SURROMIND

Language model Evaluation And Peft

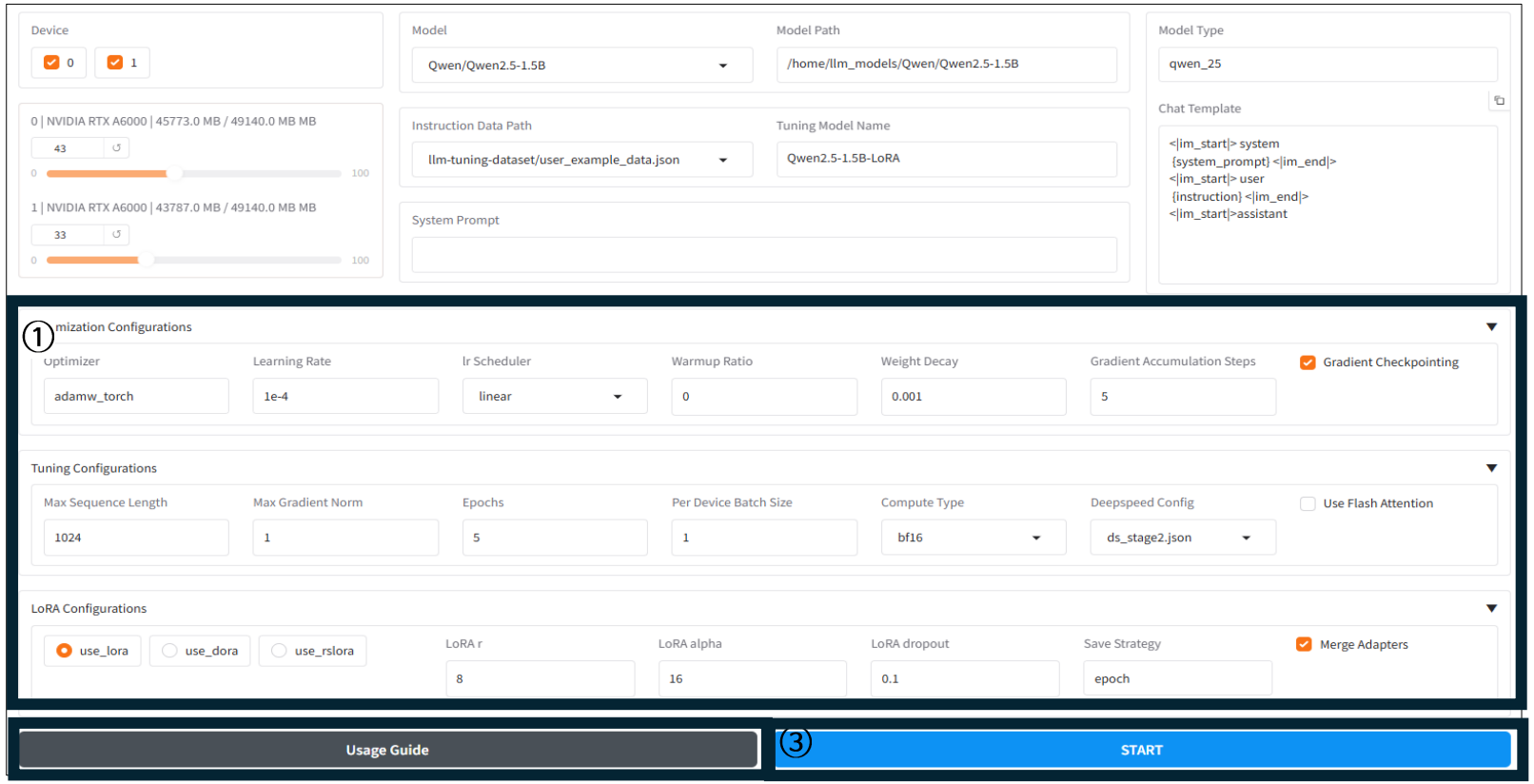
기능 설명

SURROMIND

LLM 학습 시스템: LLM PEFT(1)

기능 1: LLM PEFT 작업 수행

PEFT 기법을 활용해 평가 데이터셋에 대해 LLM의 튜닝 작업 수행



① 직관적인 GUI를 통해
사용자가 튜닝 옵션을 손쉽게 선택 가능

② 각 옵션에 대한 설명을 담은
Usage guide 제공

SURROMIND

LLM 학습 시스템: LLM PEFT(2)

