평가 (Evaluation)

RAG 성능, 무엇을 어떻게 평가해야 하는가?

Hank

Who is Speaker?

Taehan hank Kim

AI Research Engineer

Braincrew Inc. RAG 팀 팀원

Multi-Domain의 RAG 및 Agent Project 수행

- Lifelog Domain RAG
- RAG Evaluation Platform

Linked-In

GitHub

CONTENTS

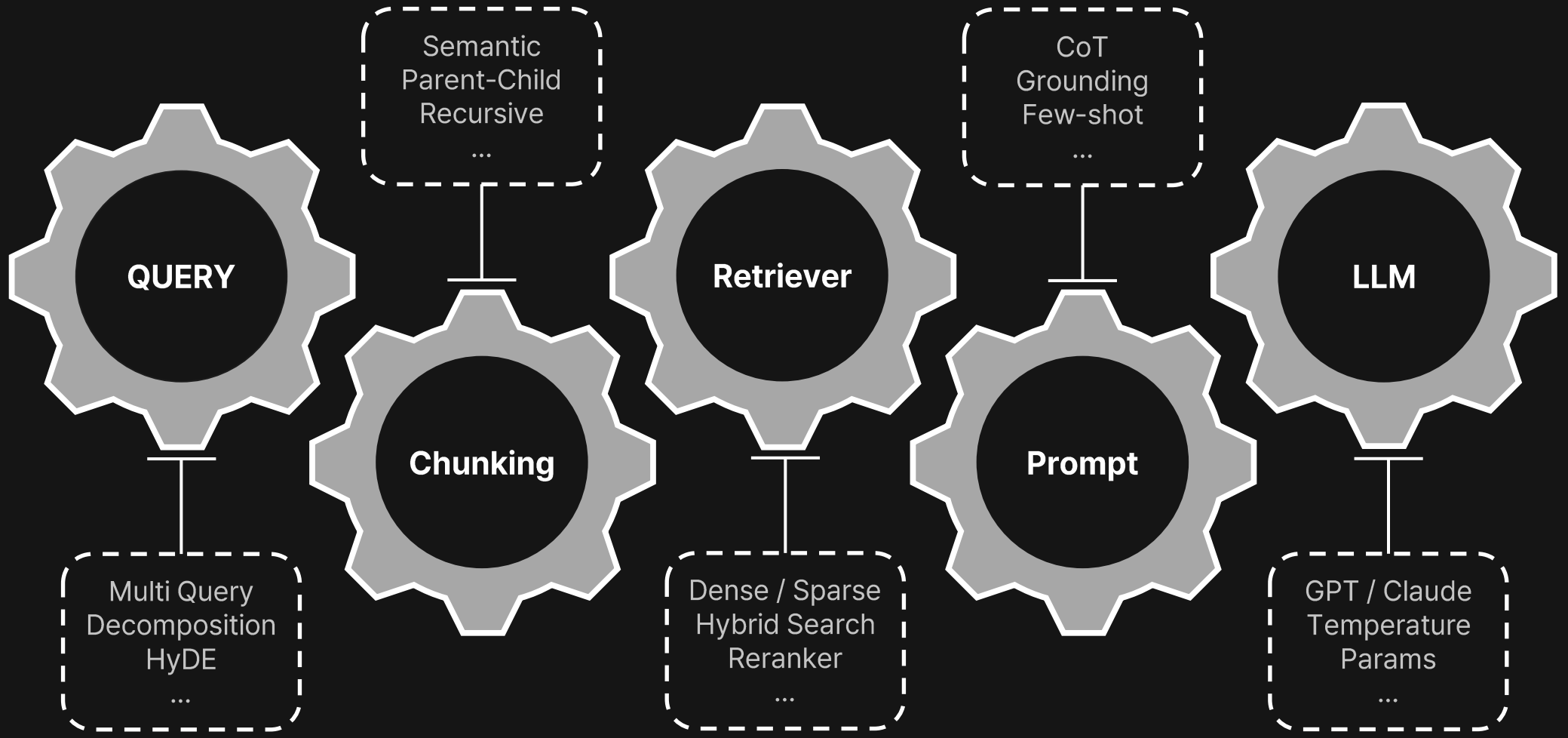
PART 01

무엇을 평가해야 하는가?

