MT-IFEval-Ko

Instruction Following Evaluation
+ Multi-Turn Evaluation
+ Korean Language
+ Langchain/Langfuse Tracing

김상진/AI Data Engineering
2025.07.18

Motivation

- 현업 연관 프로젝트를 하자
  - 데이터 생성
    - Fine-tuning을 위한 훈련 데이터셋 생성
    - 모델 평가를 위한 벤치마크 데이터셋 생성
  - 모델 평가
    - 다양한 기존 LLM의 벤치마크 평가
    - Fine-tuning된 LLM 의 성능 평가

- 결과 분석 도구/환경 필요
  - 단순 평가 결과(accuracy) 측정은 비교적 용이
  - 결과를 분석하고 개선 방안 도출이 필요함
    - Langsmith/Langfuse 를 이용해 보자
    - 기존 평가 툴을 Langchain/Langgraph 로의 변환이 필요하다

- 이번 기회에 Langchain/Langgraph/Langsmith/Langfuse 에 익숙해지자

Instruction-Following Evaluation for Large Language Models

Jeffrey Zhou[§*] Tianjian Lu[♮] Swaroop Mishra[♮] Siddhartha Brahma[♮]
Sujoy Basu[♮] Yi Luan[♮] Denny Zhou[♮] Le Hou[♮†]

[♮]Google [§]Yale University

November 15, 2023

Abstract

One core capability of Large Language Models (LLMs) is to follow natural language instructions. However, the evaluation of such abilities is not standardized: Human evaluations are expensive, slow, and not objectively reproducible, while LLM-based auto-evaluation is potentially biased or limited by the ability of the evaluator LLM. To overcome these issues, we introduce Instruction-Following Eval (**IFEval**) for large language models. IFEval is a straightforward and easy-to-reproduce evaluation benchmark. It focuses on a set of "verifiable instructions" such as "write in more than 400 words" and "mention the keyword of AI at least 3 times". We identified 25 types of those verifiable instructions and constructed around 500 prompts, with each prompt containing one or more verifiable instructions. We show evaluation results of two widely available LLMs on the market. Our code and data can be found at https://github.com/google-research/google-research/tree/master/instruction_following_eval

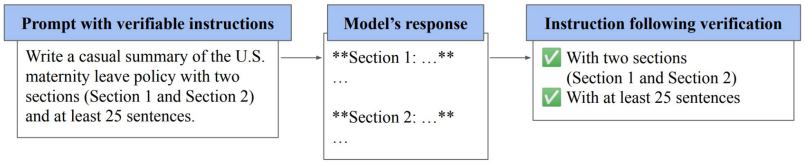


Figure 1: Instructions such as "write at least 25 sentences" can be automatically and objectively verified. We build a set of prompts with verifiable instructions, for evaluating the instruction-following ability of large language models.

Open LLM Leaderboard Archived

Average Top 5

	Rank	Type	Model	Average	IFEval	BBH	MATH	GPQA	MUSR	MMLU...
	1	🔶	MaziyarPanahi/calme-3.2-instruct-78b	52.08 %	80.63 %	62.61 %	40.33 %	20.36 %	38.53 %	70.03 %
	2	💬	MaziyarPanahi/calme-3.1-instruct-78b	51.29 %	81.36 %	62.41 %	39.27 %	19.46 %	36.50 %	68.72 %
	3	💬	dfurman/CalmeRys-78B-Orpo-v0.1	51.23 %	81.63 %	61.92 %	40.63 %	20.02 %	36.37 %	66.80 %
	4	💬	MaziyarPanahi/calme-2.4-rys-78b	50.77 %	80.11 %	62.16 %	40.71 %	20.36 %	34.57 %	66.69 %
	5	🔶	huihui-ai/Qwen2.5-72B-Instruct-abliterated	48.11 %	85.93 %	60.49 %	60.12 %	19.35 %	12.34 %	50.41 %

IFEval Top 5

	Rank	Type	Model	Average	IFEval	BBH	MATH	GPQA	MUSR	MMLU...
	40	💬	meta-llama/Llama-3.3-70B-Instruct	44.85 %	89.98 %	56.56 %	48.34 %	10.51 %	15.57 %	48.13 %
	62	💬	meta-llama/Llama-3.1-70B-Instruct	43.41 %	86.69 %	55.93 %	38.07 %	14.21 %	17.69 %	47.88 %
	7	💬	MaziyarPanahi/calme-2.1-qwen2.5-72b	47.86 %	86.62 %	61.66 %	59.14 %	15.10 %	13.30 %	51.32 %
	61	🔶	VAGOsolutions/Llama-3.1-Sauerkraut LM-70b-Instruc	43.41 %	86.56 %	57.24 %	36.93 %	12.19 %	19.39 %	48.17 %
	6	💬	Qwen/Qwen2.5-72B-Instruct	47.98 %	86.38 %	61.87 %	59.82 %	16.67 %	11.74 %	51.40 %

Strict Accuracy: 엄격하게 판단

- **Prompt-level** strict-accuracy: 현재 프롬프트에서 요구한 지시사항들을 "모두" 따랐다면 Pass
- **Inst-level** strict-accuracy: 현재 프롬프트에서 요구한 "개별" 지시사항별로 following 여부를 판단

Loose Accuracy: 관대하게 판단

- **Prompt-level** loose-accuracy: Prompt-level accuracy computed with the loose criterion.
- **Inst-level** loose-accuracy: Instruction-level accuracy computed with a loose criterion.

