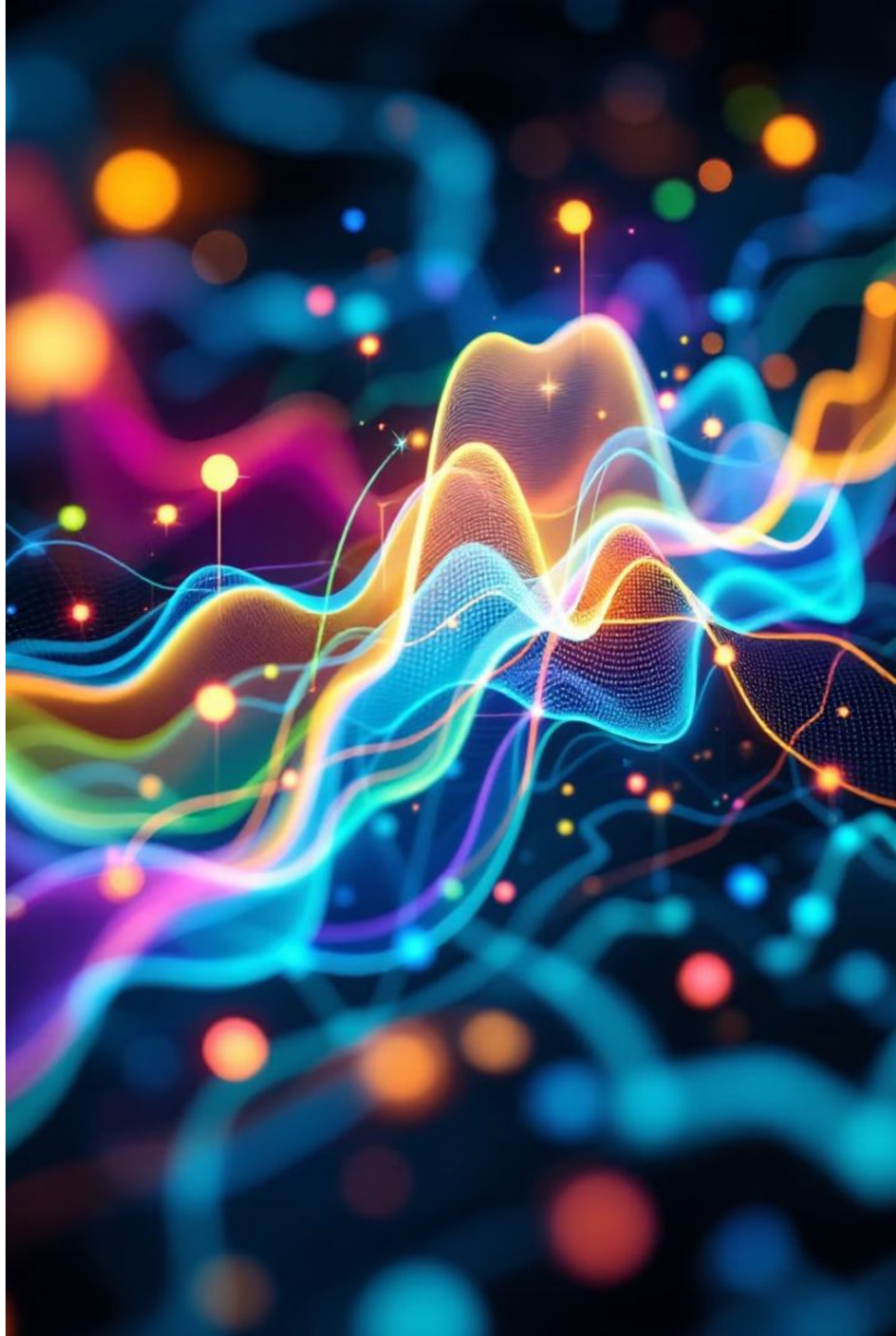
상담 지원을 위한 질의 챗봇 성능 평가 결과서



배경 및 목적

`**상담원 질의를 위한 자연어 기반 챗봇**`의 성능 분석 및 개선안 도출 위한 목적으로 수행됨

성능 평가 통해 높은 정확도의 답변 도출과
신속성 있는 상담 지원을 목표

경량화된 LLM의 성능은 지속적으로 발전하고 있으며 특정 도메인에 특화된 LLM을 직접 구축할 수 있는 단계입니다.
성능 평가는 답변의 정확성을 극대화하고 상담원의 능률 향상에 필수적인 요소입니다.
본 평가 결과는 향후 시스템 개발 및 운영에 필요한 의사 결정을 위한 중요한 정보로 활용될 것입니다.