RAG 성능 평가 3가지 단계

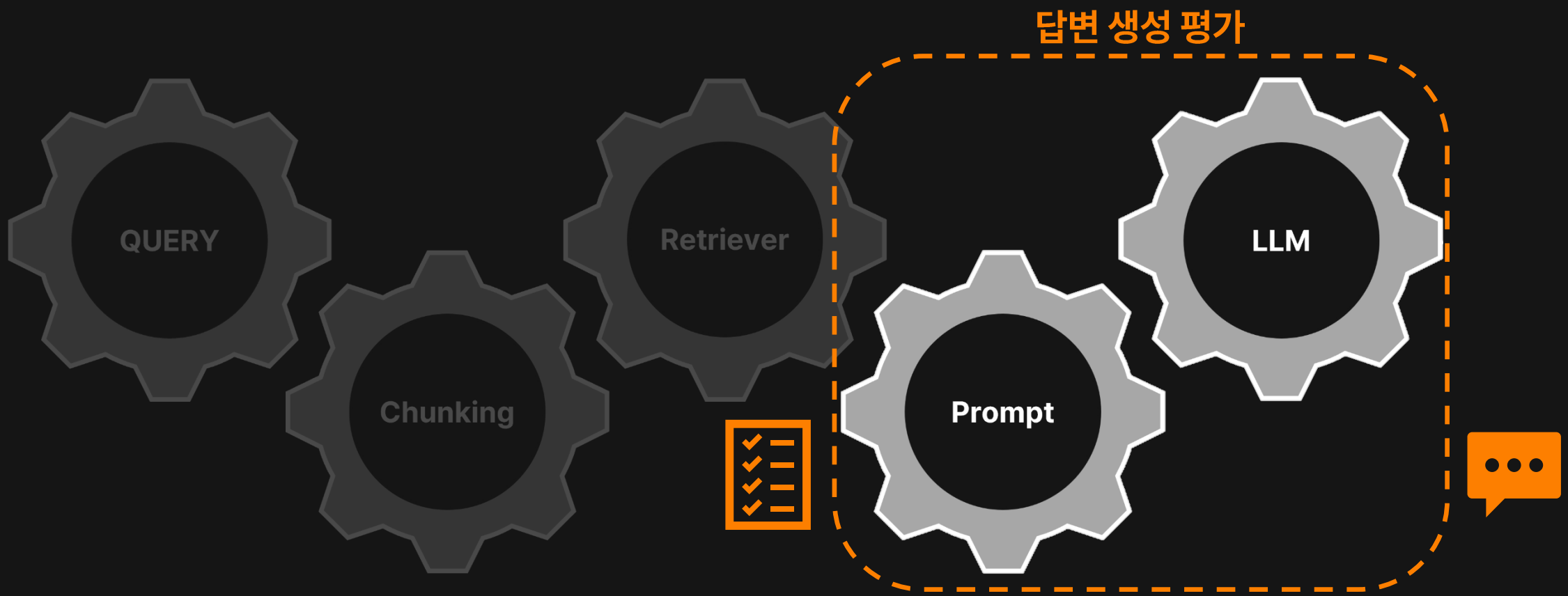
01

RAG, 무엇을 평가해야 하는가?



RAG, 무엇을 평가해야 하는가?

RAG, 무엇을 평가해야 하는가?



RAG, 무엇을 평가해야 하는가?

PART 02

어떻게 평가해야 하는가?

각 단계별 평가 지표 및 한계점

02

RAG, 어떻게 평가해야 하는가?