기능 1: LLM PEFT 작업 수행

PEFT 기법을 활용해 평가 데이터셋에 대해 LLM의 튜닝 작업 수행

③ Loss, Grad norm 그래프를 통한 학습 과정 모니터링

④ Log, Progress bar를 통한 학습 과정 모니터링

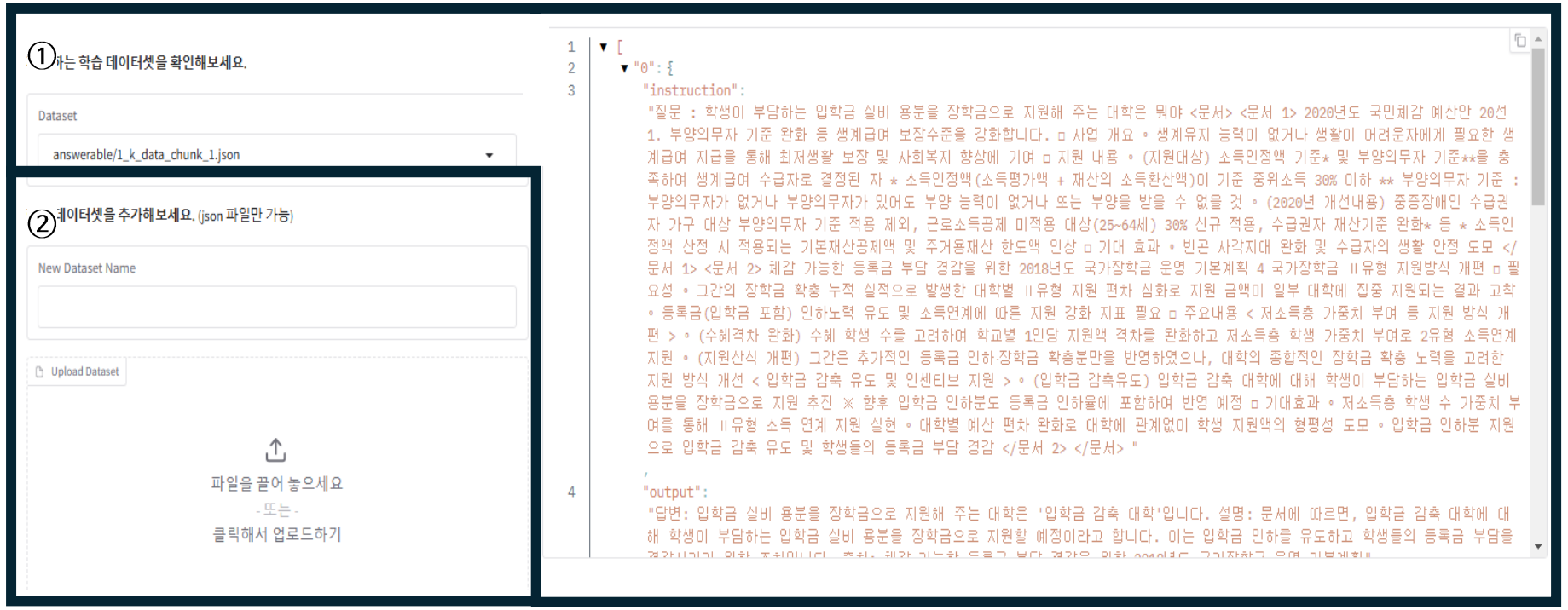
SURROMIND

LLM 학습 시스템: 학습 데이터셋 관리

기능 2: 학습 데이터셋 예시 확인, 추가

지원 데이터셋의 예시 제공

사용자가 GUI를 통하여 데이터셋을 업로드하면 자동 검증 후 지원 데이터셋에 추가



