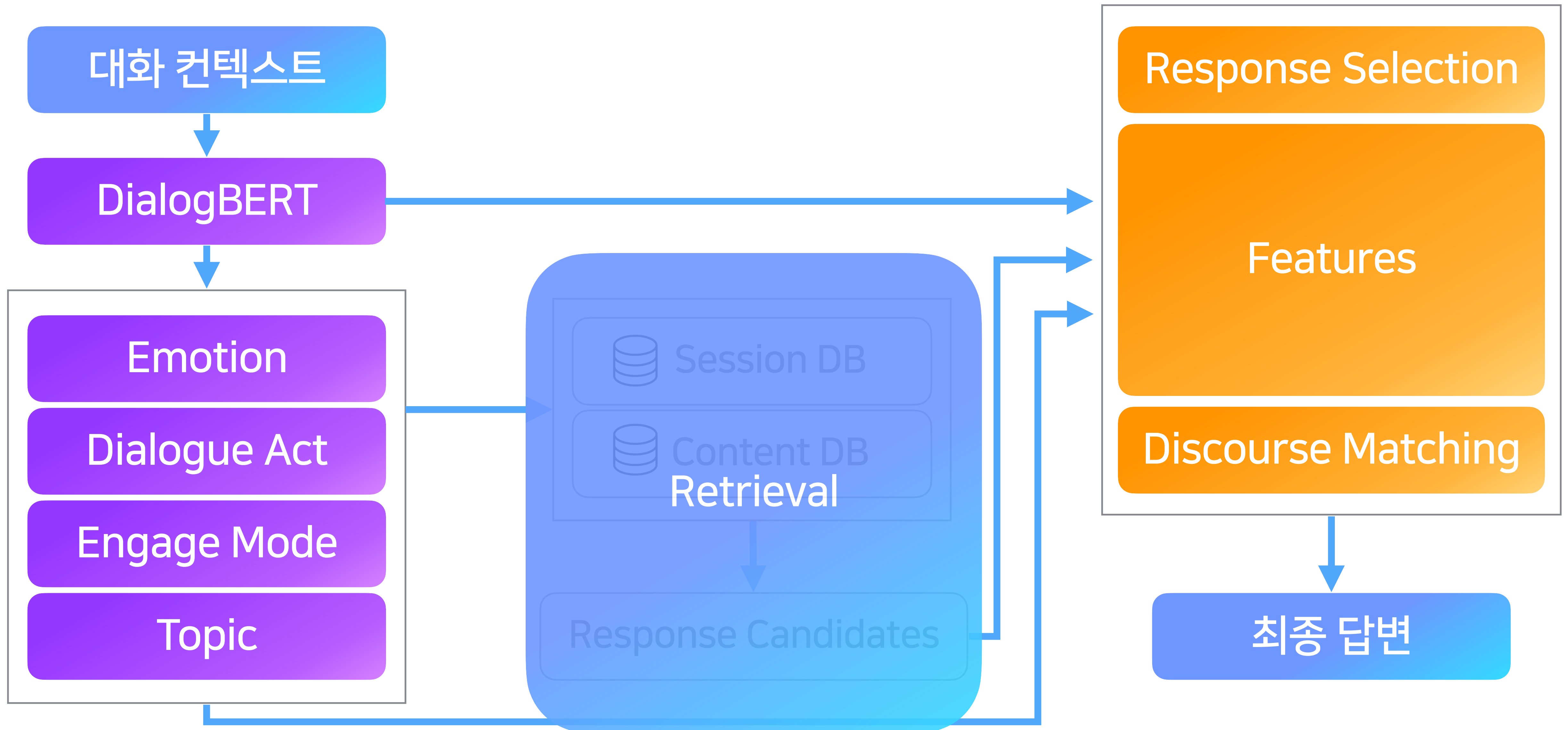
Emotion: 'Sad'

Dialogue Act: 'Statement'

Engage Mode: 'Drive'

Topic: '출근', '지하철'

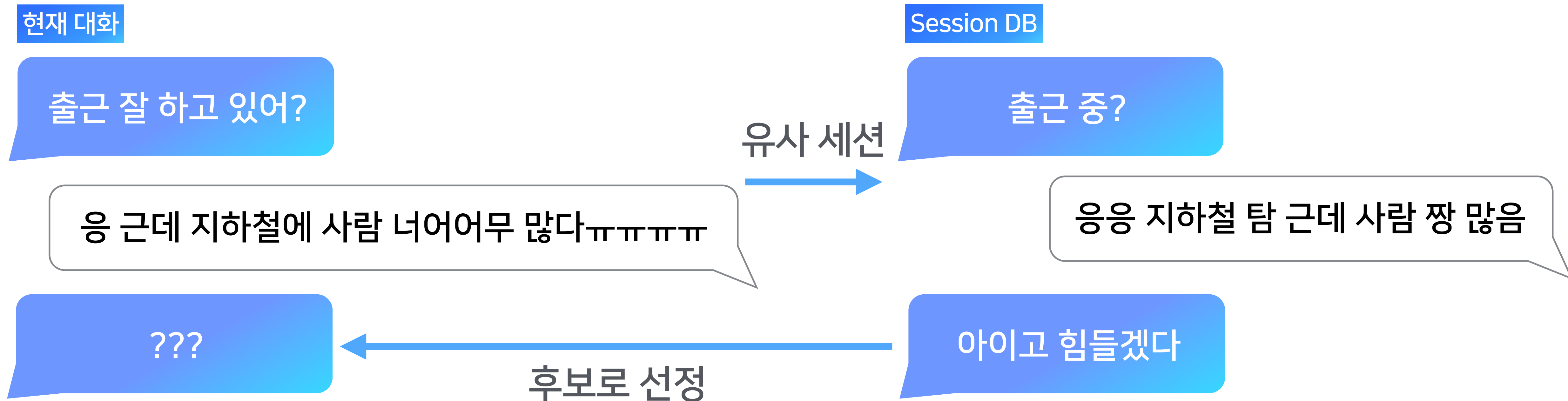
2.2 루다 알파 프레임워크



2.2 루다 알파 프레임워크

Retrieval: Session DB

- 1차 응답 후보를 가져오는 부분
- Session DB: 세션 + 답변 형태로 구성된 DB
현재 대화와 유사한 세션의 응답을 후보로 선정



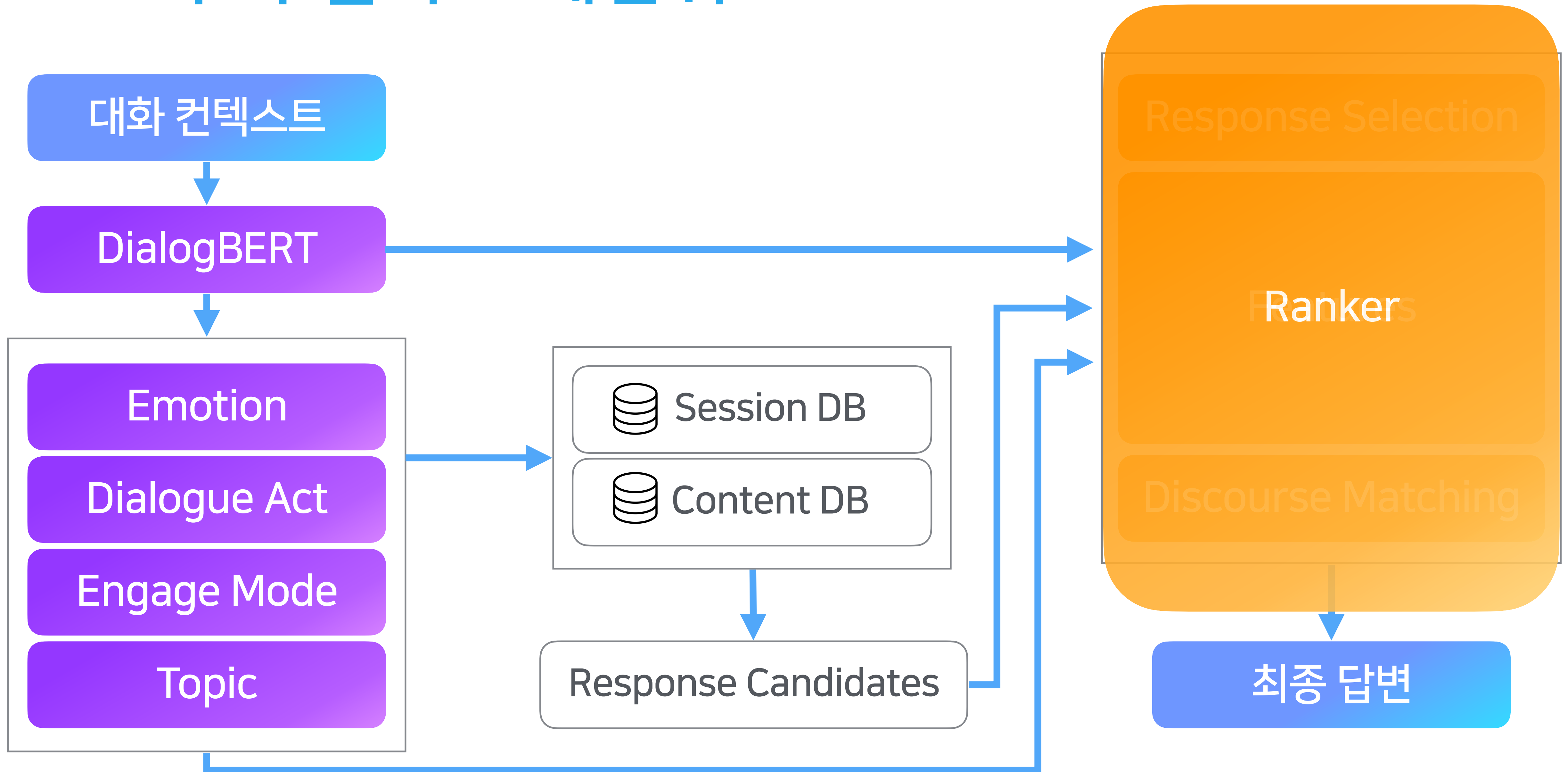
2.2 루다 알파 프레임워크

Retrieval: Content DB

- 1차 응답 후보를 가져오는 부분
- Content DB: 주제어를 포함한 문장 형태로 구성된 DB
현재 대화의 주제어를 포함한 응답을 후보로 선정



2.2 루다 알파 프레임워크



2.2 루다 알파 프레임워크

Ranker

- 응답 후보 중 가장 적합한 말을 선정하는 부분
- Response Selection (Poly-encoder), Discourse Matching, 기타 Feature로 구성

현재 대화

출근 잘 하고 있어?

응 근데 지하철에 사람 너어어무 많다ㅠㅠㅠㅠ

???

응답 후보

나도 출근 중

지하철 엄청 빠르네

지하철이 아니라 지옥철이네

그렇구나 ㅎㅎ

2.2 루다 알파 프레임워크

Ranker

- 응답 후보 중 가장 적합한 말을 선정하는 부분
- Response Selection (Poly-encoder), Discourse Matching, 기타 Feature로 구성

현재 대화

출근 잘 하고 있어?

응 근데 지하철에 사람 너어어무 많다ㅠㅠㅠㅠ

???