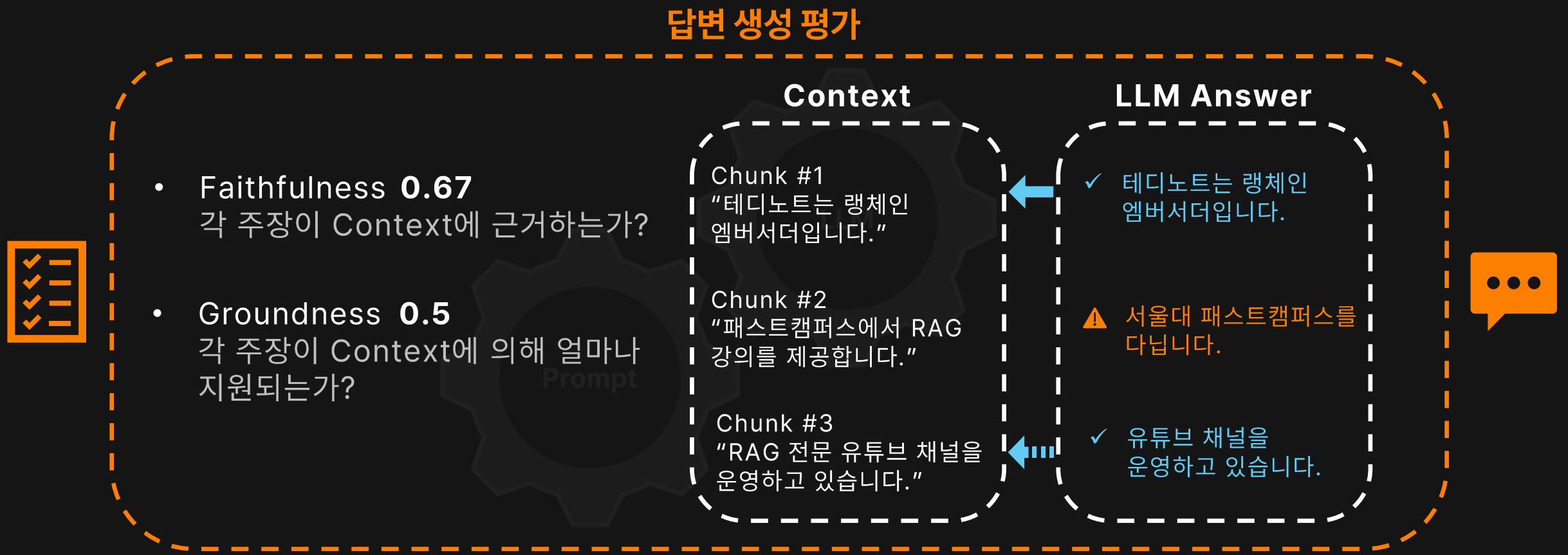
문서 검색 평가

- Precision@K **0.30**
  10개 중 3개 정답
- Recall@K **0.60**
  5개 중 3개 검색
- MRR **0.50**
  첫 정답 위치 : 2

- 검색된 문서 (K=10)

- 전체 정답 문서

RAG, 어떻게 평가해야 하는가?



RAG, 어떻게 평가해야 하는가?

End to End 평가

- Relevancy **1.00**
  답변이 질문에 적절한가?
- Correctness **0.67**
  답변이 GT를 봤을 때, 정답인가?
- Semantic Similarity **0.75**
  답변과 GT Cosine 유사도 검증

Question

"테디노트에 대해 알려주세요."

Answer

"랭체인 엠버서더이며, RAG 전문 유튜브 채널을 운영합니다."

Ground Truth

"랭체인 엠버서더, RAG 강의 제공, RAG 유튜브 채널 운영"

RAG, 어떻게 평가해야 하는가?

LLM-as-a-Judge, 만능인가?

Position Bias

7.5 8.0

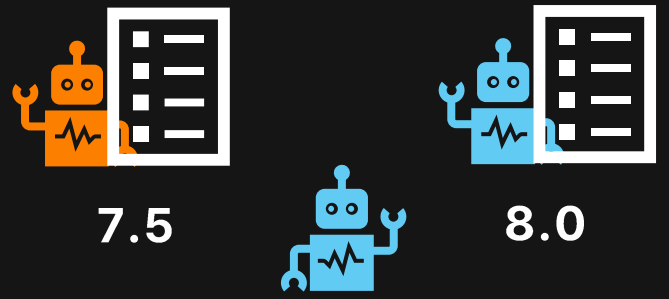
순서 편향

Verbosity Bias

7.5 8.0

답변 길이에 따른 편향

Self-enhancement



7.5 8.0

자기 편향

RAG, 어떻게 평가해야 하는가?