① 학습 데이터셋 확인

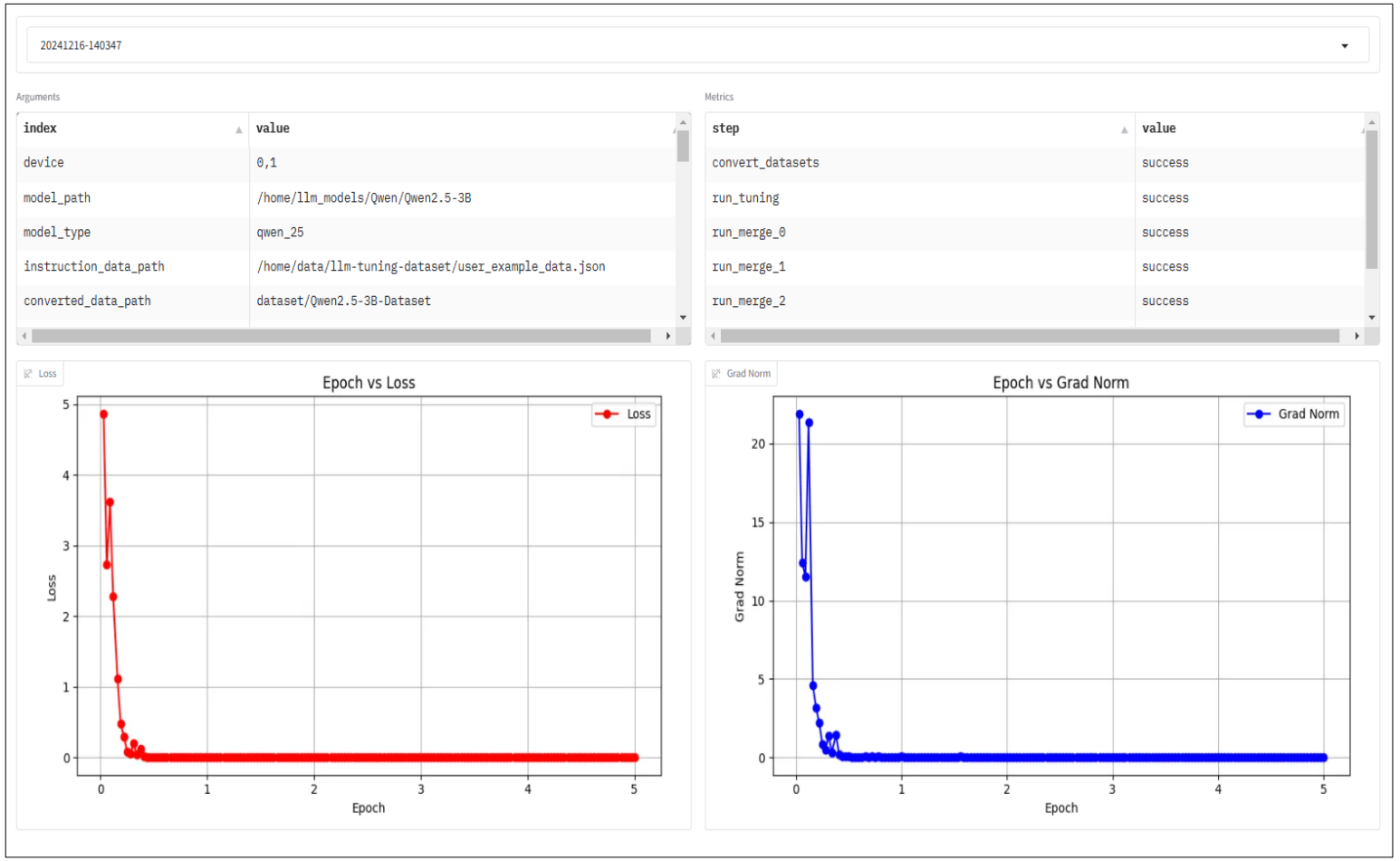
② 학습 데이터셋
검증 후 추가

SURROMIND

LLM 학습 시스템: 이전 학습 기록 조회

기능 3: 이전 학습 기록 조회

과거 학습 작업의 입력 변수, 성공 여부, Loss 및 gradient norm 그래프 제공



이전 학습 기록 확인

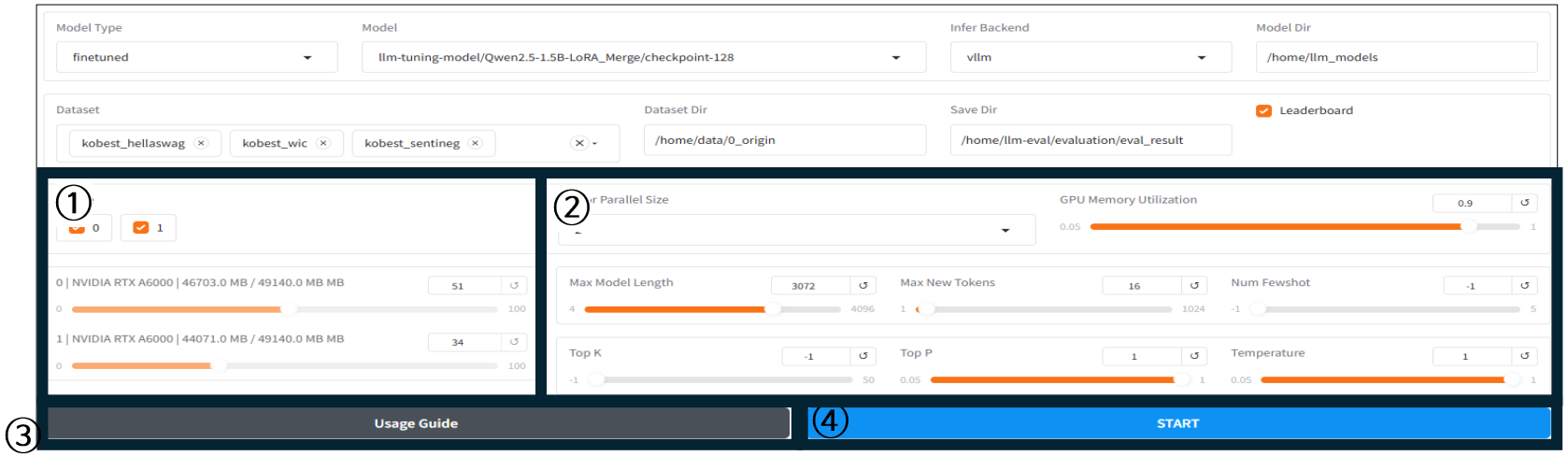
- Arguments
- 성공/실패 여부
- Loss, Grad norm 그래프