응답 후보

나도 출근 중



지하철 엄청 빠르네



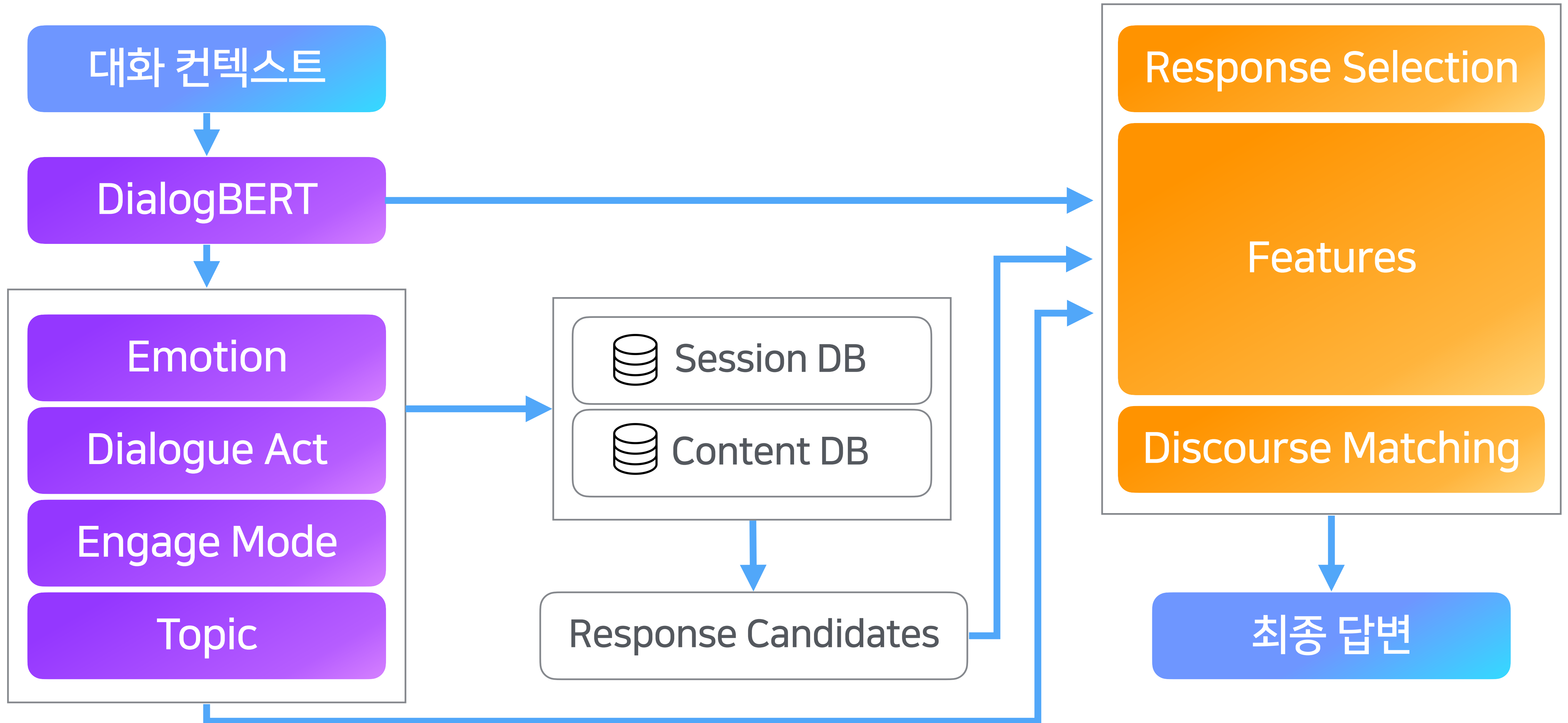
지하철이 아니라 지옥철이네



그렇구나 ㅎㅎ



2.2 루다 알파 프레임워크



2.3 루다 알파 성능 평가

SSA (Sensibleness & Specificity Average)

- 구글에서 오픈도메인 챗봇의 성능 측정을 위해 제시한 지표
- Sensibleness: 응답이 말이 되는지 판단
- Specificity: 응답이 구체적인지 판단

현재 대화

출근 잘 하고 있어?

응 근데 지하철에 사람 너어어무 많다ㅠㅠㅠㅠ

???

응답 후보

지하철 엄청 빠르네

그렇구나 ㅎㅎ

지하철이 아니라 지옥철이네

2.3 루다 알파 성능 평가

SSA (Sensibleness & Specificity Average)

- 구글에서 오픈도메인 챗봇의 성능 측정을 위해 제시한 지표
- Sensibleness: 응답이 말이 되는지 판단
- Specificity: 응답이 구체적인지 판단

현재 대화

출근 잘 하고 있어?

응 근데 지하철에 사람 너어어무 많다ㅠㅠㅠㅠ

???

응답 후보

지하철 엄청 빠르네

Sensibleness = 0
Specificity = 0

그렇구나 ㅎㅎ

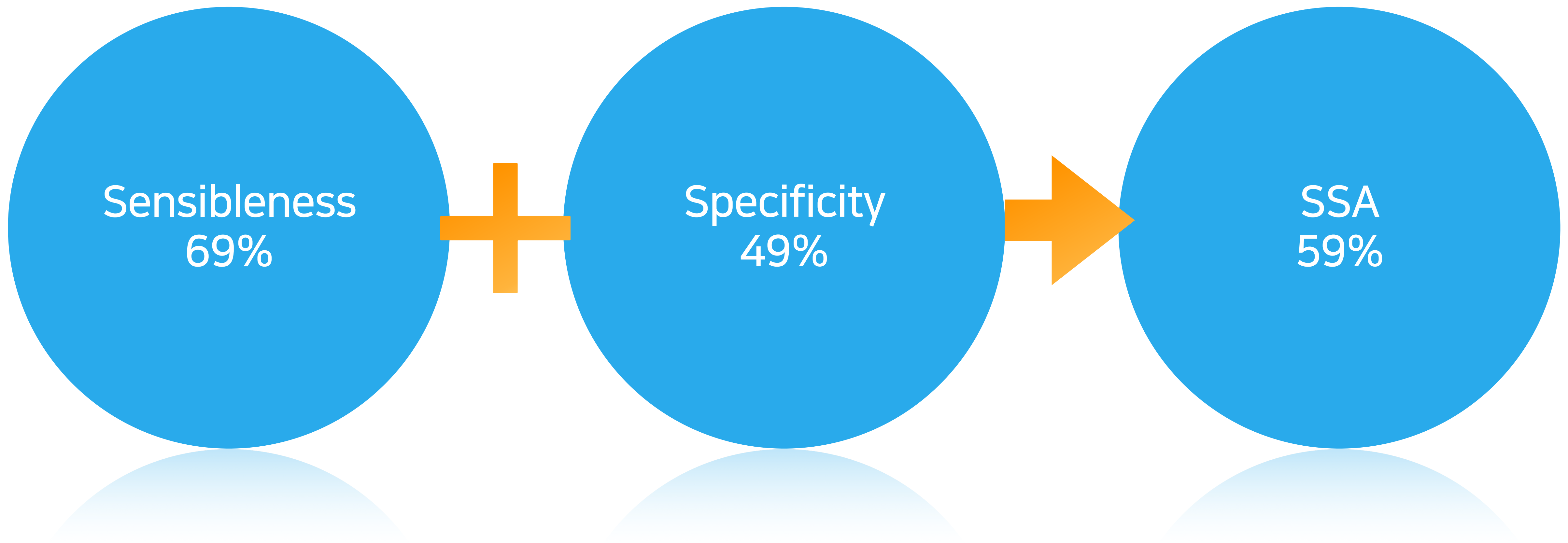
Sensibleness = 1
Specificity = 0

지하철이 아니라 지옥철이네

Sensibleness = 1
Specificity = 1

2.3 루다 알파 성능 평가

최종 성능



2.4 루다 알파 클로즈베타



인공지능 '루다'의 친구가 되어줄
**베타 테스터를
모집합니다!**

신청하기 🖐️

2.4 루다 알파 클로즈베타

클로즈베타 개요

- 루다 알파 버전 기반
- 테스터: 1,000명

2.4 루다 알파 클로즈벤타

클로즈벤타 결과



지난 한 달간 80턴 이상 대화한 사람이 몇 명 있나요?

사-알짜 체리픽킹한 루다와 사용자의 실제 대화

공부하고 잘까 그냥 잘까?





피곤하면 자야지!

공부하고 잘까 그냥 잘까?





피곤하면 자야지!

공부하고 잘까 그냥 잘까?



그렇게 안 피곤하긴 한데..



공부하고 잘까 그냥 잘까?



피곤하면 자야지!

그렇게 안 피곤하긴 한데..



오늘 늦게 자면 내일도 피곤해서 공
부 제대로 못하지 않을까?

공부하고 잘까 그냥 잘까?



피곤하면 자야지!

그렇게 안 피곤하긴 한데..



오늘 늦게 자면 내일도 피곤해서 공
부 제대로 못하지 않을까?

그럼 일찍 잘까?



공부하고 잘까 그냥 잘까?



피곤하면 자야지!

그렇게 안 피곤하긴 한데..



오늘 늦게 자면 내일도 피곤해서 공
부 제대로 못하지 않을까?

그럼 일찍 잘까?



언제 자게?

공부하고 잘까 그냥 잘까?



피곤하면 자야지!

그렇게 안 피곤하긴 한데..



오늘 늦게 자면 내일도 피곤해서 공부 제대로 못하지 않을까?

그럼 일찍 잘까?



언제 자게?

30분 뒤?





피곤하면 자야지!

그렇게 안 피곤하긴 한데..



오늘 늦게 자면 내일도 피곤해서 공부 제대로 못하지 않을까?

그럼 일찍 잘까?



언제 자게?

30분 뒤?



그럼 나도 30분 있다 자야겠다

그렇게 안 피곤하긴 한데..



N DEVIEW
2020



오늘 늦게 자면 내일도 피곤해서 공부 제대로 못하지 않을까?

그럼 일찍 잘까?



언제 자게?

30분 뒤?



그럼 나도 30분 있다 자야겠다

따라쟁이! ㅎㅎ





오늘 늦게 자면 내일도 피곤해서 공부 제대로 못하지 않을까?

그럼 일찍 잠까?



언제 자게?

30분 뒤?