The Specialist Gap: LLM은 전문가가 아니다

→ 도메인 전문 지식이 필요한 영역에서 LLM Judge의 한계

출처: 2025 ACM IUI, "Limitations of the LLM-as-a-Judge Approach for Evaluating LLM Outputs in Expert Knowledge Tasks"

PART 03

Braincrew 적용 사례

03

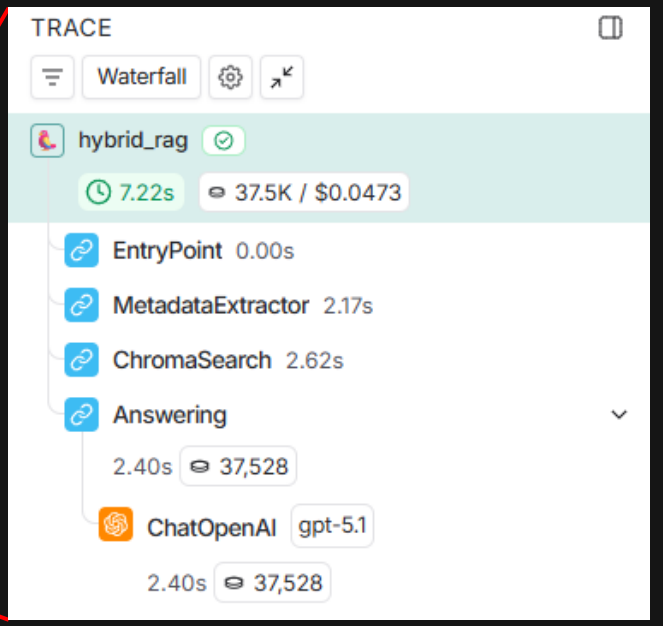
Braincrew 적용 사례

공통된 평가 체계 구축 및 Tracing

실험 Run 결과 확인

	Inputs	Reference Outputs	Outputs	Correctness 0.821 AVG	Latency 6.10 P50
1	8/19 바나나 머핀을 먹은 장소... #0348 →	컨퍼런스 홀	2025년 8월 19일에 바나나 ... 1/3	0.6667 μ	7.28s
2	8월 9일 숙소 도착한게 몇시쯤... #03bb →	8월 9일 11시 경	제공된 활동 기록에는 8월 11... 1/3	0.8333 μ	7.55s
3	8월 31일 점심에 만들어 먹은 ... #0440 →	비빔국수	8월 31일 점심(12시~14시)에... 1/3	1.00 μ	6.58s
4	가계부 앱 정리는 주로 무슨 요... #04fc →	일요일	가계부 앱(또는 가계부 정리 ... 1/3	1.00 μ	7.63s
5	여의도공원에서 자전거 대여를... #054a →	따릉이 앱	여의도공원에서 자전거를 대... 1/3	1.00 μ	5.97s
6	8월 15일 집에서 먹은 맥주는? #0847 →	카스 프레시	8월 15일에 집(평창 집)에서 ... 1/3	0.8333 μ	5.91s
7	8/31에 본 유튜브 채널 이름이 ... #098a →	유튜브 미주은 브리핑	2025년 8월 31일에 본 유튜... 1/3	0.6667 μ	5.74s
8	8월 30일 저녁 먹을 때 탄산수 ... #0ae1 →	트레비 탄산수	8월 30일 저녁 식사 때 마신 ... 1/3	0.00 μ	7.22s
9	8/23에 냉동실에서 꺼낸 왕교... #0b6b →	비비고	8월 23일에 냉동실에서 꺼낸 ... 1/3	1.00 μ	5.96s