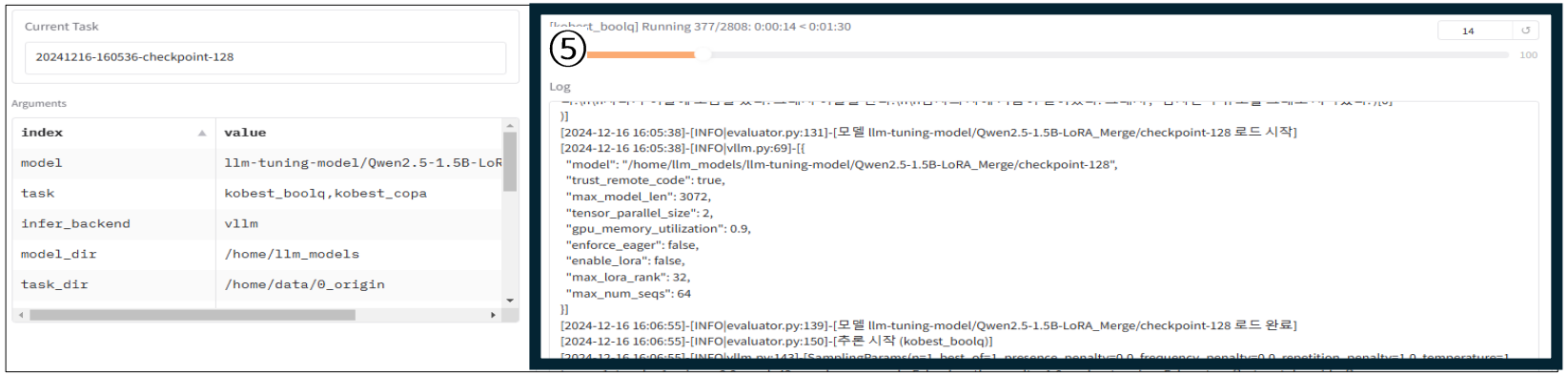
SURROMIND

LLM 평가 시스템: LLM 평가

기능 1: 한국어 벤치마크 데이터셋에 대한 LLM 평가

특정 데이터셋에 대해 LLM의 추론 작업을 수행하고 평가 지표 계산