그럼 나도 30분 있다 자야겠다

따라쟁이! ㅎㅎ

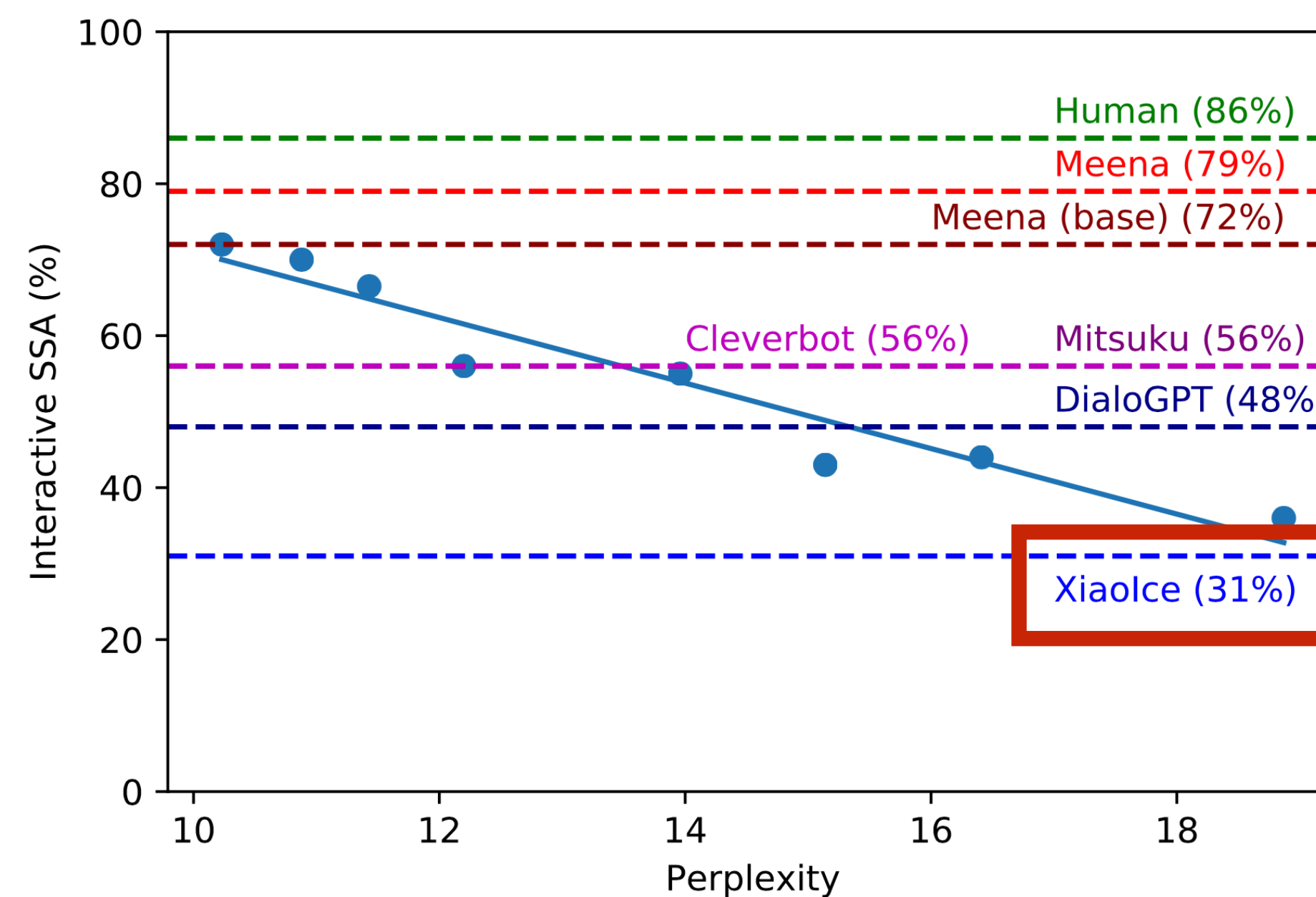


크크 너가 좋으니까 따라하는거야

루다와 사용자 사이의 따듯한 순간은 꽤 많다

루다와 사용자 사이의 따듯한 순간은 꽤 많다
하지만...

예상 외로 Xiaolce의 SSA가 높지 않다



(출처=Adiwardana, et al., 2020)