The list of 25 verifiable instructions

Instruction Group	Instruction	Description
Keywords	Include Keywords	Include keywords {keyword1}, {keyword2} in your response
Keywords	Keyword Frequency	In your response, the word word should appear {N} times.
Keywords	Forbidden Words	Do not include keywords {forbidden words} in the response.
Keywords	Letter Frequency	In your response, the letter {letter} should appear {N} times.
Language	Response Language	Your ENTIRE response should be in {language}, no other language is allowed.
Length Constraints	Number Paragraphs	Your response should contain {N} paragraphs. You separate paragraphs using the markdown divider: * * *
Length Constraints	Number Words	Answer with at least / around / at most {N} words.
Length Constraints	Number Sentences	Answer with at least / around / at most {N} sentences.
Length Constraints	Number Paragraphs + First Word in i-th Paragraph	There should be {N} paragraphs. Paragraphs and only paragraphs are separated with each other by two line breaks. The {i}-th paragraph must start with word {first_word}.
Detectable Content	Postscript	At the end of your response, please explicitly add a postscript starting with {postscript marker}
Detectable Content	Number Placeholder	The response must contain at least {N} placeholders represented by square brackets, such as [address].
Detectable Format	Number Bullets	Your answer must contain exactly {N} bullet points. Use the markdown bullet points such as: * This is a point.
Detectable Format	Title	Your answer must contain a title, wrapped in double angular brackets, such as <<poem of joy>>.

Instruction Group	Instruction	Description
Detectable Format	Choose From	Answer with one of the following options: {options}
Detectable Format	Minimum Number Highlighted Section	Highlight at least {N} sections in your answer with markdown, i.e. *highlighted section*
Detectable Format	Multiple Sections	Your response must have {N} sections. Mark the beginning of each section with {section_splitter} X.
Detectable Format	JSON Format	Entire output should be wrapped in JSON format.
Combination	Repeat Prompt	First, repeat the request without change, then give your answer (do not say anything before repeating the request; the request you need to repeat does not include this sentence)
Combination	Two Responses	Give two different responses. Responses and only responses should be separated by 6 asterisk symbols: ******.
Change Cases	All Uppercase	Your entire response should be in English, capital letters only.
Change Cases	All Lowercase	Your entire response should be in English, and in all lowercase letters. No capital letters are allowed.
Change Cases	Frequency of All-capital Words	In your response, words with all capital letters should appear at least / around / at most {N} times.
Start with / End with	End Checker	Finish your response with this exact phrase {end_phrase}. No other words should follow this phrase.
Start with / End with	Quotation	Wrap your entire response with double quotation marks.
Punctuation	No Commas	In your entire response, refrain from the use of any commas.

Multi-language

7/FEB/2025

M-IFEval: Multilingual Instruction-Following Evaluation

Antoine Dussolle[1,2*], Andrea Cardeña Díaz[1], Shota Sato[1], Peter Devine[1],

[1]Lightblue KK., [2]Arkema,
antoine.dussolle@gmail.com
andrea.cdiaz@hotmail.es
{shota.sato,peter}@lightblue-tech.com

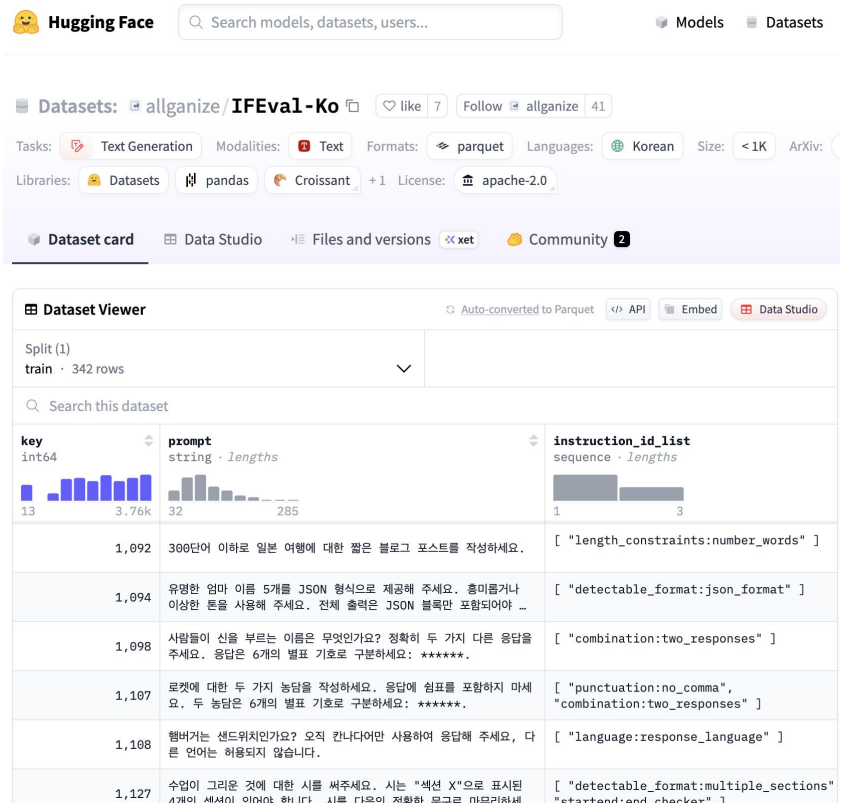
Multi-turn/Multi-language

13/NOV/2024

Multi-IF: Benchmarking LLMs on Multi-Turn and Multilingual Instructions Following