평가 대상 선정

1 **RAG 성능**

KMS 데이터셋에서

데이터 반환 정확성 확인

- RAGas 성능 확인

2 **LLM Prompt 성능**

상담원에게 필요한 정보에

대한

직관적 제공 여부 확인

정성적 판단

3 **Cache Layer 성능**

자주 질의하는, 혹은 동일한 답변을 요구하는

여러 질의에 대한 빠른 성능 제공 여부 - 신속성 판단

평가 항목 선정

처리 속도

질의문(Query) 요청 시부터 반환 시까지 처리 시간 분석

확장성

프롬프트 수정에 따른 상담원 질의 UX 변화를 분석

정확성

RAG와 LLM의 문서 기반 정확한 답변을 도출하는 지 Precision / Recall / 정성 평가

비용 효율성

서버 배포 및 환경 구축, 운영 및 유지보수 비용 등을 분석

평가 방법론

1

성능 테스트 도구를 활용하여 시스템에 다양한 부하를 가하고 성능 지표를 측정합니다.

2

실제 데이터를 사용하여 시스템의 성능을 평가하고, 시스템의 실제 환경에서의 성능을 파악합니다.

3

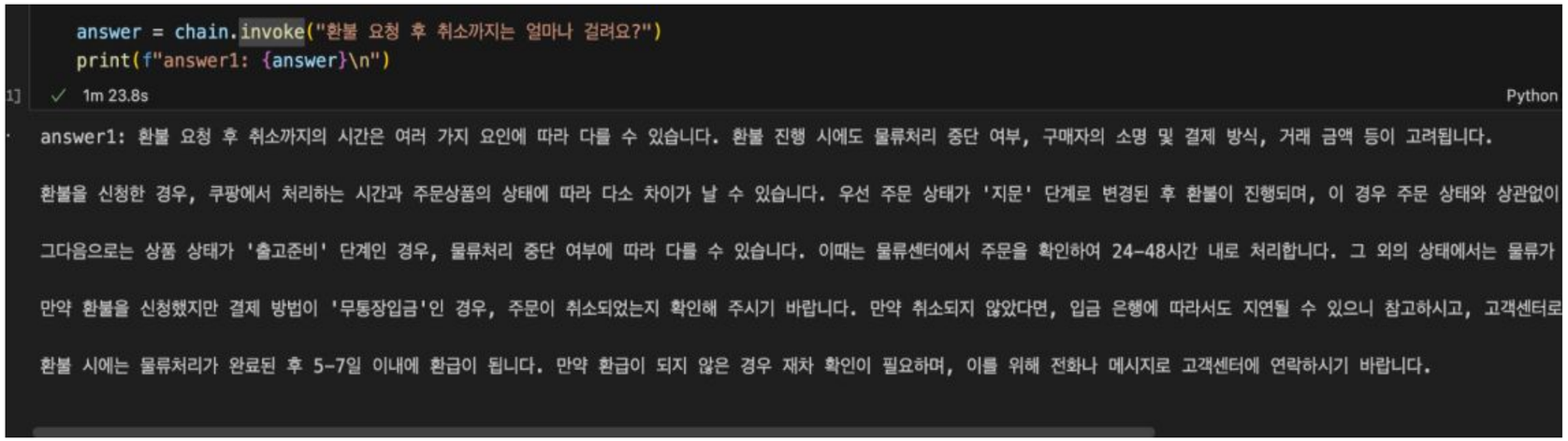
평가 결과를 분석하고 실제 사용 UX 시나리오를 기반으로 단점을 파악하여 개선 방안을 도출

[RAG + LLM] Langchain 성능 평가 :

ChromaDB

OllamaEmbed

Ollama3.1:8b



평가 항목	결과	평가
Embedding 시간	10분 내외	실제 성능과 무관
확장성	확장 시 Embed, Query 처리 시간 증가	미흡
성능(정확도)	질의에 대해 정확한 답변 반환	미흡
Query 처리 시간	1분 30초 내외	미흡