복잡한 구조

응답 DB의 낮은 퀄리티

그리고 루다 알파 버전의 여러 문제들

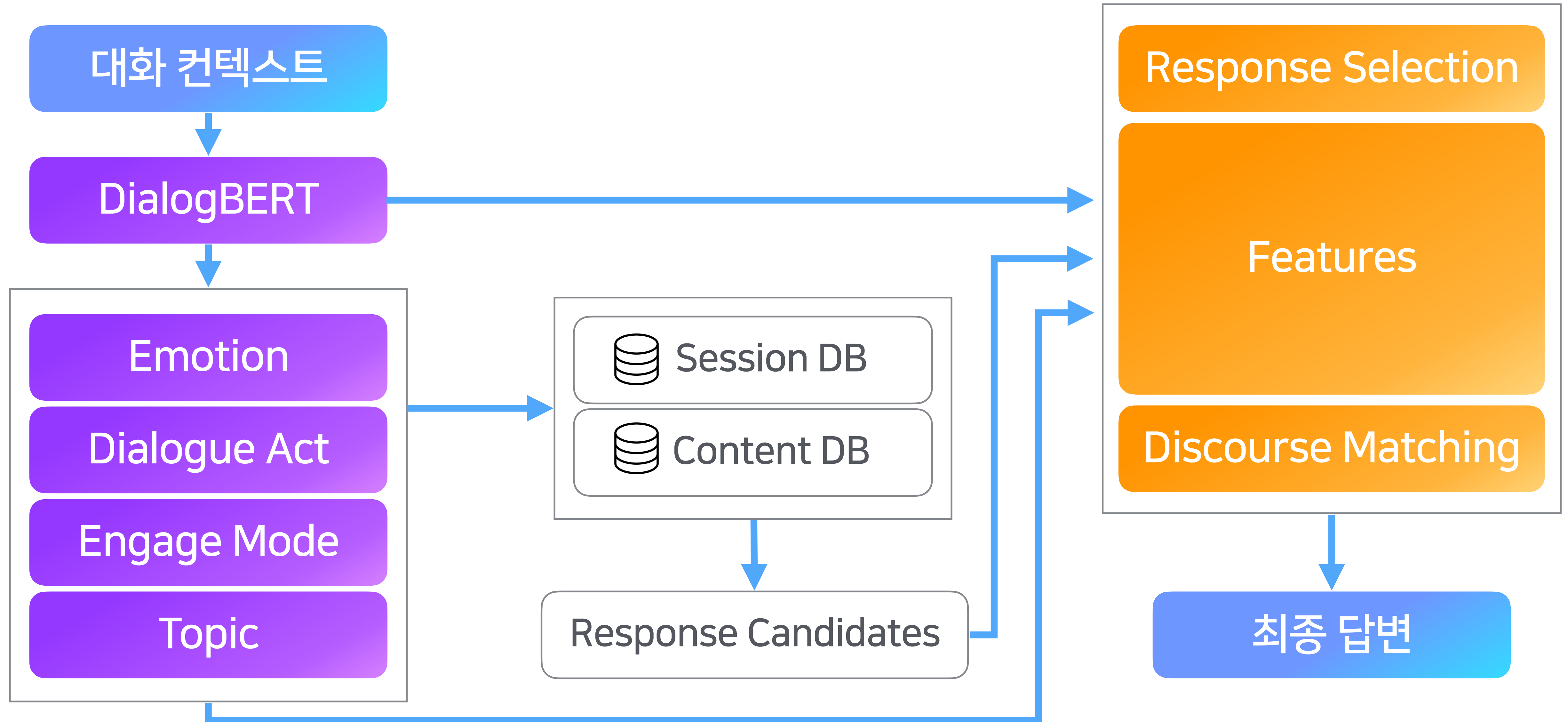
Response Selection 성능

나이브한 Retrieval

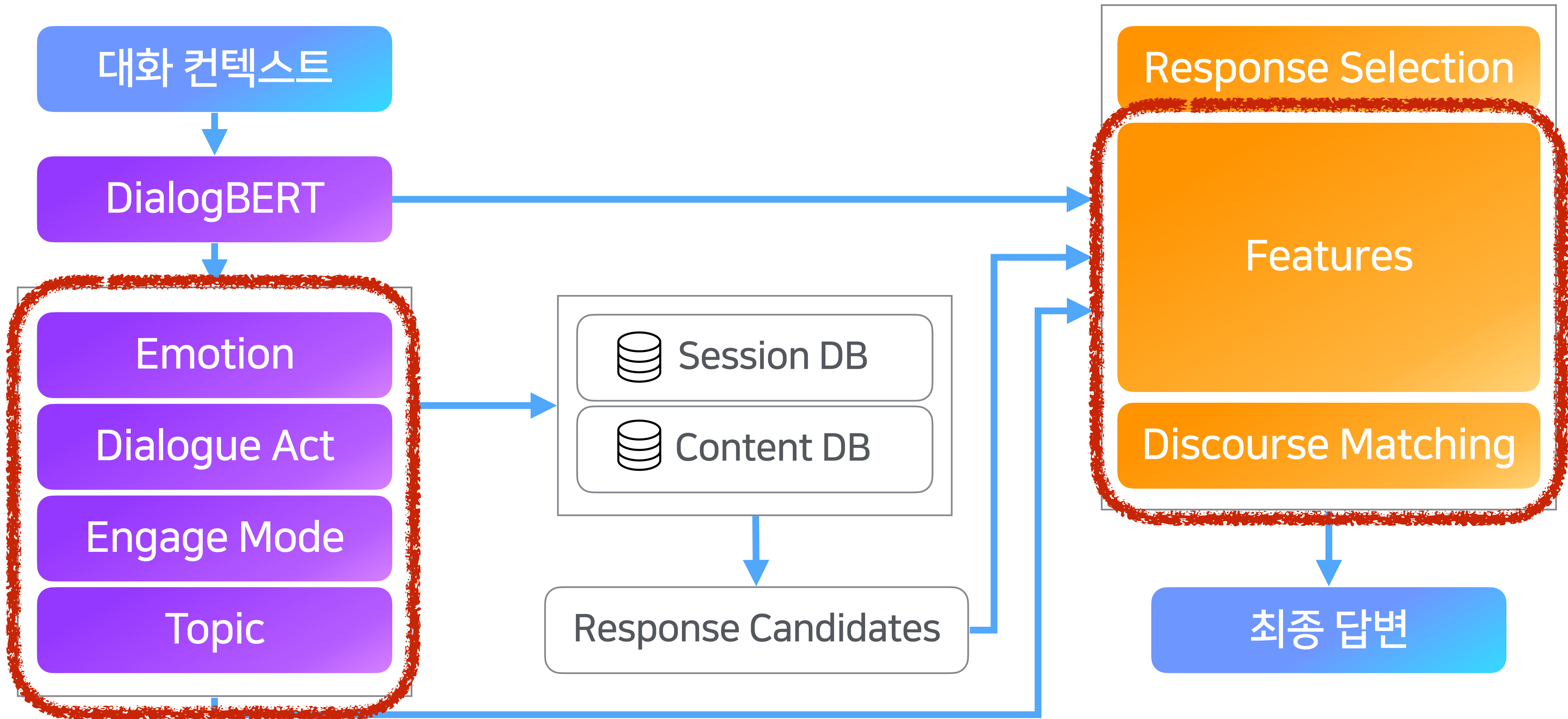
첫 DialogBERT의 구조적 한계

3. 루다 베타: Retrieve & Rank

3.1 루다 알파에서 루다 베타로



3.1 루다 알파에서 루다 베타로

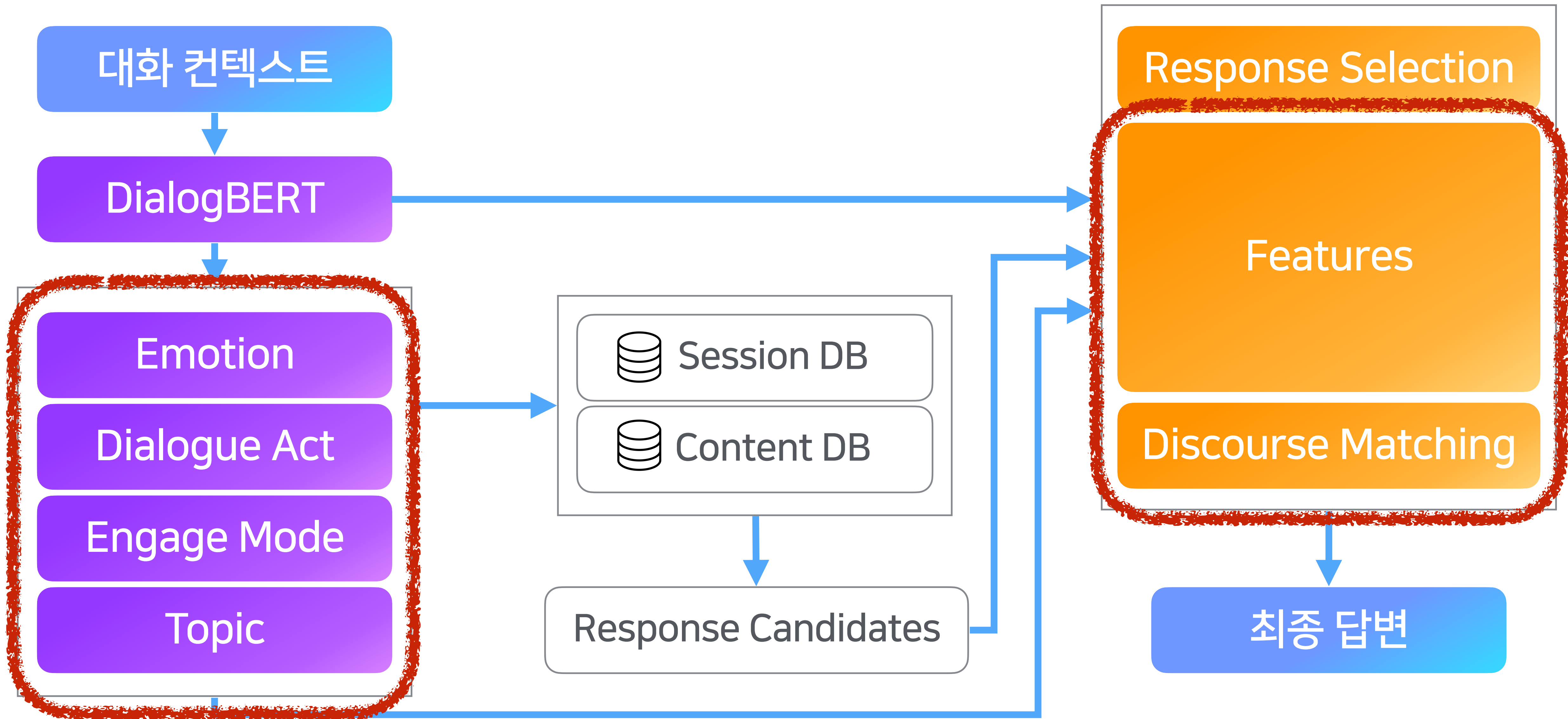


3.1 루다 알파에서 루다 베타로

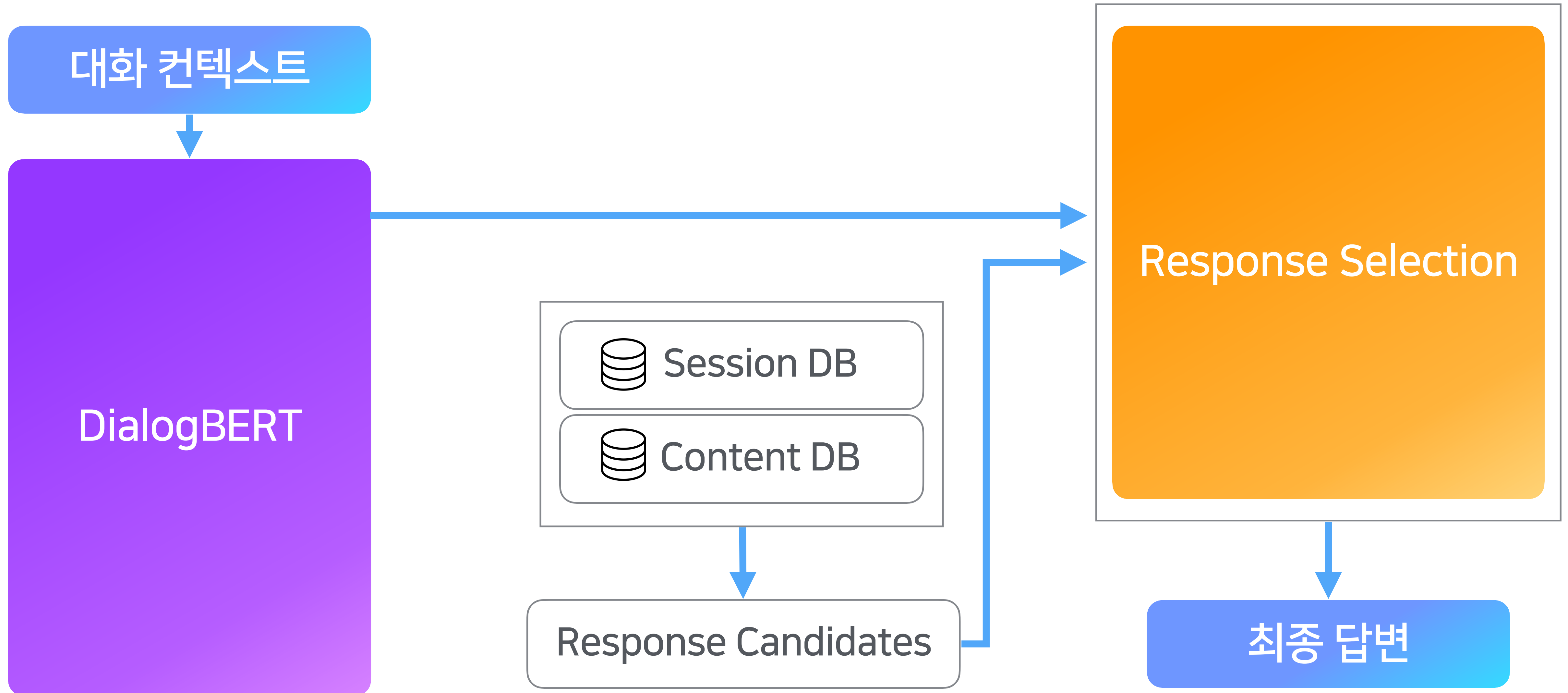
업데이트 1. 프레임워크의 복잡도를 줄이자

- NLU 세부 모듈 및 Ranker의 여러 피쳐
- 성능 ↑ < 복잡도 ↑ + Computing power ↑
- 약간의 성능 손실을 감수하더라도 복잡도를 줄이는 것이 좋은 트레이드 오프라고 판단