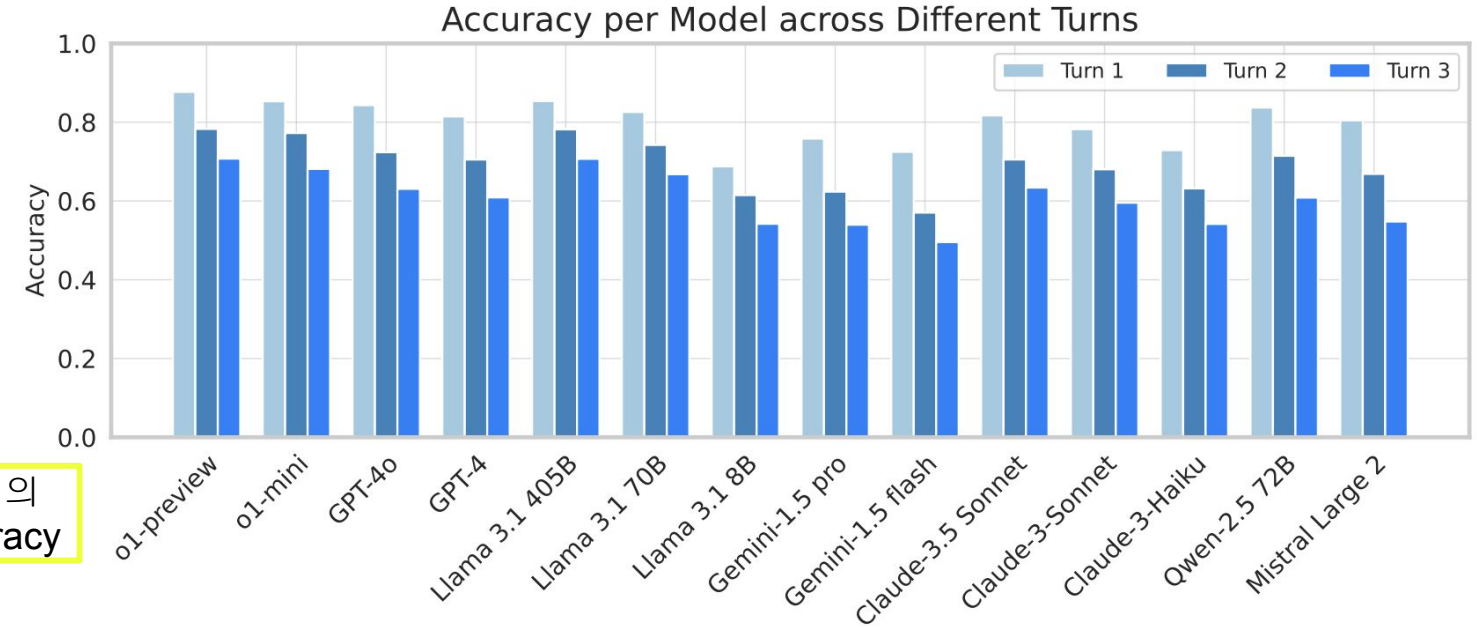
Yun He*, Di Jin*, Chaoqi Wang*, Chloe Bi*, Karishma Mandyam, Hejia Zhang, Chen Zhu, Ning Li, Tengyu Xu, Hongjiang Lv, Shruti Bhosale, Chenguang Zhu, Karthik Abinav Sankararaman, Eryk Helenowski, Melanie Kambadur, Aditya Tayade, Hao Ma, Han Fang, Sinong Wang