① GPU 상태 확인 후
사용 Device 선택

② 직관적인 GUI를 통해 사용자가
평가 옵션을 손쉽게 선택

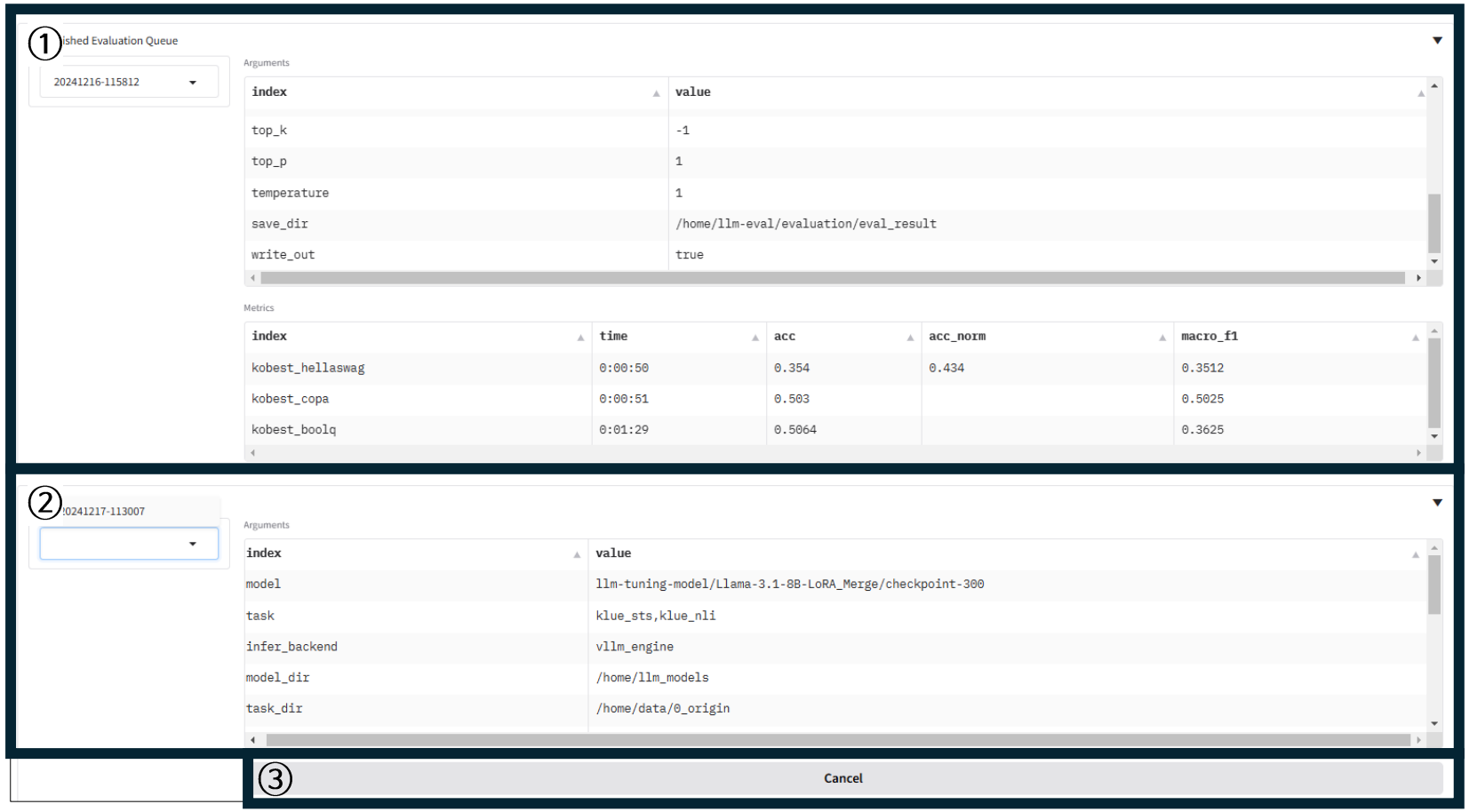
③ 각 옵션에 대한 설명을 담은
Usage guide 제공
④ 평가 작업 예약