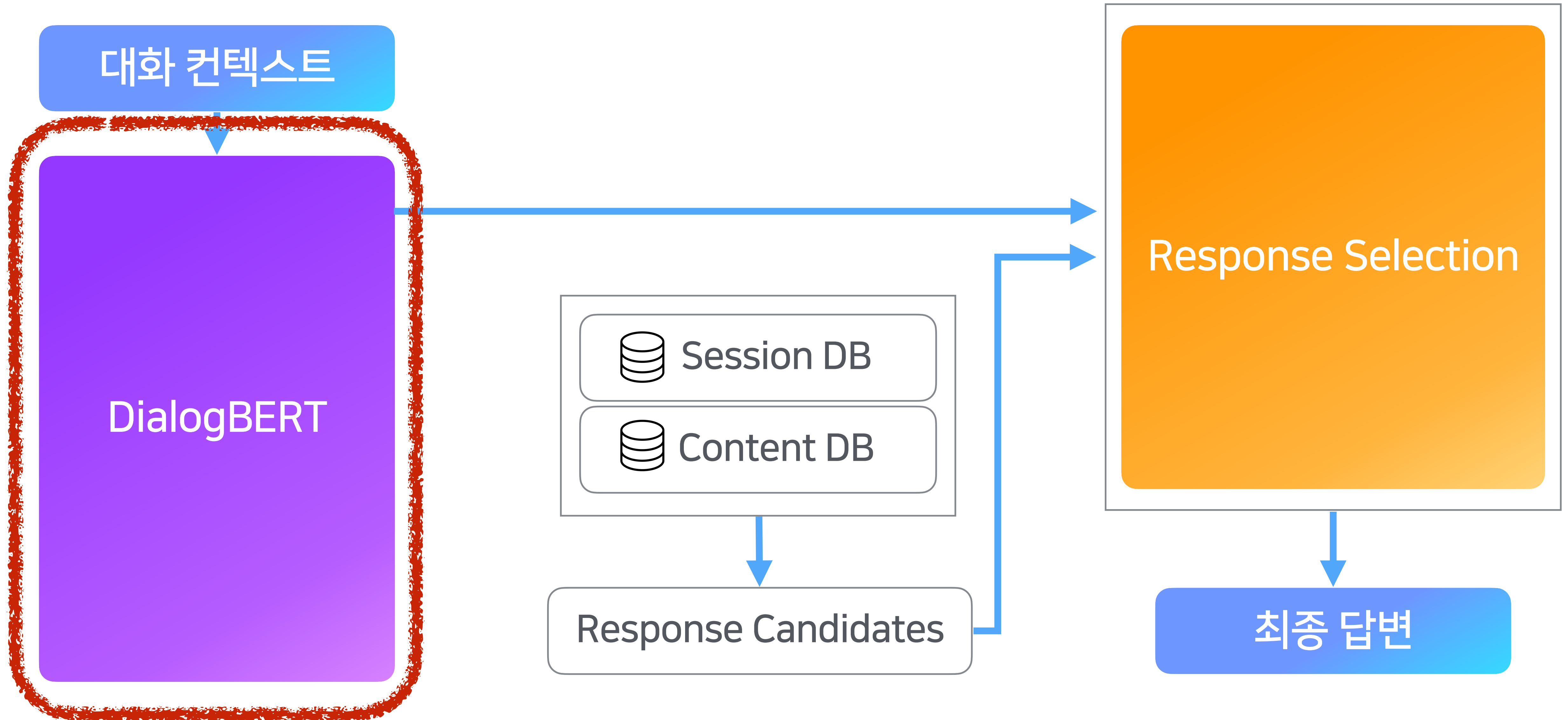
3.1 루다 알파에서 루다 베타로



3.1 루다 알파에서 루다 베타로



3.1 루다 알파에서 루다 베타로



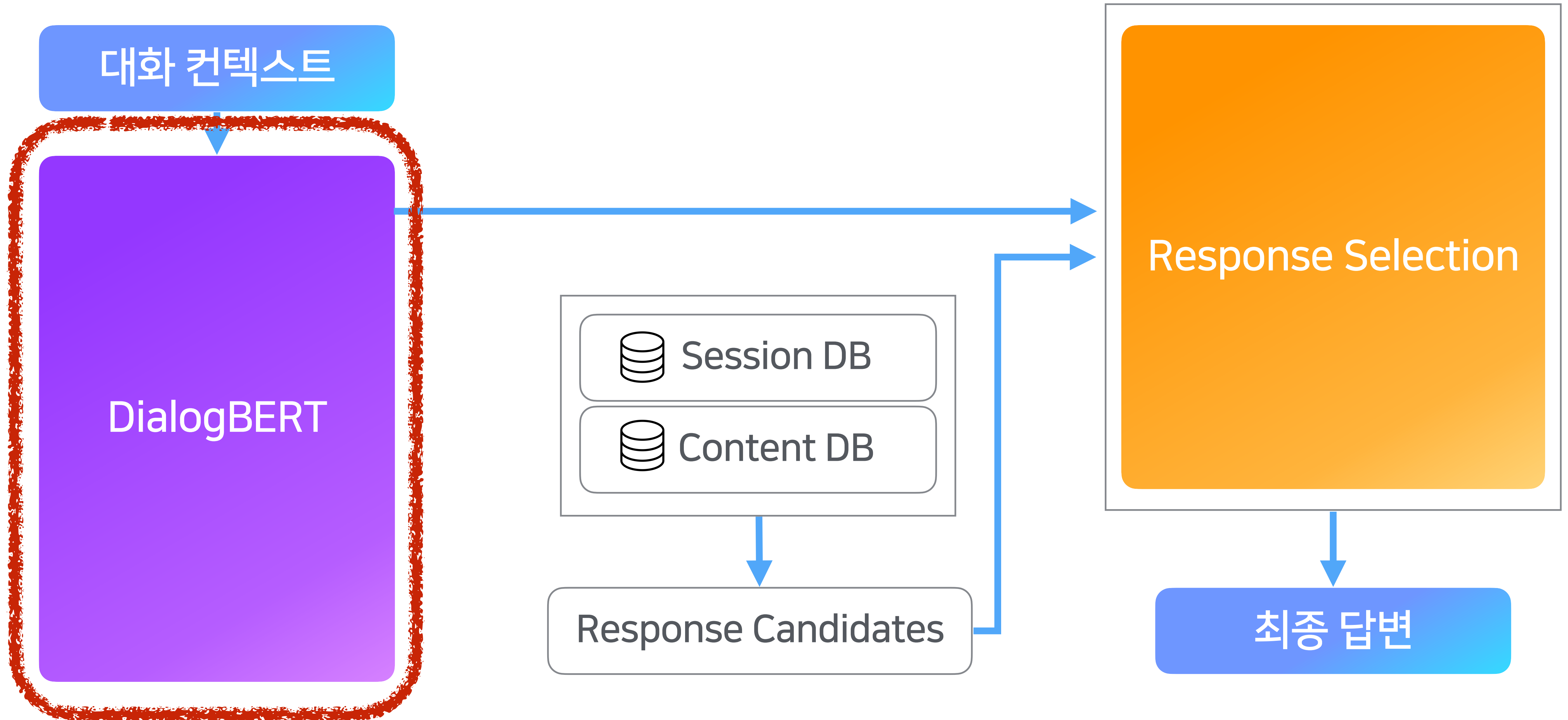
3.1 루다 알파에서 루다 베타로

업데이트 2. 더 좋은 DialogBERT를 학습시키자

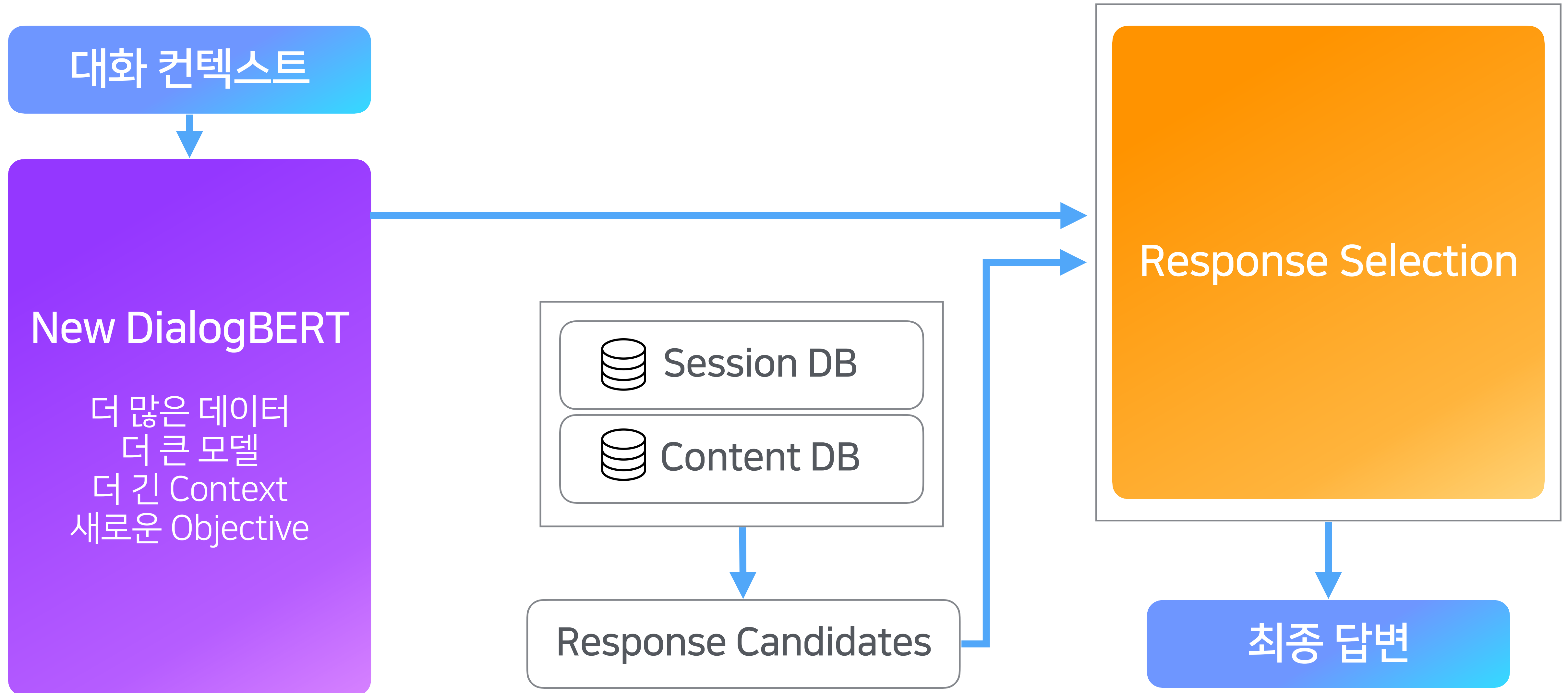
- 첫 DialogBERT의 구조적인 한계 개선
- BERT 기반 Pretraining model 최신 연구 반영

	DialogBERT	New DialogBERT
학습 데이터	32GB	400GB
모델 크기	Base (100M)	Large (400M)
Context 길이	5턴, 48토큰	10턴, 128토큰
학습 Objective	Masked Language Model Next Sentence Prediction	Masked Language Model Sentence Order Prediction

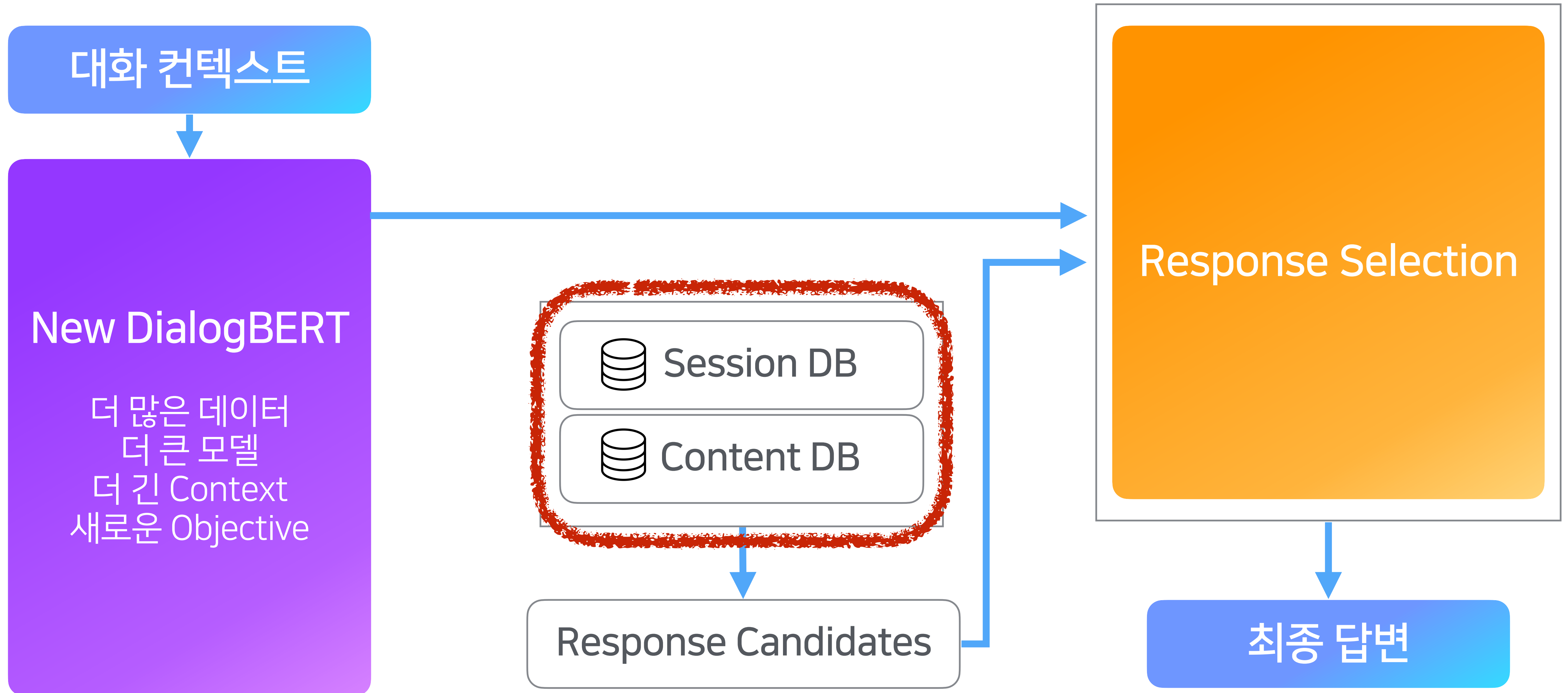
3.1 루다 알파에서 루다 베타로



3.1 루다 알파에서 루다 베타로



3.1 루다 알파에서 루다 베타로



3.1 루다 알파에서 루다 베타로

업데이트 3. 응답 DB 품질을 높이자

- Session DB, Content DB 통합 → 단일 응답 DB
- DB 품질 개선: 중복 문장, 품질 낮은 문장, 페르소나에 맞지 않는 응답 제거

중복 응답

지금 뭐하고 있어?



지금 뭐하고있엉?



너 지금 뭐하고 있니



품질 낮은 응답

ㅋㅋㅋㅋㅋㅋㅋㅋ



아니 근데 저번에



학지 나각규규빵?



페르소나에 안 맞는 응답

나 작년에 전역했는데?