Meta GenAI
*Co-first Author



한국어, multi-turn 평가 데이터셋 부재

Multi-IF: Multi-turn & Multilingual Instruction-Following Evaluation

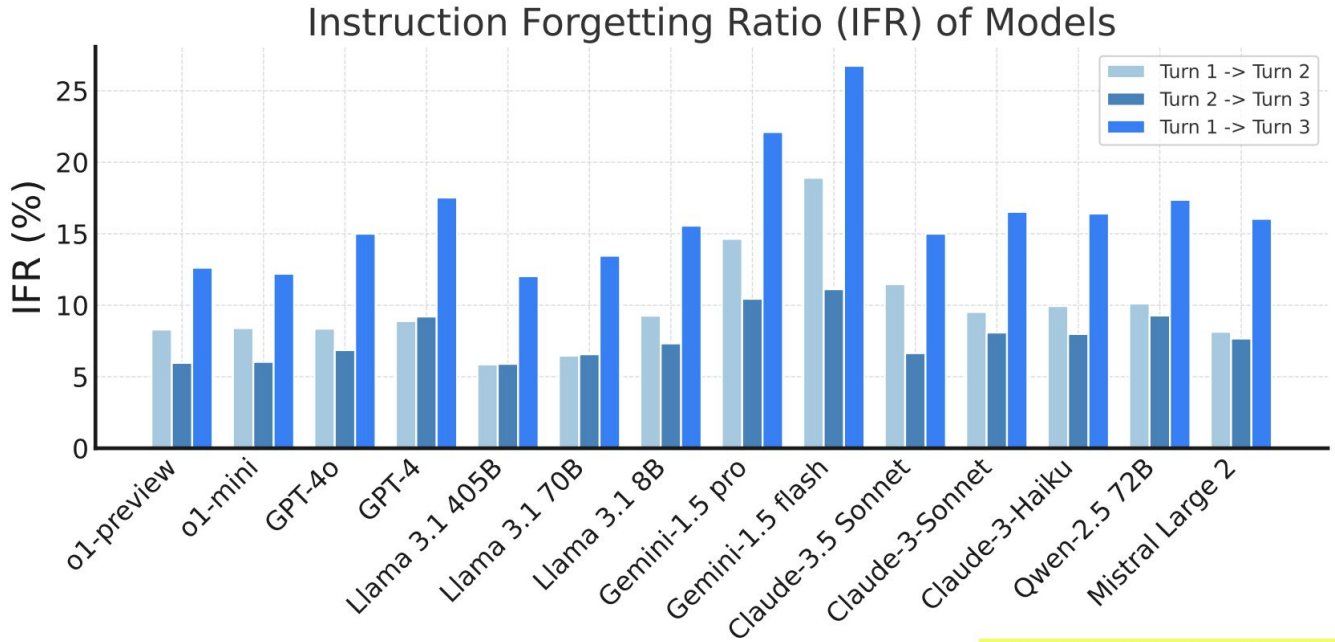


모든 언어의 평균 accuracy

Figure 6 The impact of multi-turns on instruction following. "Accuracy" means the average of the final metric of the all languages; The final metric is the average of the four accuracy scores: instruction-level strict accuracy, conversation-level strict accuracy, instruction-level loose accuracy, and conversation-level loose accuracy.

Turn 이 늘어날수록 instruction following이 떨어진다 → 이전 instruction을 잊는다.

Multi-IF: Multi-turn & Multilingual Instruction-Following Evaluation



turn-3에서 turn-1을 계속 기억하고 following 하는가?

이걸 측정하는 python코드는 github에 공개되지 않음

Multi-turn에서 turn이 늘어날수록 최초 instruction을 잘 잊는 모델은?
→ Gemini 모델들

Instruction Forgetting Ratio

이전 turn에서 follow했으나, 후속 turn에서 not following한 instruction 수

$$\text{IFR} = \left(\frac{\text{Number of Previously Followed Instructions Not Followed in Subsequent Turns}}{\text{Total Number of Instructions Followed in Previous Turn}} \right) \times 100$$

이전 turn 에서 follow했던 instruction 수

MT-IFEval-Ko 구성

데이터셋 준비

- Langfuse 서버 연동
- Input/Output/Metadata 활용

배치 평가 수행

- Langfuse evaluator 작성
  - Langchain 기반
- Langfuse tracing 구현
  - Multi-turn 고려
- Batch evaluation
  - 전체 dataset 일괄 평가