[RAG + LLM]
Langchain 성능 평가 :
PGVector
OllamaEmbed
Ollama3.1:8b

```python
def merge_pages(pages):
    merged = "\n\n".join(page.page_content for page in pages)
    return merged

chain = (
    {"query": RunnablePassthrough(), "context": retriever | merge_pages}
    | prompt
    | llm
    | StrOutputParser()
)

answer = chain.invoke("조금 전에 제가 주문을 했는데 주문이 된 건지 문자가 없어서 그래요.")
print(f"answer1: {answer}\n")
```

[18] ✓ 29.3s

```
answer1: 주문이되었습니다. 카드 결제수단을 설정한 경우 주문시 자동으로 카드결제가 진행됩니다.

그러나 카드 결제 시 3Dsecure 인증이 실패하는 경우에는 문자알림을 받지 못할 수 있습니다.

쿠팡 고객센터에 문의하시면 더욱 빠르게 확인하실 수 있으니 도움이 필요하신 경우 이용해 주세요.
```

평가 항목	결과	평가
Embedding 시간	10분 내외	실제 성능과 무관
확장성	확장 시 Embed, Query 처리 시간 증가	미흡
성능(정확도)	질의에 대해 정확한 답변 반환	우수
Query 처리 시간	30초 내외	미흡

[RAG + LLM] Langchain 성능 평가 : ChromaDB UpstageEmbed Gemini API

```
query = '조금 전에 제가 주문을 했는데 주문이 된 건지 문자가 없어서 그래요.'
persist_db.similarity_search(query)

retrieved_docs = persist_db.similarity_search(query, k=3)

for doc in retrieved_docs:
    pprint(doc.page_content)
```

Executed at 2024.12.27 10:38:28 in 1s 42ms

```
('Question: Q[주문] 주문하지 않은 상품을 받았는데 누가 주문한지 알 수 있나요?\n'
 'Answer: 배송출발 또는 배송완료 안내 문자를 받은 경우 문자에서 주문자의 이름을 확인할 수 있습니다. 단 해당 주문자가 누구인지 '
 '모르거나 안내 문자를 받지 못했다면 운송장에 기재된 주문번호(주문번호 확인 불가 시 운송장 번호)를 확인하여 쿠팡 '
 '고객센터(1577-7011)로 연락 주시기 바랍니다. 상품의 정상 배송 확인 및 주문자와 연락을 통해 주문자 확인을 도와드리겠습니다.\n'
 'keywords: ORDER\n'
 'count: 0')
('Question: Q[무통장 입금] 무통장입금을 했는데 입금확인은 언제 되나요?\n'
 'Answer: 무통장입금(가상계좌)으로 입금한 경우 약 10분 이내로 입금 내역이 확인됩니다. 입금이 완료되면 고객님께 문자 메시지 또는 '
 '모바일 앱 알림이 발송됩니다. 입금 후 1시간 뒤에도 주문 내역에서 '입금대기중' 상태가 지속되면 쿠팡 고객센터(1577-7011)로 '
 '연락 주시기 바랍니다.\n'
 'keywords: ORDER\n'
 'count: 0')
('Question: Q[배송] 상품을 이미 받았는데 발송되었다는 문자가 왔습니다.\n'
 'Answer: 쿠팡에서 배송 정보를 전산에 입력하면 발송안내 문자메시지가 전송됩니다. 그러나 간혹 운송장번호 입력이 늦어지는 경우가 '
```

평가 항목	결과	평가
Embedding 시간	33초	실제 성능과 무관
확장성	확장 시 처리 시간 변화 없음	우수
성능(정확도)	질의에 대해 정확한 답변 반환	우수
Query 처리 시간	1.7초	우수

[RAG + LLM]
Langchain 성능 평가 :
PineconeDB
UpstageEmbed
Gemini API

```
query2 = '휴대폰 결제는 어떻게 해?'

retrieved_docs2 = database.similarity_search(query2, k=3)

for doc in retrieved_docs2:
    pprint(doc.page_content)
```