상세 Trace



→ 점수 이상 발견 시, 해당 Run Trace 즉시 확인 및 원인 분석

Braincrew 적용 사례

평가 LLM 고도화: 도메인 특화 프롬프팅

최종 버전 ——— v11

도메인 용어 반영 ——— v3

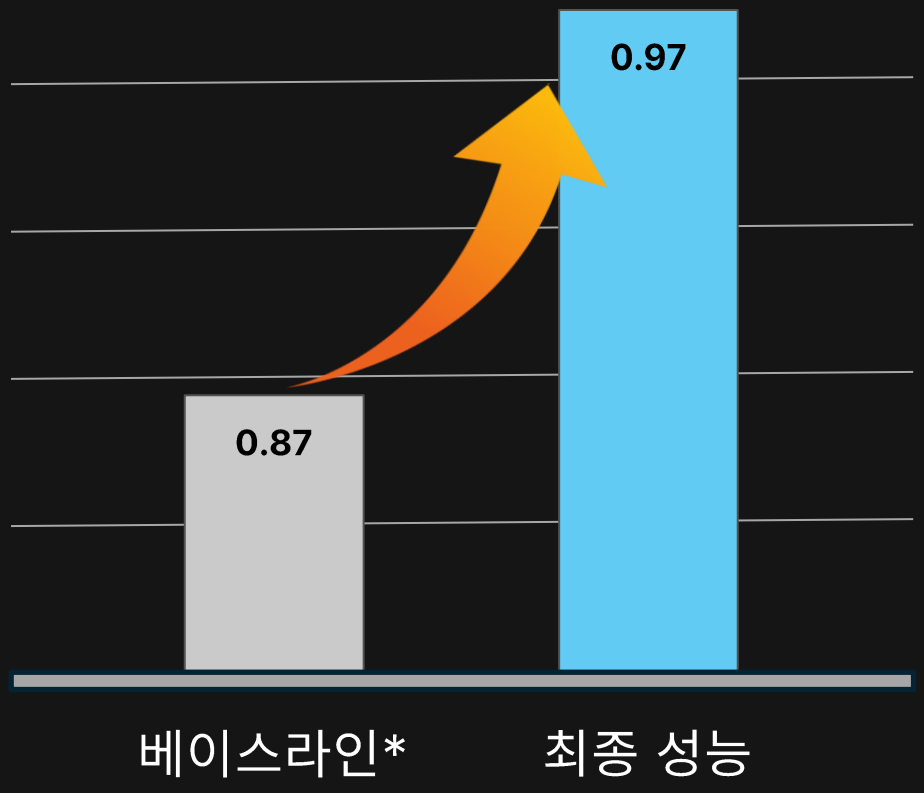
오판 케이스 발견 및 수정 ——— v2

기본 설정 ——— v1

Name	Visibility	Prompt Type	Created By	Last Updated
eval_[redacted]_evalset_20251218_...	Private	StructuredPrompt	Teddy Lee	2025. 12. 22. 오후 4:29...
eval_[redacted]_evalset_20251218_...	Private	ChatPromptTemplate	Teddy Lee	2025. 12. 22. 오후 4:29...
eval_[redacted]_evalset_20251218_...	Private	StructuredPrompt	Teddy Lee	2025. 12. 19. 오전 12:2...
eval_[redacted]_evalset_20251218_...	Private	StructuredPrompt	Teddy Lee	2025. 12. 19. 오전 12:2...
eval_[redacted]_evalset_correctne...	Private	StructuredPrompt	Teddy Lee	2025. 12. 19. 오전 12:2...
eval_[redacted]_trusty_evaluator_f...	Private	StructuredPrompt	JaeHun Choi	2025. 12. 17. 오후 11:26...
eval_[redacted]ce_trusty_evaluator...	Private	ChatPromptTemplate	JaeHun Choi	2025. 12. 17. 오후 9:49:...
eval_[redacted]_test_9368ad5c	Private	StructuredPrompt	JaeHun Choi	2025. 12. 17. 오후 9:09:13
eval_[redacted]ce_test_y2fh4dq0	Private	ChatPromptTemplate	JaeHun Choi	2025. 12. 17. 오후 9:09:12
eval_[redacted]_test_evaluator_12d...	Private	StructuredPrompt	JaeHun Choi	2025. 12. 17. 오후 9:07:...