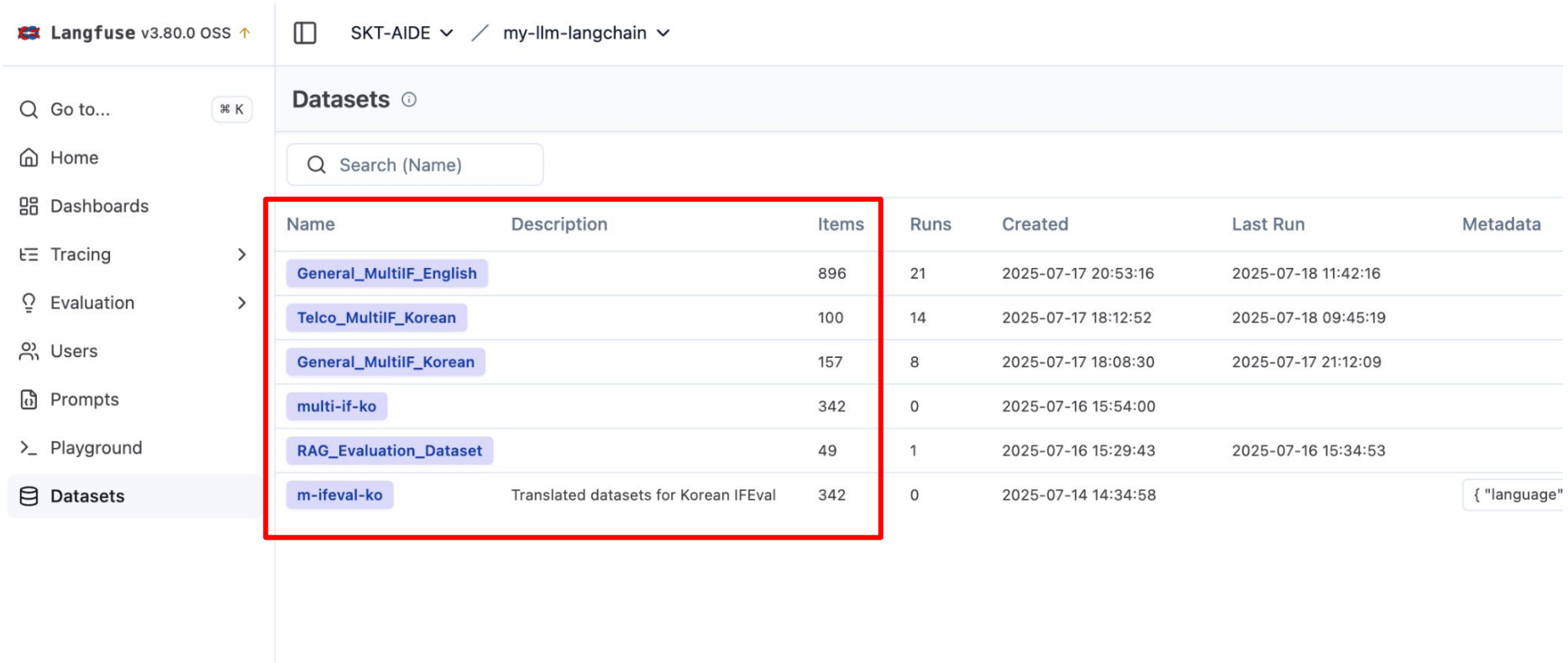
데모 평가 수행

- Gradio 데모 작성
- Run-time evaluation
  - 특정 item 평가

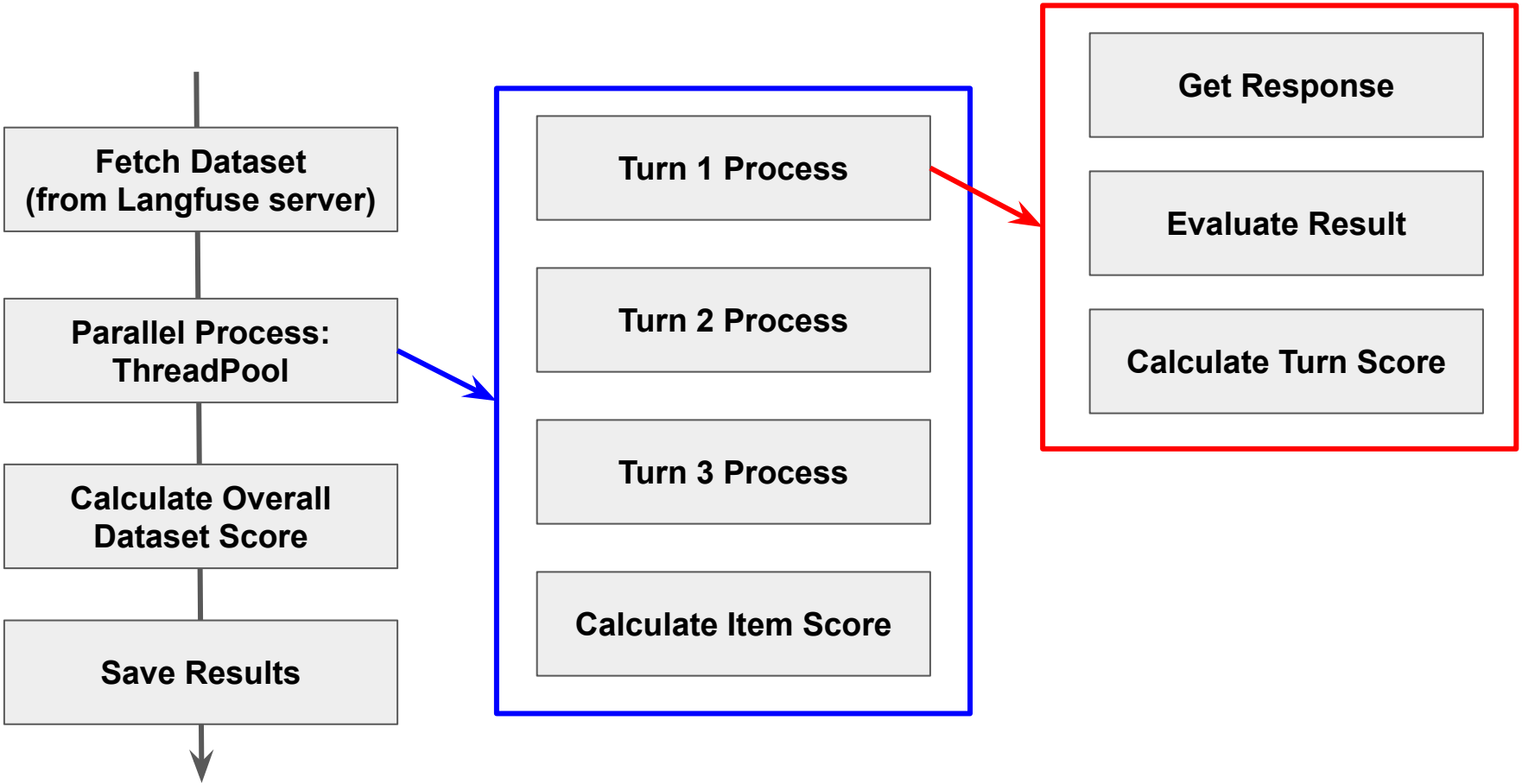
MT-IFEval Dataset Example


```
{
  "INDEX": "TelcoIF_001",
  "turn": "turn_1",
  "turn_1_prompt": {
    "role": "user",
    "content": "114에서 온 부재중 전화의 사유를 확인할 수 있나요?\n위 질문에 대한 답변을 반드시 영어로만 작성하세요."
  },
  "turn_1_instruction_id_list": [
    "language:response_language"
  ],
  "turn_1_kwargs": [
    {
      "language": "en"
    }
  ],
  "turn_2_prompt": {
    "role": "user",
    "content": "응답의 가장 마지막 문단을 \"P.S.\"로 시작하는 추신 형태로 적어주세요."
  },
  "turn_2_instruction_id_list": [
    "language:response_language",
    "detectable_content:postscript"
  ],
  "turn_2_kwargs": [
    {
      "language": "en"
    },
    {
      "postscript_marker": "P.S."
    }
  ],
```