Executed at 2024.12.27 10:32:48 in 875ms

```
('Question: Q[휴대폰결제] 휴대폰 결제란 무엇이며 어떻게 결제해야 하나요?\n'
 'Answer: 휴대폰 결제란 휴대폰 명의자가 문자 서비스로 본인 인증을 한 후 정해진 한도 내에서 결제하는 후불제 결제 수단입니다. 결제 '
 '대금은 다음 달 휴대폰 요금에 부과되어 청구됩니다. 결제 한도는 통신사 정책에 따라 각 고객마다 다를 수 있습니다. 결제 한도 확인 및 '
 '증액 가능 여부는 이용하시는 통신사로 직접 문의해 주시기 바랍니다. \n'
 ' 휴대폰 결제 방법\n'
 '① 주문/결제 페이지에서 결제 수단으로 [휴대폰] 선택\n'
 '② 이용하는 통신사 선택 후 [결제하기] 클릭\n'
 '③ 휴대폰 번호 휴대폰 명의자 개인정보 입력 후 [승인번호 요청] 클릭\n'
 '④ 문자(SMS)로 받은 승인번호 입력 후 [결제하기] 클릭 \n'
 ' 참고사항 \n'
 '다음의 경우에는 휴대폰 결제가 불가합니다.\n'
 '• 미성년자 사업자 외국인 명의로 등록된 휴대폰\n'
 '• 요금 미납 또는 체납 상태의 휴대폰\n'
 '• 정액 요금제 선불 휴대폰 등\n'
```

평가 항목	결과	평가
Embedding 시간	39.6초	실제 성능과 무관
확장성	확장 시 처리 시간 변화 없음	우수
성능(정확도)	질의에 대해 정확한 답변 반환	우수
Query 처리 시간	6~7초	보통

[RAG + LLM]
Langchain 성능 평가 :

Faiss DB
Snowflake-Arctic

```
ollama_similar_docs = ollama_db.similarity_search('조금 전에 제가 주문을 했는데 주문이 된 건지 문자가 없어서 그래요.', K=3)

for doc in ollama_similar_docs:
    print(f"snowflake: {doc.page_content}")
    print("="*100)
```

Executed at 2024.12.27 11:18:50 in 2s 167ms

```
snowflake: Question: Q[주문] 주문하지 않은 상품을 받았는데 누가 주문한지 알 수 있나요?
Answer: 배송출발 또는 배송완료 안내 문자를 받은 경우 문자에서 주문자의 이름을 확인할 수 있습니다. 단 해당 주문자가 누구인지 모르거나
  운송장 번호)를 확인하여 쿠팡 고객센터(1577-7011)로 연락 주시기 바랍니다. 상품의 정상 배송 확인 및 주문자와 연락을 통해 주문자 확
keywords: ORDER
count: 0
====================================================================================================
snowflake: Question: Q[배송] 상품을 이미 받았는데 발송되었다는 문자가 왔습니다.
Answer: 쿠팡에서 배송 정보를 전산에 입력하면 발송안내 문자메시지가 전송됩니다. 그러나 간혹 운송장번호 입력이 늦어지는 경우가 있으며
  상품 수령 후에 발송안내 문자를 받아도 주문한 상품이 이중으로 배송되지는 않습니다.
keywords: DELIVERY
count: 0
====================================================================================================
snowflake: Question: Q[쿠페이] 결제를 하지 않았는데 주문이 되었어요.
Answer: 아래의 경우에 해당하지 않는지 확인해 주시기 바랍니다.
```

평가 항목	결과	평가
Embedding 시간	9분 56초	실제 성능과 무관
확장성	확장 시 처리 시간 변화 없음	우수
성능(정확도)	질의에 대해 정확한 답변 반환	우수
Query 처리 시간	2.1초	우수