Braincrew 적용 사례

최종 결과



베이스라인* : E2E 평가 단계 이전 RAG 파이프라인

투입

인원 : 3명
기간 : 3일

실험 : 45회
평가 : 11번 프롬프트 개선

결과

Human Eval: 87% -> 97%
10%p 상승

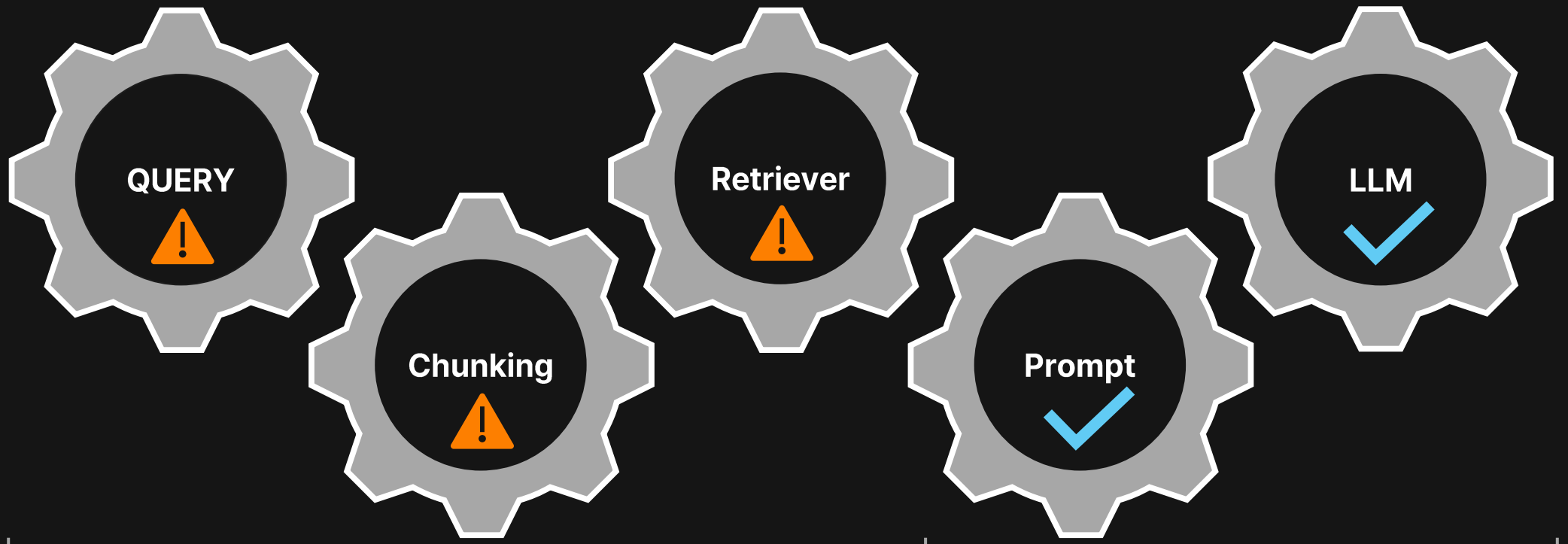
→ **신뢰할 수 있는 평가 체계가 있었기에 가능한 결과**

PART 04

Braincrew’s Next Step

04

LangSmith의 한계점



로직 구현 필요

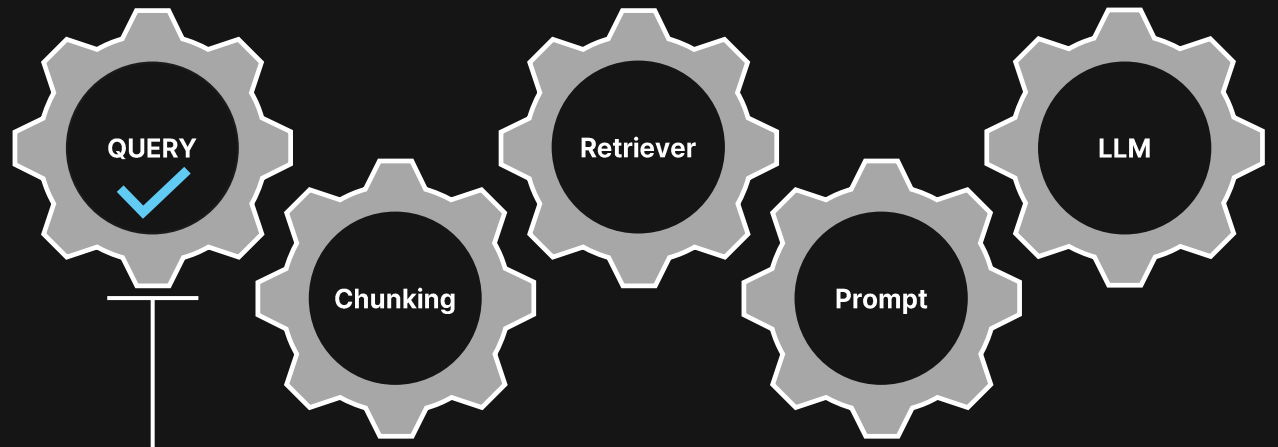
Pricision@K, Recall@K, MRR, nDCG 등 계산 로직