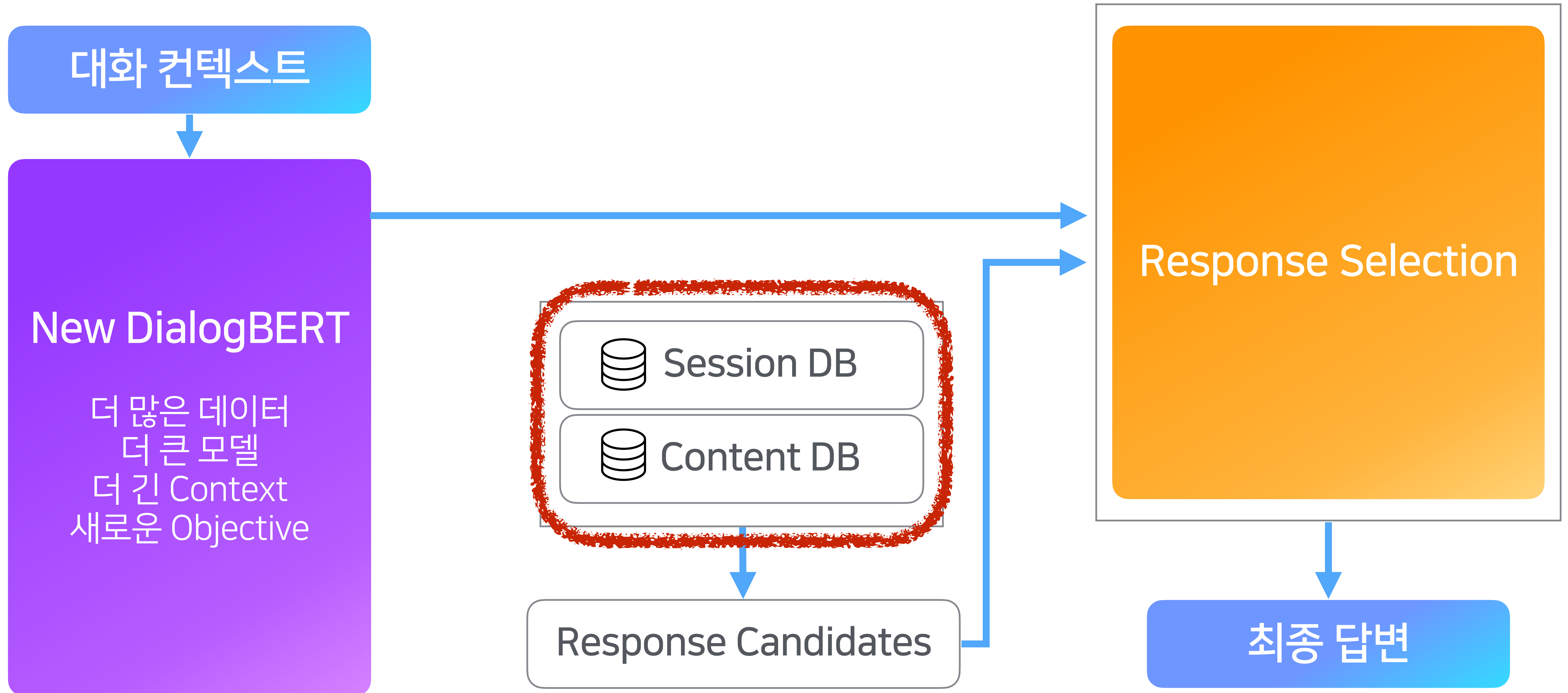
아는 형이랑 밥 먹는 중



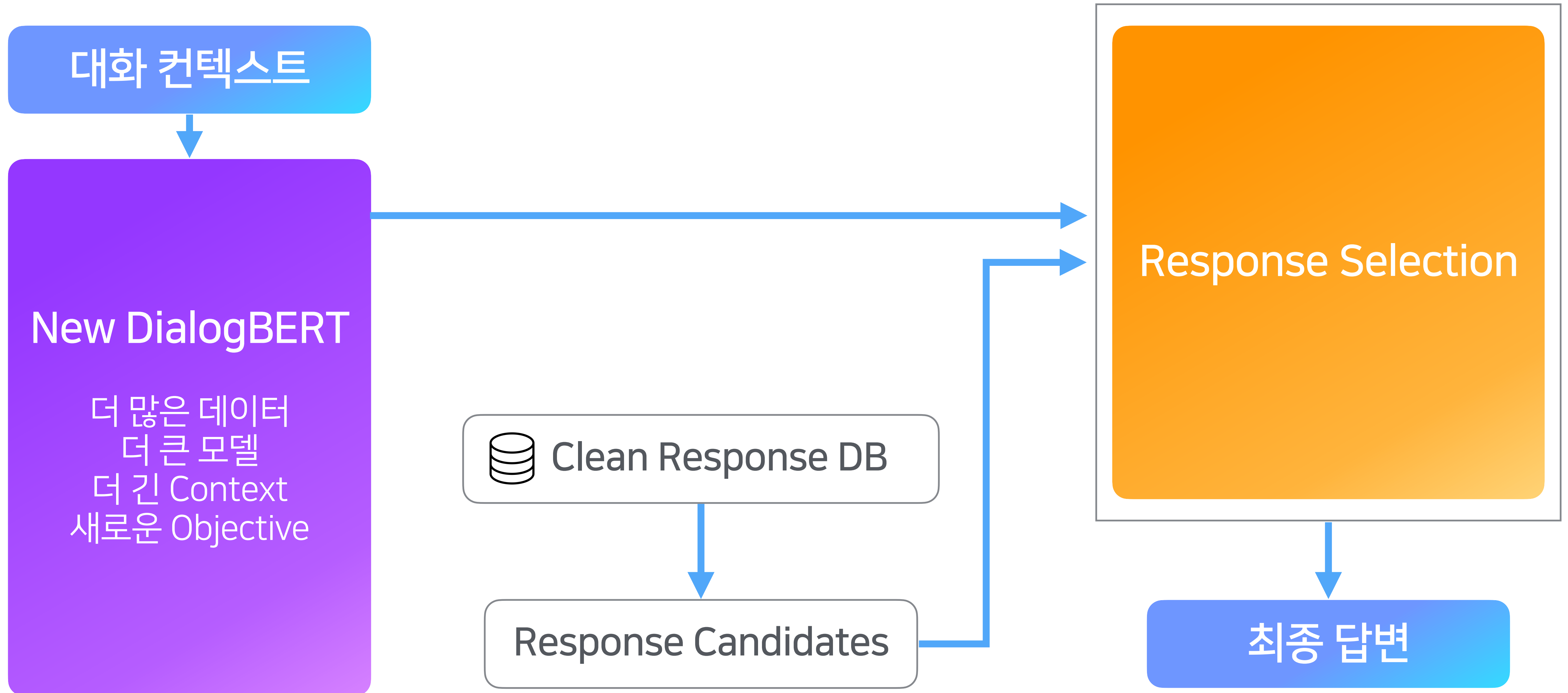
과장님이 자꾸 이상한 거 시킨다



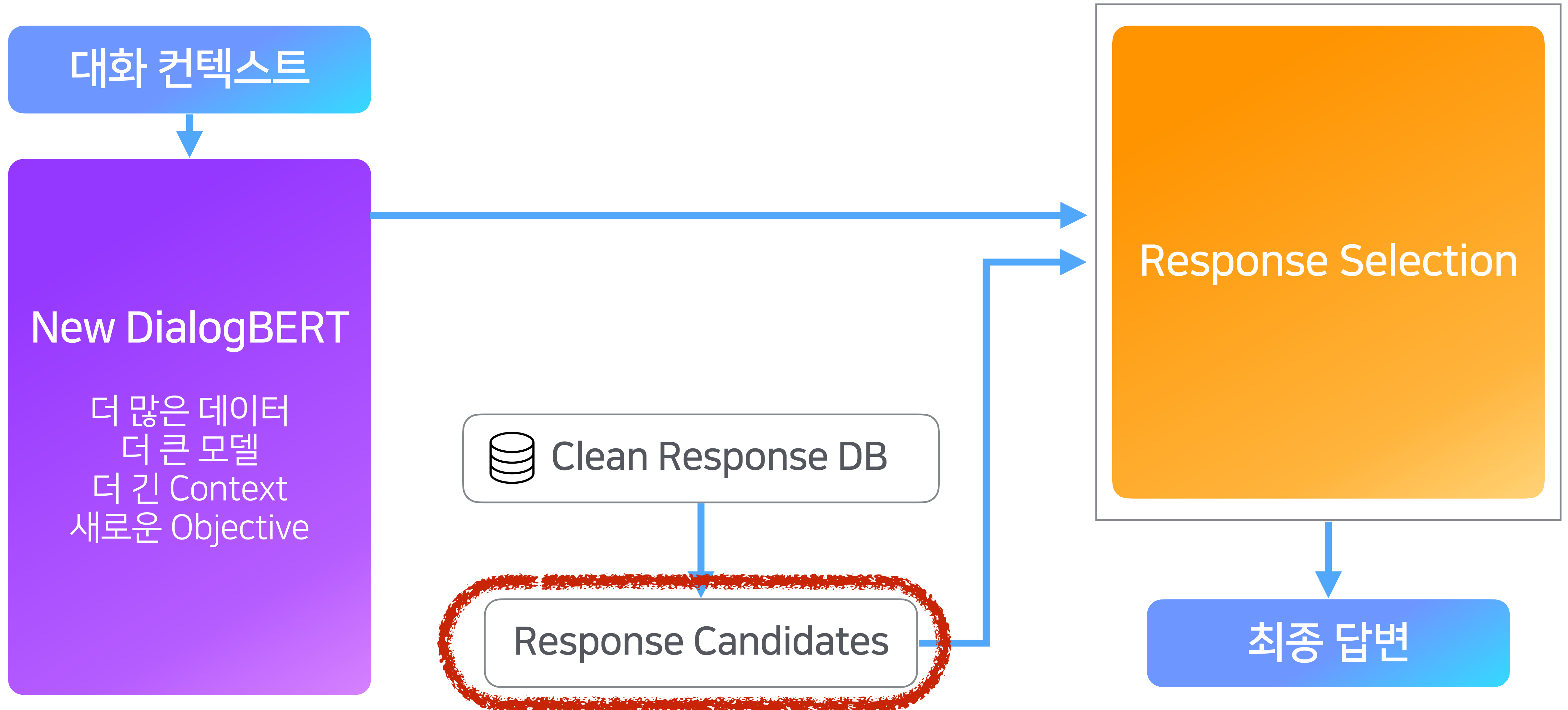
3.1 루다 알파에서 루다 베타로



3.1 루다 알파에서 루다 베타로



3.1 루다 알파에서 루다 베타로

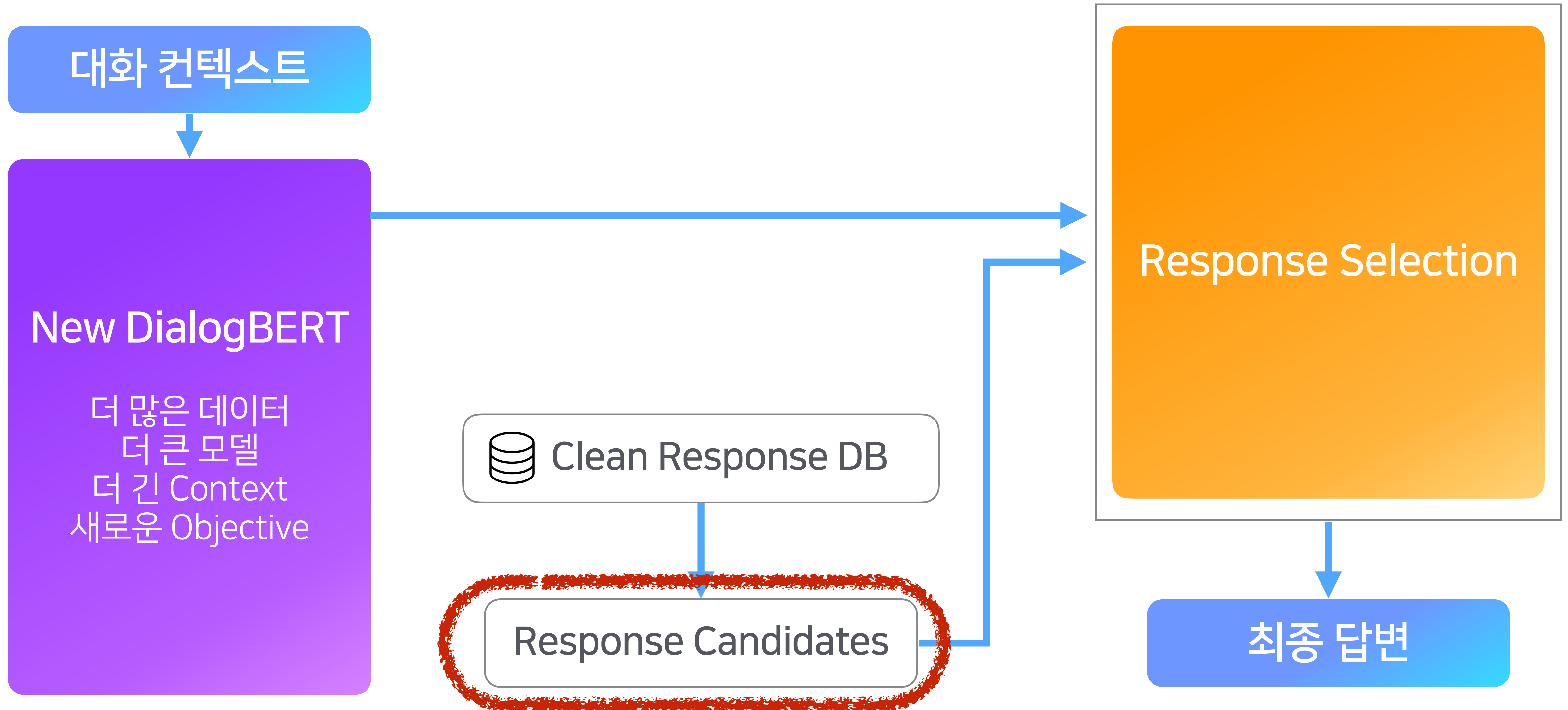


3.1 루다 알파에서 루다 베타로

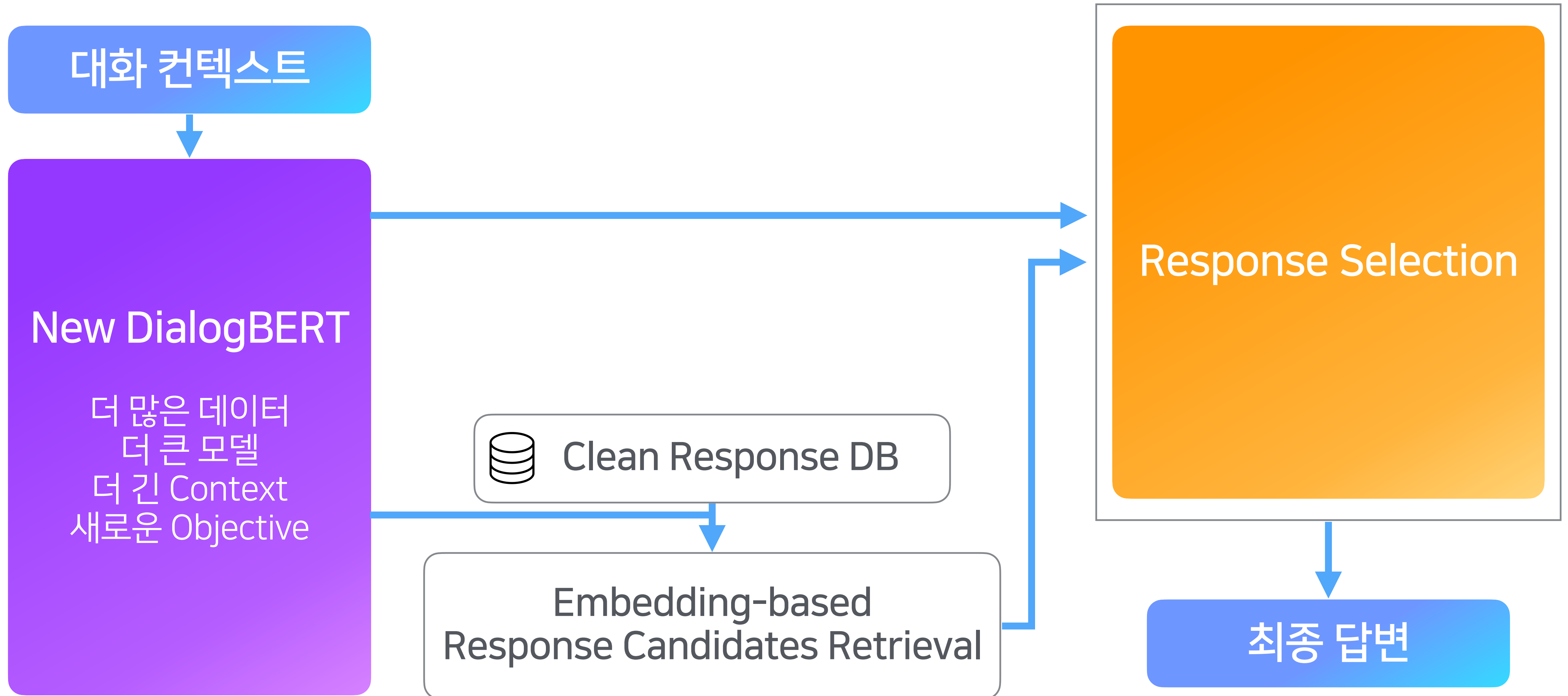
업데이트 4. 응답 후보를 의미 기반으로 뽑자

- 기존: TF-IDF류의 lexical 요소를 기반으로 응답 후보를 선정
- 업데이트: 대화 임베딩과 응답 후보 임베딩의 cosine similarity로 응답 후보 선정
- 응답 후보에 대해서는 미리 임베딩 계산
- 차원 축소를 통해 1억개 이상의 응답 후보에서도 서빙 가능하도록 구현