최종 모델 :
Pinecone / UpstageEmbed / Gemini

1

VectorStore :: Pinecone

임베딩 시간은 모델 서비스 이전 과정

문서 Retrieve 정확도 우수

클라우드 기반 서비스이기 때문에 구축 과정 간단함

2

Embedding Function :: UpstageEmbedding

임베딩 시간은 모델 서비스 이전 과정

문서 Retrieve 정확도 우수, 실시간성 우수

클라우드 기반 서비스이기 때문에 구축 과정 간단함

3

LLM : Gemini

응답에 대한 관련성 높음

문서 Retrieve부터 적절한 한글 답변까지의 실시간성

클라우드 기반 서비스이기 때문에 구축 과정 간단함

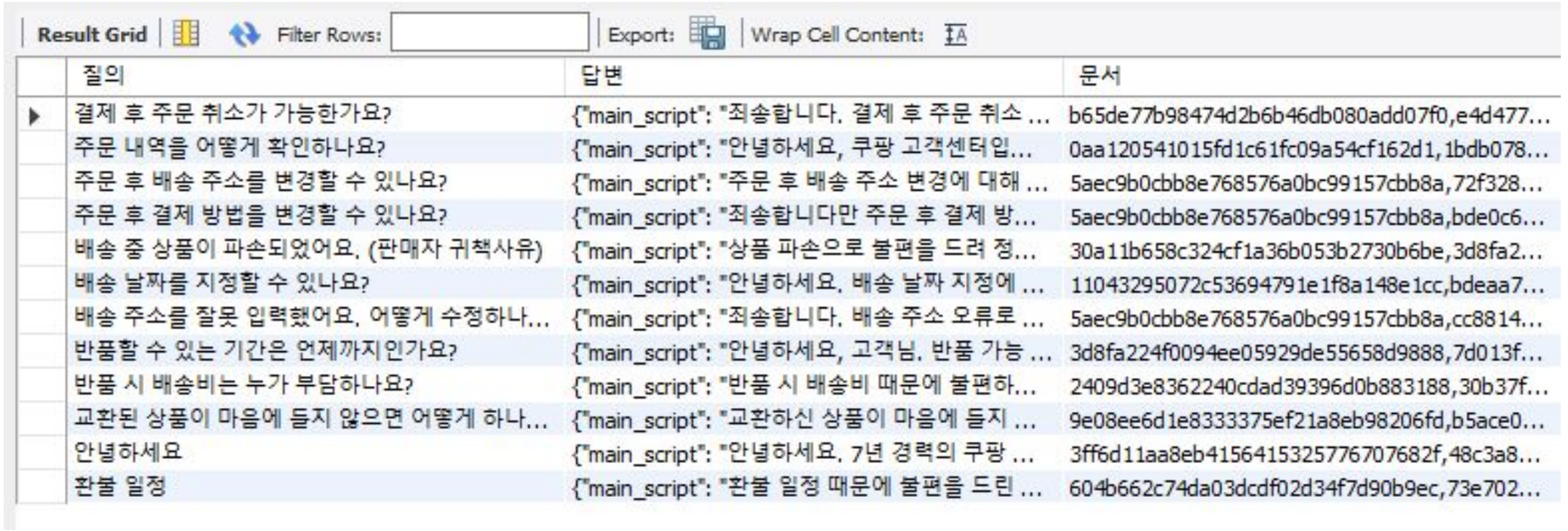
LLM Prompt 개선

성능 평가

```
template = """
[context]: {context}
---
[질의]: {query}

7년 이상의 경력을 가진 상담사라고 생각하고, 위의 [context] 정보 내에서 [질의]에 대해 상담사 입장에서 사용자가 만족할 수 있을 정도로 성의있게 답해주세요.
최대한 문장을 쉼표로 끊어서 대답하기 보다는 온점으로 문장을 끊어주세요.
문장의 마무리는 '~요' 보다는 '~다'로 끝나는 쪽이 전문적으로 보입니다.

또한, 상담사는 가능한 선에서 직접 확인+안내+해결을 도와주는 직원이므로 직접 확인 후 해결까지 돕는 방향으로 작성해 주세요.
그리고, 사용자의 편의를 위해 서비스 특성 상 쿠션어를 사용하시면 좋습니다.
쿠션어의 예시는 다음과 같습니다.
예시)
불편을 드려 죄송합니다.
번거로우시겠지만~
~하는 점 양해 부탁드립니다.
~할 예정입니다.
~를 부탁드립니다.

위 사항들을 종합해서 2~3줄로 상담사가 활용하기 좋게 대본을 만들어 주세요.
대본을 만들면 대본의 마침표가 나올때마다 보기 쉽게 실제 줄바꿈을 해주세요.

만약, 조건별로 안내 내용이 다른 경우
1차 응대 (양해멘트 or 1차 안내 등) + 정보 확인 멘트로 대본을 구성하면 됩니다.
정보 확인 멘트는 "정확한 상담을 위해 주문하신 주문 번호 확인 부탁드립니다." 입니다.
문서의 아래에 각 조건별 대응 방법을 기술해 주세요.
그리고 조건별 안내 내용은 간결하게 내용만말해줘.

단, 제일 중요한 것은 [context] 정보에 없는 내용을 답해서는 안됩니다. [context]에 정보가 없거나 문서들의 유사성이 0.2 이하로 떨어질 경우,
"문의주신 내용은 확인이 필요하여 지금 답변드리기 어려울 것 같습니다. 번거로우시겠지만 확인 후에 다시 연락드려도 괜찮을까요?" 라고 답해주세요.


format instructions: {format_instruction}
sub_scenarios에 들어갈 조건별 대응 방법들이 많은 경우, 더 추가해도 됩니다.
"""
```