LLM-as-a-Judge

LangSmith Evaluator 활용

PReP: RAG evaluation Platform

브레인크루 내부 평가 플랫폼



📊 **Multi Query Results:**
Multi Query Evaluation Metrics

Strategy	Diversity	Coverage	Relevance
Multi-Query	0.9200	0.8800	1.0000

📊 **Hyde Results:**
Hyde Evaluation Metrics

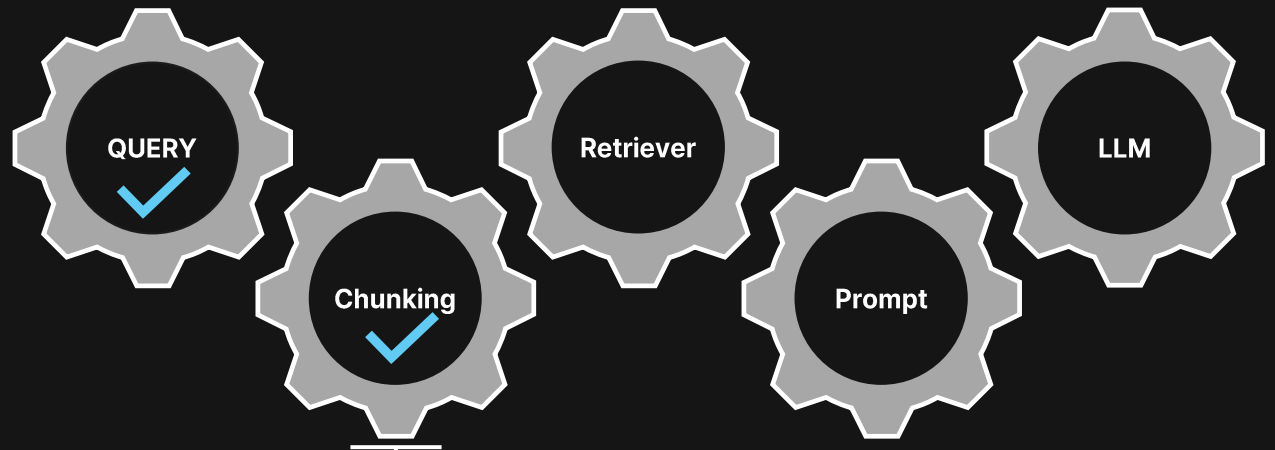
Strategy	Relevance	Specificity	Factuality	Coherence
HyDE	1.0000	0.9400	0.9600	1.0000

📊 **Decomposition Results:**
Decomposition Evaluation Metrics

Strategy	Completeness	Granularity	Independence	Answerability
Query Decomposition	0.9800	1.0000	0.8800	1.0000