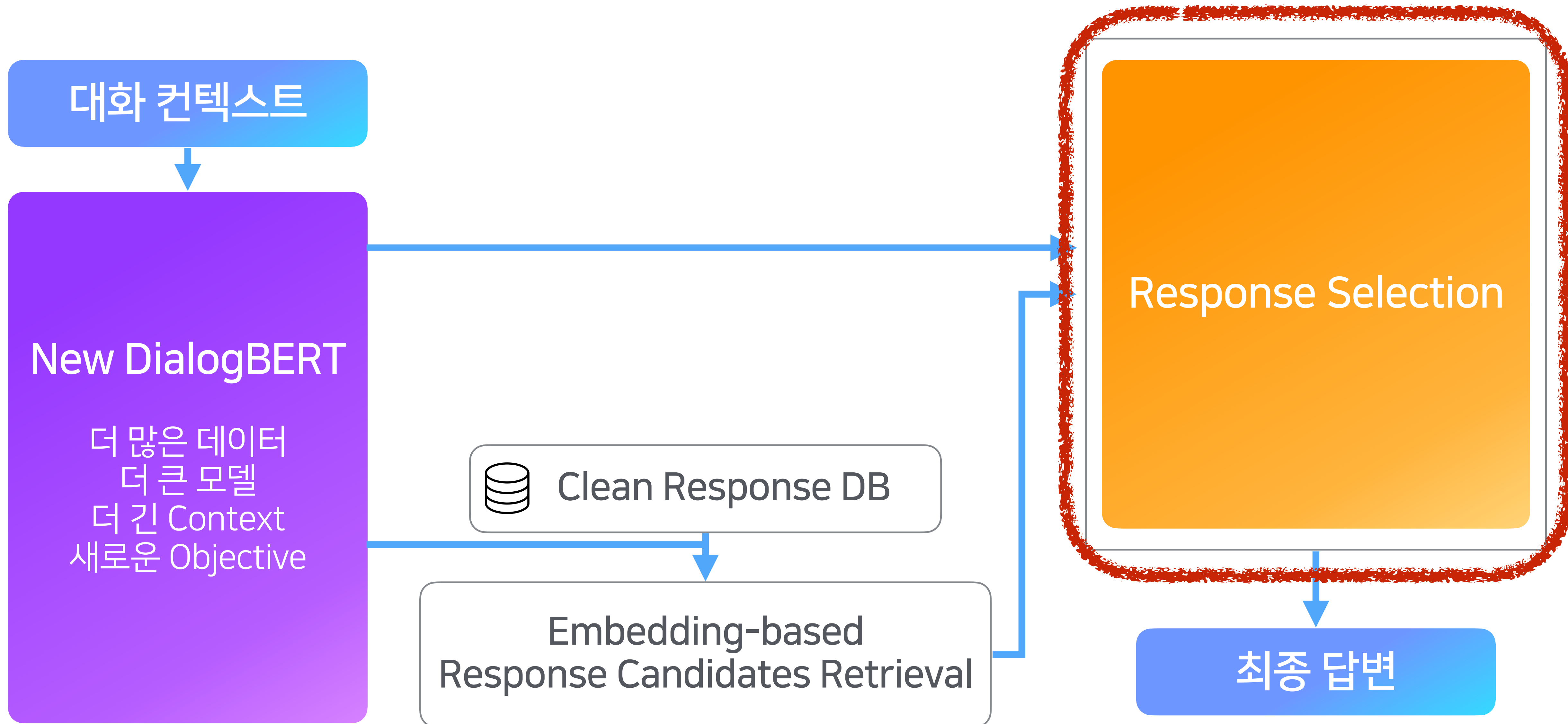
3.1 루다 알파에서 루다 베타로



3.1 루다 알파에서 루다 베타로



3.1 루다 알파에서 루다 베타로

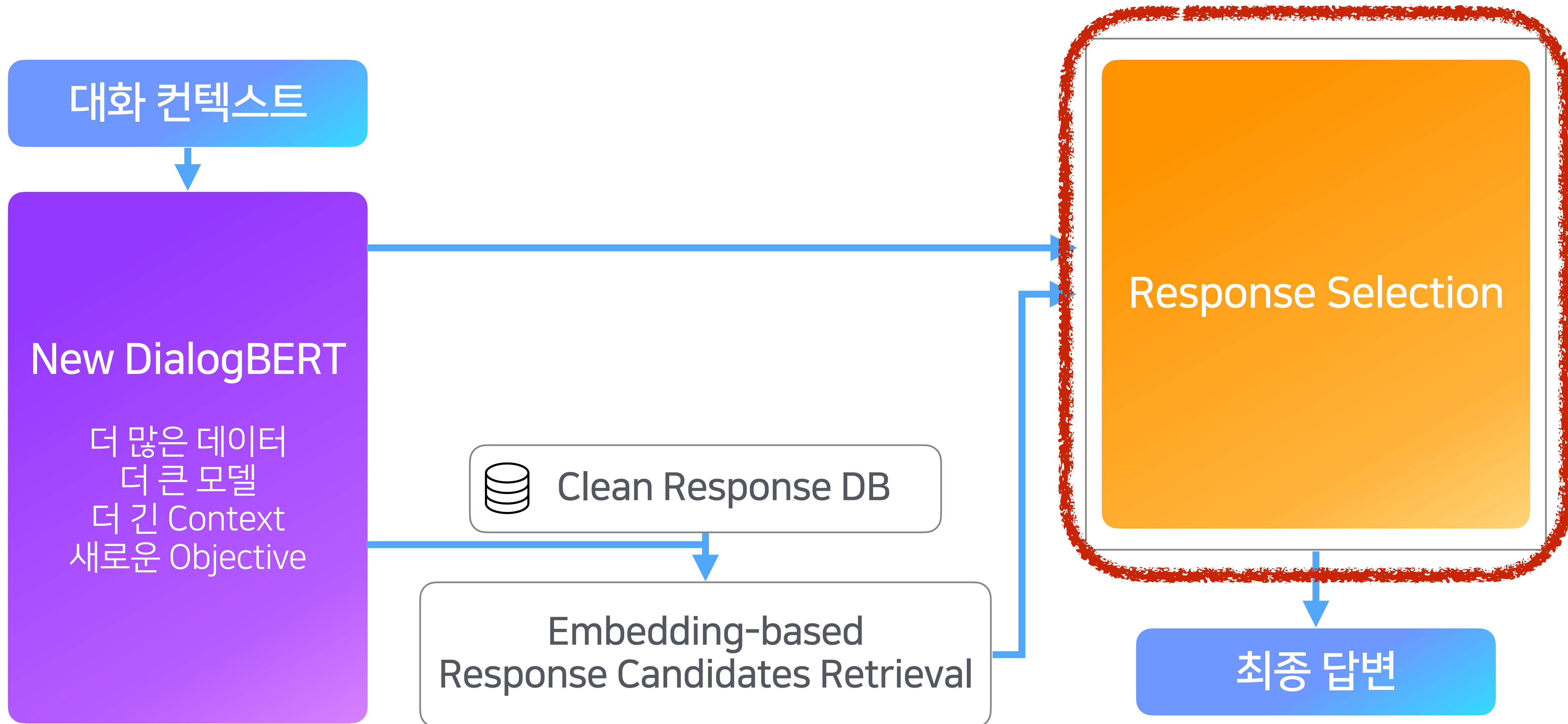


3.1 루다 알파에서 루다 베타로

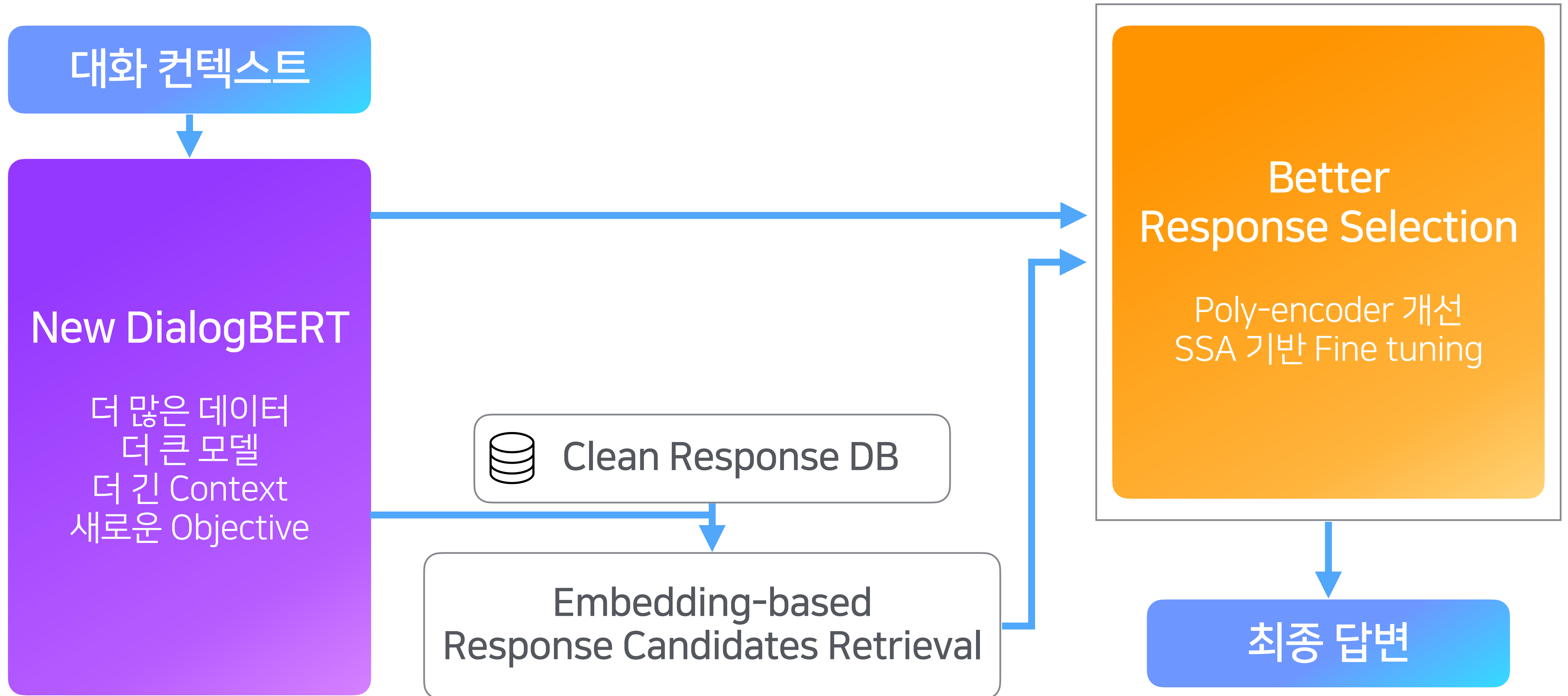
업데이트 5. Fine-tuning으로 Response Selection 성능을 개선하자

- Poly-encoder의 개량: 직전 턴, 같은 화자의 턴 별도 학습
- Fine-tuning: 24,000세션에 대해 응답 후보에 대한 SSA 레이블링 데이터로 학습

3.1 루다 알파에서 루다 베타로



3.1 루다 알파에서 루다 베타로



3.2 루다 베타 성능 평가

최종 성능

	Xiaolce	Google Meena	루다 알파	루다 베타
Sensibleness	45%	86%	69%	83%
Specificity	17%	63%	49%	73%
SSA	31%	78%	59%	78%

3.2 루다 베타 성능 평가

루다 정식 버전

- 루다 베타 + 여러 제품 요소
- 연말 출시 예정
- '과연 몇 명의 사람들과, 얼마나 많이, 얼마나 오래 얘기하고 친해질 수 있을까?'

4. 핵심 교훈

4.1 신뢰할만한 벤치마크를 만들자

오픈도메인 대화의 특성상 좋은 벤치마크를 만드는 게 쉽지 않음

- "Context는 몇 개여야 할까?"
- "Positive는 몇 개여야 할까? Negative는 어떤 종류로 몇 개여야 할까?"
- 좋은 벤치마크는 리서치의 등불이자 나침반

Probing task의 필요성

- Ribeiro et al., 2020
- 정량적인 결과와 함께 어떤 걸 잘 하고, 못 하는지 훑아보아야 함

앞으로 오픈도메인 벤치마크와 체크리스트를 최대한 공개할 예정

4.2 엄밀한 실험이 더 빠르다

제품을 목표로 하는 머신러닝 리서치의 딜레마

- 리서치의 엄밀함 vs. 제품 릴리즈를 위한 속도