평가 항목	결과	평가
처리 속도	답변 형식 맞추기 위해 조금 느려짐	보통
확장성	데이터 형식 변동에 맞춰 구성 가능	우수
성능	Prompt 개선만으로 상담사 UX 개선	우수
비용 효율성	데이터셋 추가 처리 없이 개선	우수

Cache Layer
도입 성능 평가



Result Grid | Filter Rows: | Export: | Wrap Cell Content:

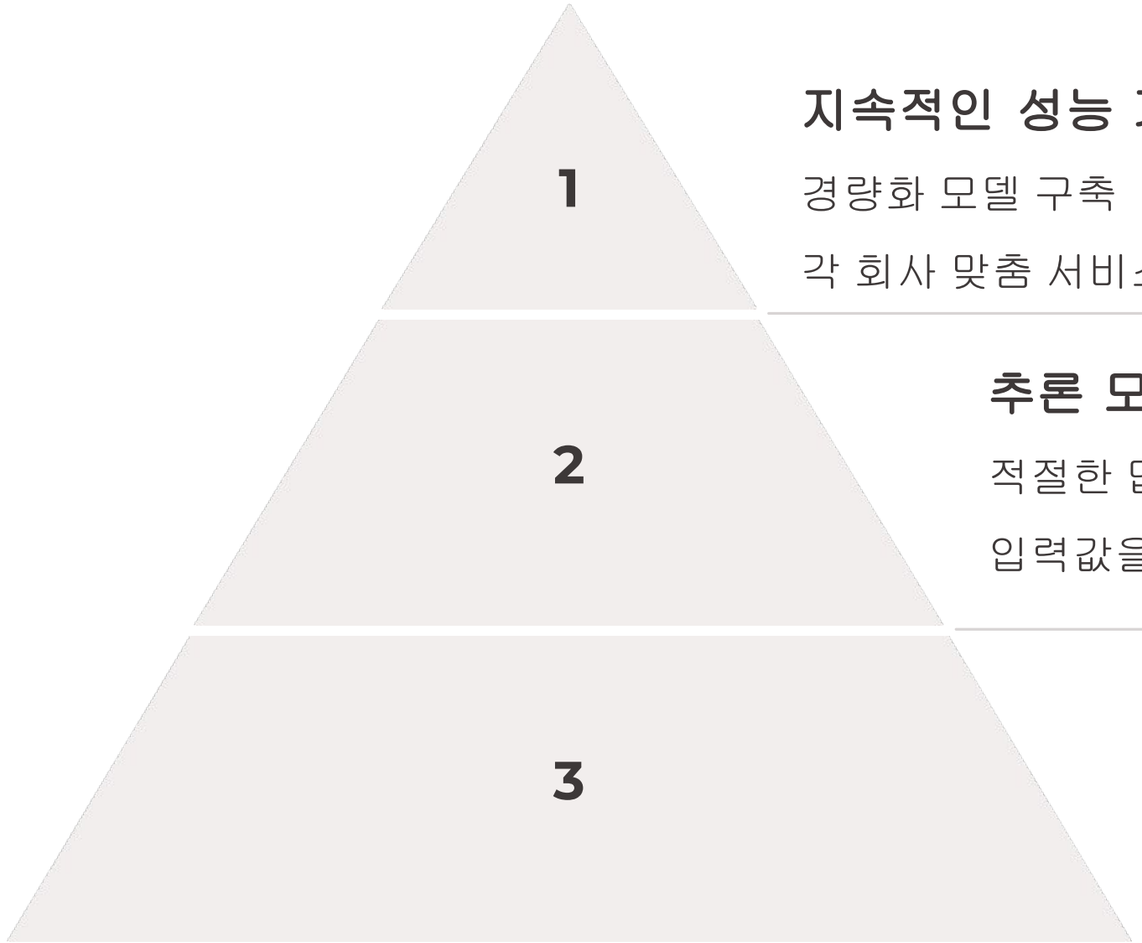
질의	답변	문서
결제 후 주문 취소가 가능한가요?	{"main_script": "죄송합니다. 결제 후 주문 취소 ...	b65de77b98474d2b6b46db080add07f0,e4d477...
주문 내역을 어떻게 확인하나요?	{"main_script": "안녕하세요, 쿠팡 고객센터입...	0aa120541015fd1c61fc09a54cf162d1,1bdb078...
주문 후 배송 주소를 변경할 수 있나요?	{"main_script": "주문 후 배송 주소 변경에 대해 ...	5aec9b0cbb8e768576a0bc99157cbb8a,72f328...
주문 후 결제 방법을 변경할 수 있나요?	{"main_script": "죄송합니다만 주문 후 결제 방...	5aec9b0cbb8e768576a0bc99157cbb8a,bde0c6...
배송 중 상품이 파손되었어요. (판매자 귀책사유)	{"main_script": "상품 파손으로 불편을 드려 정...	30a11b658c324cf1a36b053b2730b6be,3d8fa2...
배송 날짜를 지정할 수 있나요?	{"main_script": "안녕하세요. 배송 날짜 지정에 ...	11043295072c53694791e1f8a148e1cc,bdeaa7...
배송 주소를 잘못 입력했어요. 어떻게 수정하나...	{"main_script": "죄송합니다. 배송 주소 오류로 ...	5aec9b0cbb8e768576a0bc99157cbb8a,cc8814...