Langfuse Dataset

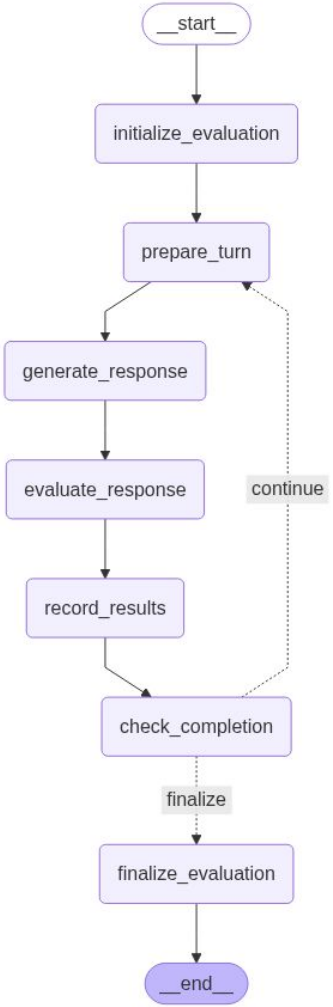


Evaluation Process



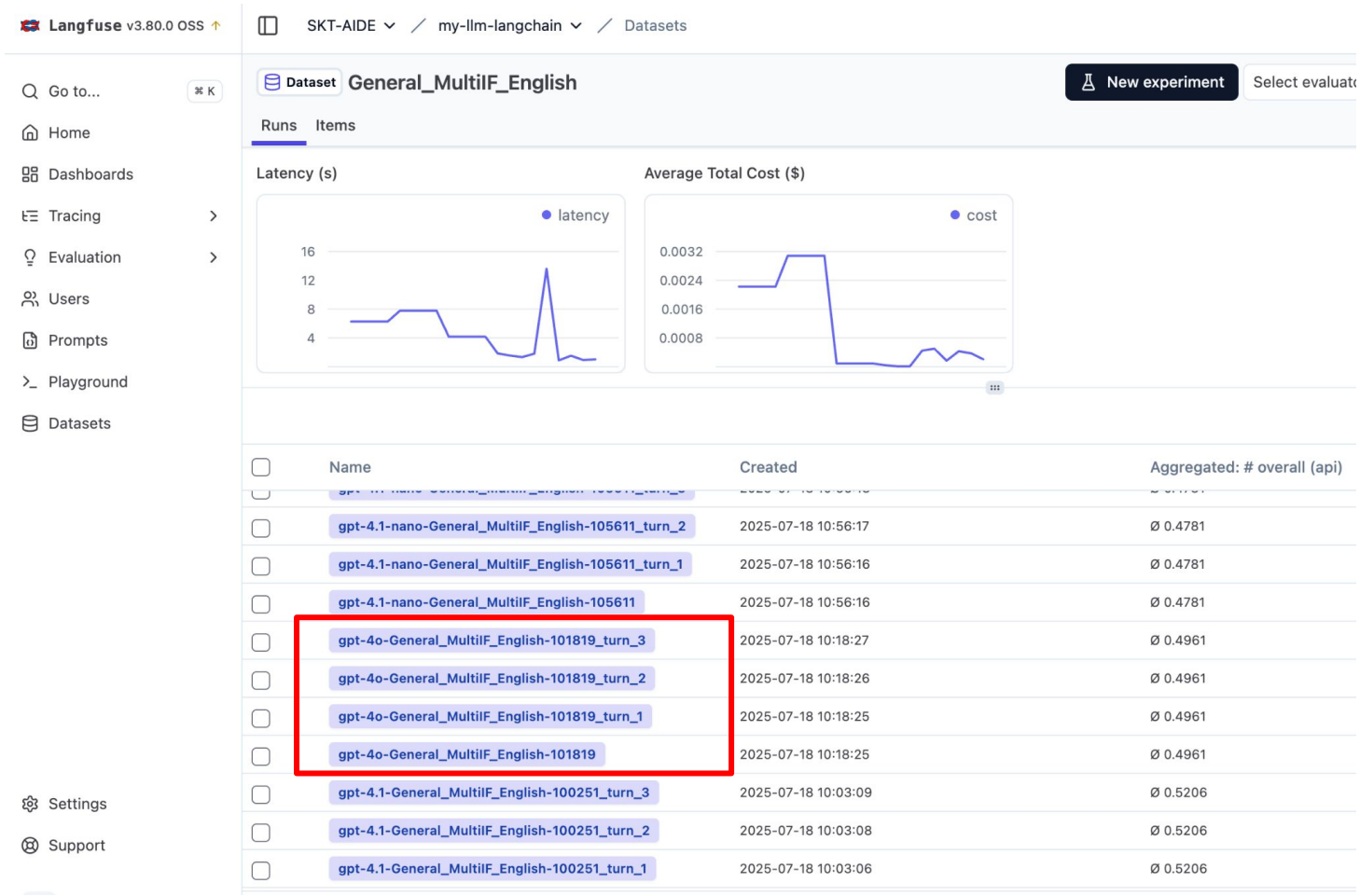
Multi-turn Messages

Langgraph Implementation

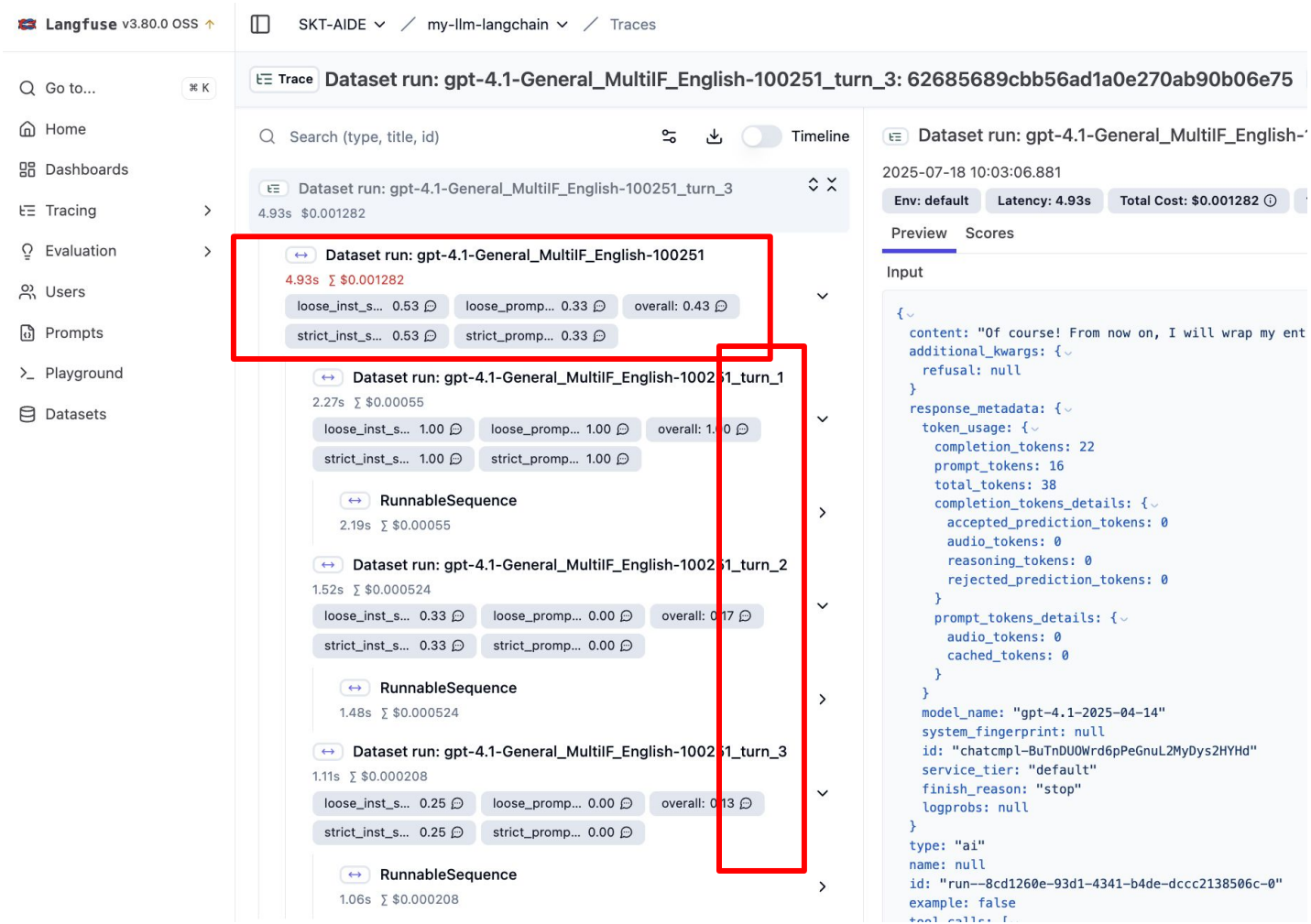



```python
def _create_workflow(self):
    """LangGraph 워크플로우 생성"""
    workflow = StateGraph(GraphState)

    # 노드 추가
    workflow.add_node("initialize_evaluation", self._initialize_evaluation)
    workflow.add_node("prepare_turn", self._prepare_turn)
    workflow.add_node("generate_response", self._generate_response)
    workflow.add_node("evaluate_response", self._evaluate_response)
    workflow.add_node("record_results", self._record_results)
    workflow.add_node("check_completion", self._check_completion)
    workflow.add_node("finalize_evaluation", self._finalize_evaluation)

    # 엣지 추가
    workflow.add_edge(START, "initialize_evaluation")
    workflow.add_edge("initialize_evaluation", "prepare_turn")
    workflow.add_edge("prepare_turn", "generate_response")
    workflow.add_edge("generate_response", "evaluate_response")
    workflow.add_edge("evaluate_response", "record_results")
    workflow.add_edge("record_results", "check_completion")

    # 조건부 엣지 추가
    workflow.add_conditional_edges(
        "check_completion",
        self._should_continue,
        {
            "continue": "prepare_turn",
            "finalize": "finalize_evaluation"
        }
    )

    workflow.add_edge("finalize_evaluation", END)

    # 메모리 체크포인트 설정
    memory = MemorySaver()

    return workflow.compile(checkpointer=memory)
```