PReP: RAG evaluation Platform

브레인크루 내부 평가 플랫폼

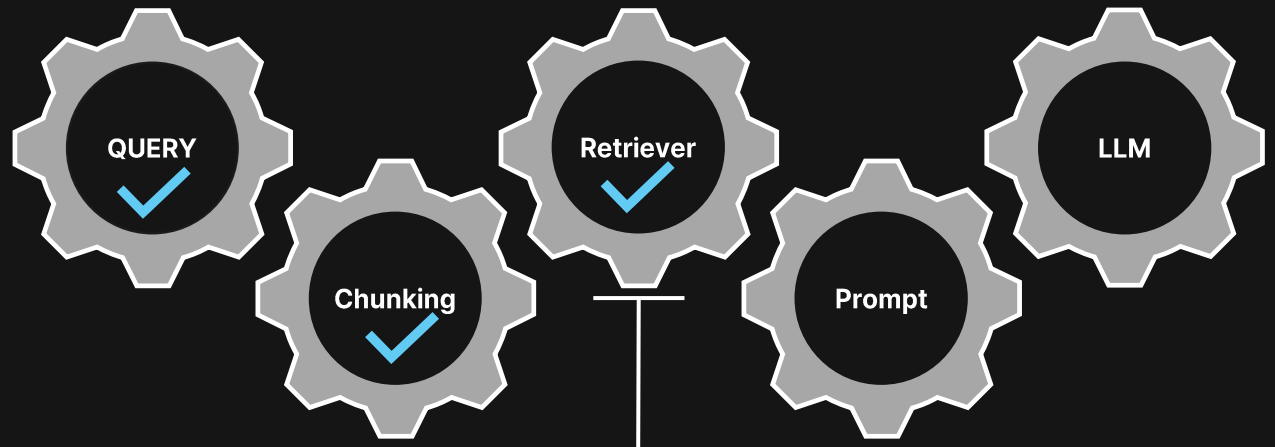


Chunking Evaluation Results

iou_mean	iou_std	recall_mean	recall_std	precision_omeg…	precision_ome…	precision_mean	precision_std	chunker
0.01752579788…	0.01419275701…	0.961430287192…	0.17046205171…	0.053859297981…	0.03544702947…	0.017536946376…	0.01419652744…	FixedTokenChun…

PReP: RAG evaluation Platform

브레인크루 내부 평가 플랫폼



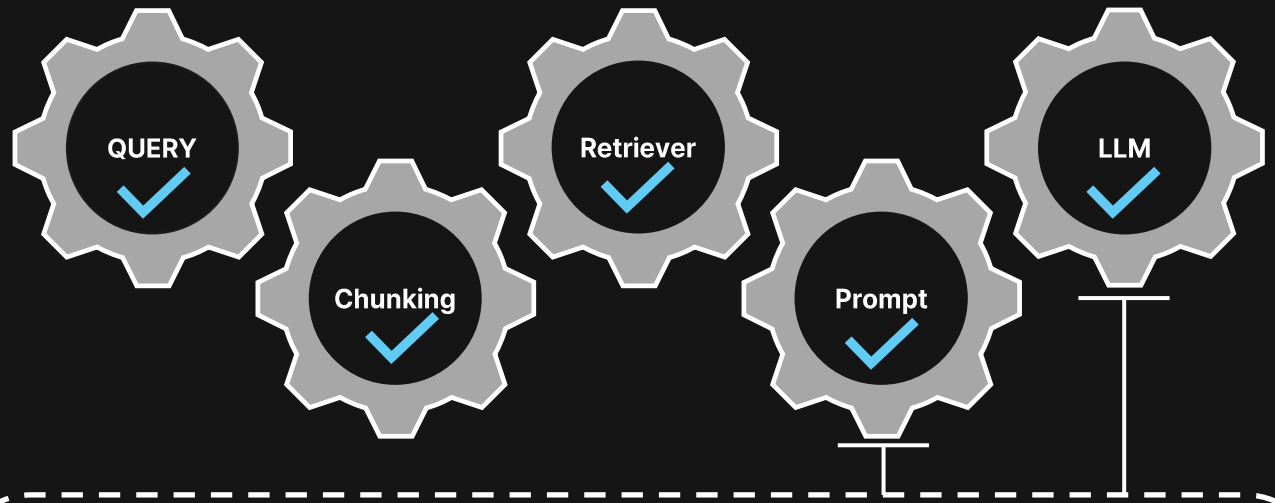
📊 **Retrieval Evaluation Results:**

Retrieval Performance Metrics

Top-K	Precision	Recall	F1	Ndcg	Hit_rate	MRR
1	0.6667	0.6667	0.6667	0.6667	0.6667	0.8333
3	0.3333	1.0000	0.5000	0.8770	1.0000	0.8333
5	0.3333	1.0000	0.5000	0.8770	1.0000	0.8333
10	0.3333	1.0000	0.5000	0.8770	1.0000	0.8333
20	0.3333	1.0000	0.5000	0.8770	1.0000	0.8333

PReP: RAG evaluation Platform

브레인크루 내부 평가 플랫폼



📊 **G-Eval Results:**

G-Eval Metrics

Metric	Average Score	Total Samples
Answer Relevancy	1.0000	3

앞으로의 방향

Thank You