반품할 수 있는 기간은 언제까지인가요?	{"main_script": "안녕하세요, 고객님. 반품 가능 ...	3d8fa224f0094ee05929de55658d9888,7d013f...
반품 시 배송비는 누가 부담하나요?	{"main_script": "반품 시 배송비 때문에 불편하...	2409d3e8362240cdad39396d0b883188,30b37f...
교환된 상품이 마음에 들지 않으면 어떻게 하나...	{"main_script": "교환하신 상품이 마음에 들지 ...	9e08ee6d1e8333375ef21a8eb98206fd,b5ace0...
안녕하세요	{"main_script": "안녕하세요. 7년 경력의 쿠팡 ...	3ff6d11aa8eb41564153257767076 82f,48c3a8...
환불 일정	{"main_script": "환불 일정 때문에 불편을 드린 ...	604b662c74da03dcdf02d34f7d90b9ec,73e702...

평가 항목	결과	평가
처리 속도	hit 시 6~7초에서 1~2초로 단축	우수
확장성	데이터 유형 무관 - 모두 문자열로 저장	우수
질의 구조	가변 길이의 문서를 md5 함수로 고정 길이의 문서로 해시	
비용 효율성	캐시를 위한 서버가 아닌, DB로만 구축 SQL 통한 간단한 적재 및 질의 구조	우수

RAGas 성능 평가

평가 항목	결과	평가
faithfulness	답변이 사실(Context)에 근거하여 정확한 답변 반환	우수
context_recall	답변의 재현율이 높다.	우수
context_precision	질문에 대한 답변과 관련있는 문서를 잘 찾는다.	우수

향후 발전 방향



지속적인 성능 개선

경량화 모델 구축

각 회사 맞춤 서비스 최적화 위해 용어 치환 혹은 설명 추가 과정

추론 모델 적용

적절한 답변만 제공하는 것뿐만 아니라,

입력값을 제공하면, 원하는 값에 대한 예상 수치를 계산하여 반환

로그 기반 통계 분석

자주 질의하는 질의, 특정 기간에 몰리는 질의를

기술통계 분석하여 인사이트 제공