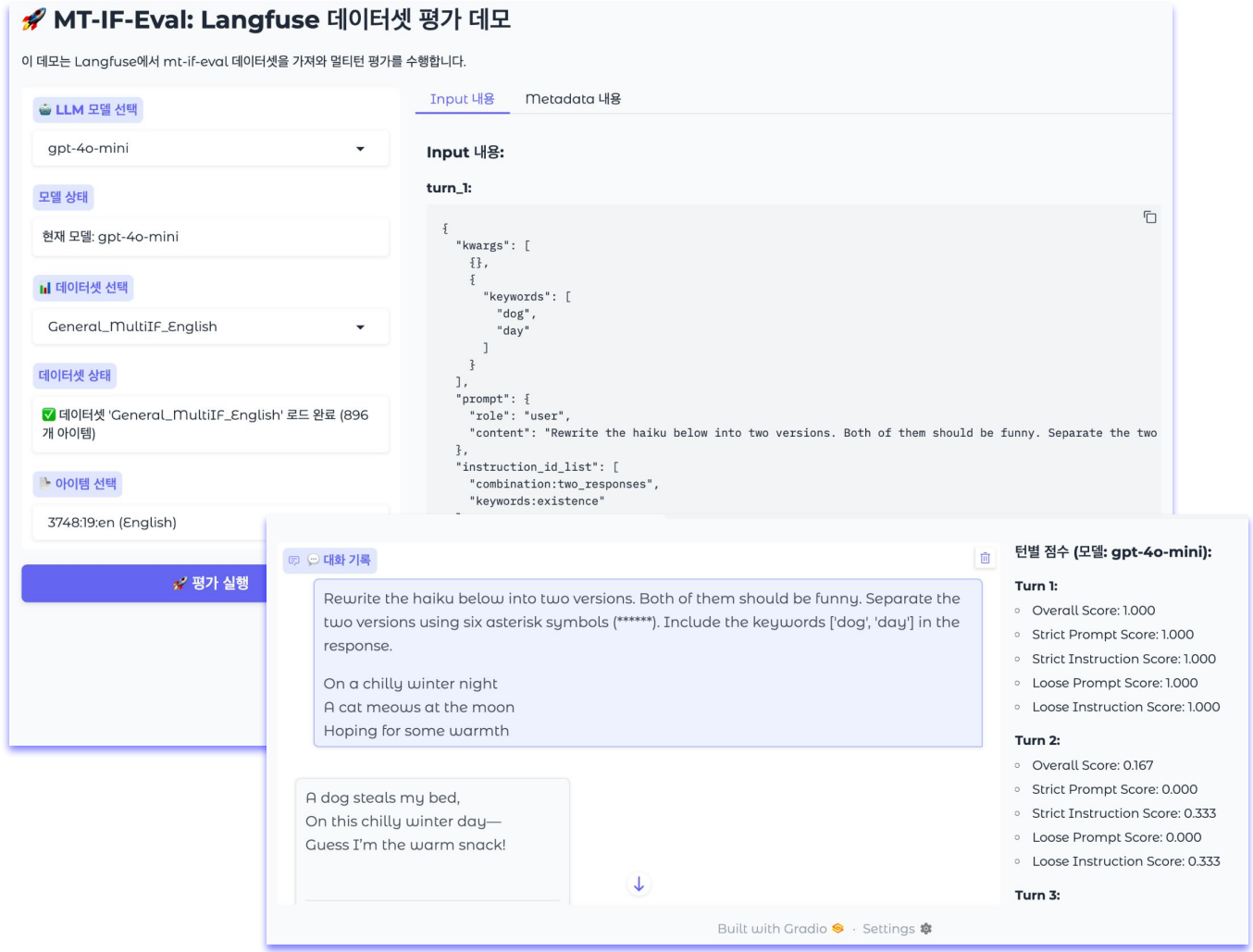
Langfuse Dataset Score



Langfuse Tracing



Gradio Run-time Demo



Summary

- Langchain/Langgraph 기반 workflow 이해
  - chain/node 연결 기반의 workflow
  - 코드 가독성 향상
- Langfuse/Langsmith 모니터링 연동
  - 평가 결과 분석 효율화
- 현업 업무에 연계 방안
  - 평가 데이터셋 개발에 활용
    - 실시간 프롬프트 테스트 및 평가 결과 확인
    - 어플리케이션별 평가 데이터셋 개발 효율화
  - 평가 결과 분석에 활용
    - LLM 모델별 성능 비교 분석 효율화
    - Fine-tuning LLM 모델 성능 개선을 위한 분석
    - Prompt-optimization 을 위한 분석

TODO

Langchain → Langgraph

- Langgraph 형태로 변환 (디버깅 중)

Gradio Dataset Studio

- Turn별 prompt/instruction/kwargs 입력/편집
- Turn별 LLM 평가 결과 실시간 확인
- 개별 데이터셋을 Langfuse 데이터셋에 업데이트/관리

Korean MT-IFEval Dataset

- 한국어 평가를 위한 데이터셋 개발 (general-domain)
- Telco 서비스용 LLM 모델 평가용 데이터셋 개발 (application-domain)