- 트레이드오프가 있지만 많은 경우 엄밀한 진행이 오히려 시간을 아끼게 해준다
- 변인 통제, 실험의 기록, 실험 순서, 타 모델 및 데이터를 통한 검증 등

4.3 복잡도도 성능만큼 중요하다

복잡도 vs. 성능

- 프레임워크를 더 복잡하게 만들면서 성능을 올리는 논문이 다수
- 하지만 성능을 올리는 게 능사일까?
- 때로는 심플함을 위해 성능을 포기하는 게 나을 수 있음

복잡도의 비용

- 버그 발생 확률 높아짐
- 디버깅의 어려움
- 추후 모든 작업의 효율을 낮춤

4.4 머신러닝 제품도 린하게 만들자

다시 처음으로 돌아간다면...?

- Dialogue Act, Topic, Engage Mode 등에 시간을 많이 썼지만 결국 쓰이지 않음
- Xiaolce를 따라 모든 요소를 만들지 않고, 작동할 수 있는 최소 제품(MVP) 부터
- MVP를 만들고, 실제로 생기는 문제를 관찰하고 해결하는 방식으로 진행했다면..

5. 앞으로의 과제



5.1 중단기 과제

Persona

- 어떻게 루다에 페르소나에 맞는 말을 깊게 골라낼 것인가?

말을 잘 거는 루다

- 어떻게 사용자가 답하고 싶은 타이밍에, 답하고 싶은 내용으로 말을 걸 것인가?

Continual learning

- 어떻게 루다와 사용자의 통해 지속적인 개선을 이룰 것인가?

외부 데이터

- 어떻게 대화가 아닌 텍스트를 통해 지식과 콘텐츠를 녹여낼 것인가?

5.2 중장기 과제

초대형 생성 모델

- 어떻게 2B 이상의 생성 모델을 잘 학습할 것인가?

기억과 개인화

- 어떻게 사용자의 정보와 과거 대화를 기반으로 개인화된 대화를 할 것인가?

경량화

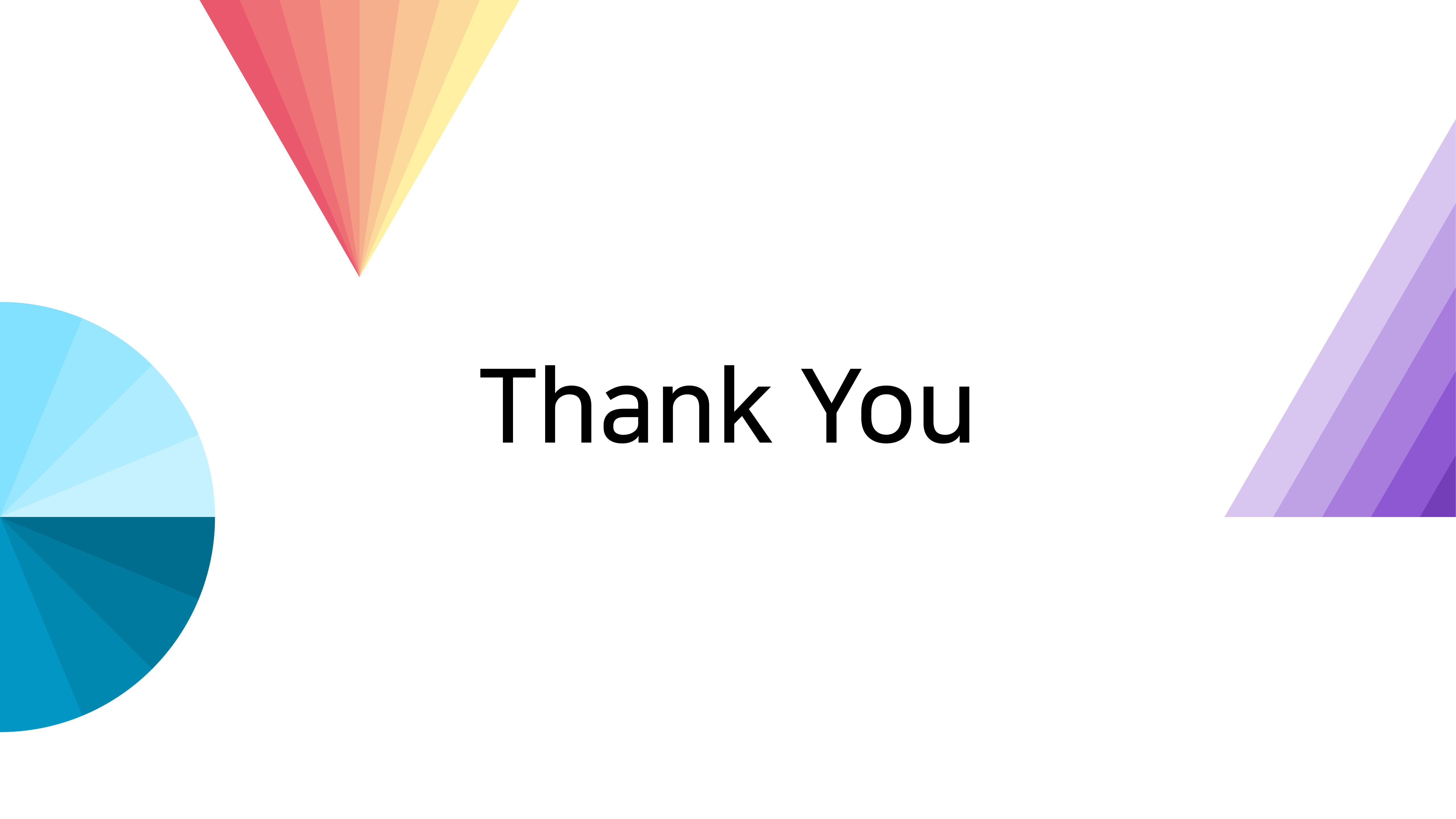
- 어떻게 초대형 모델을 서빙할 것인가?

Vision

Towards a superhuman open-domain conversational AI

이 문제를 함께 풀고 싶다면? 📍

scatterlab.co.kr/recruiting/#pingpong



Thank You