⑤ 평가 진행 시, 진행 과정을
progress bar와 log를 통해
모니터링

SURROMIND

LLM 평가 시스템: 작업 예약

기능 2: 평가 작업 예약
평가 작업 예약 및 예약 취소 기능 제공



① 완료된 평가 작업의 입력 Arguments 및 결과 지표 확인

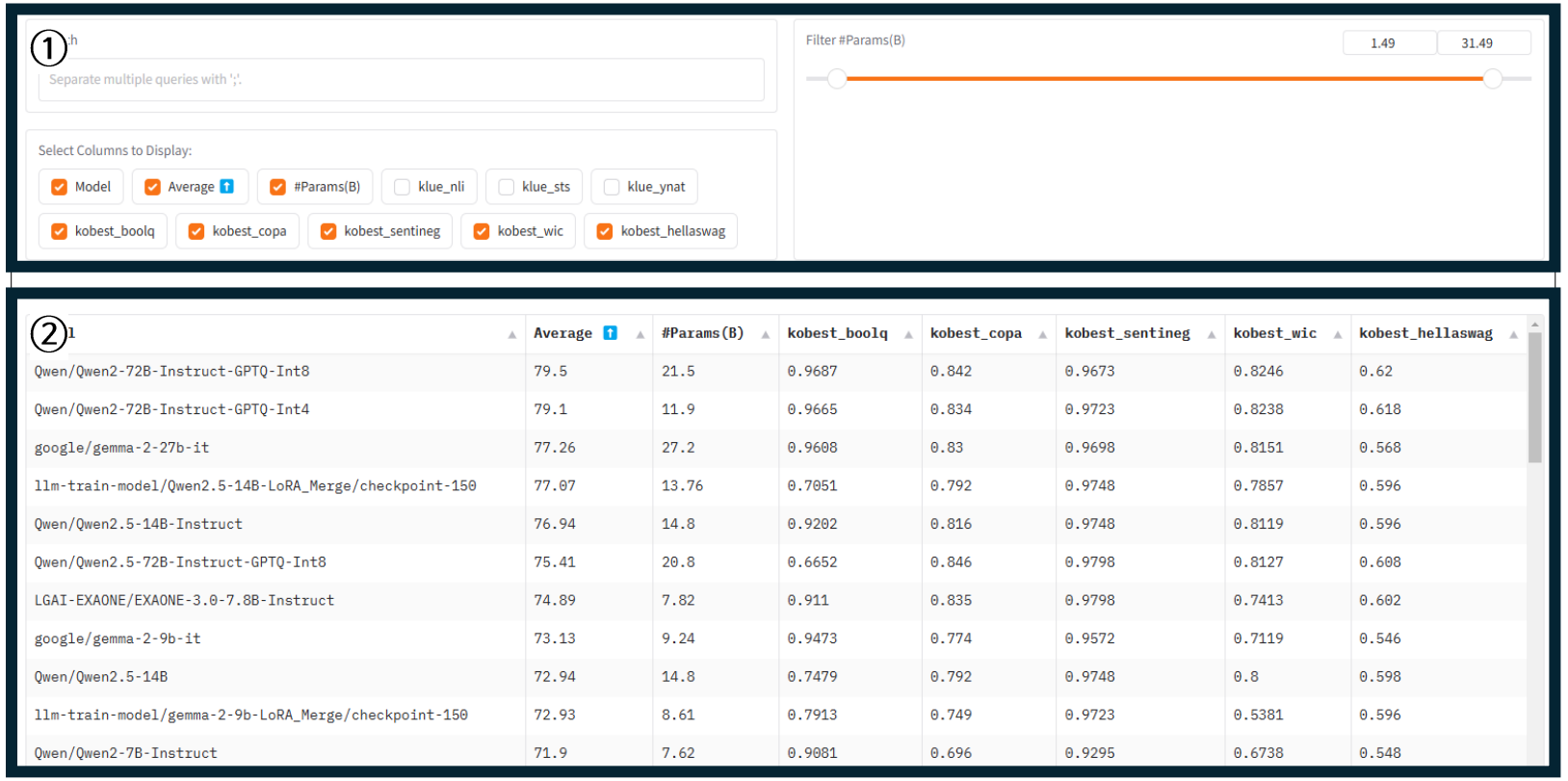
② 예약된 작업 목록 확인

③ 예약 취소

SURROMIND

LLM 평가 시스템: LLM 리더보드

기능 3: LLM 리더보드 제공
평가 결과를 리더보드 형태로 시각화하여 제공



① 파라미터 수 및 Task, 모델명에 따라 결과 필터링 가능

② 모델별 평가 결과를 리더보드 형태로 확인 가능

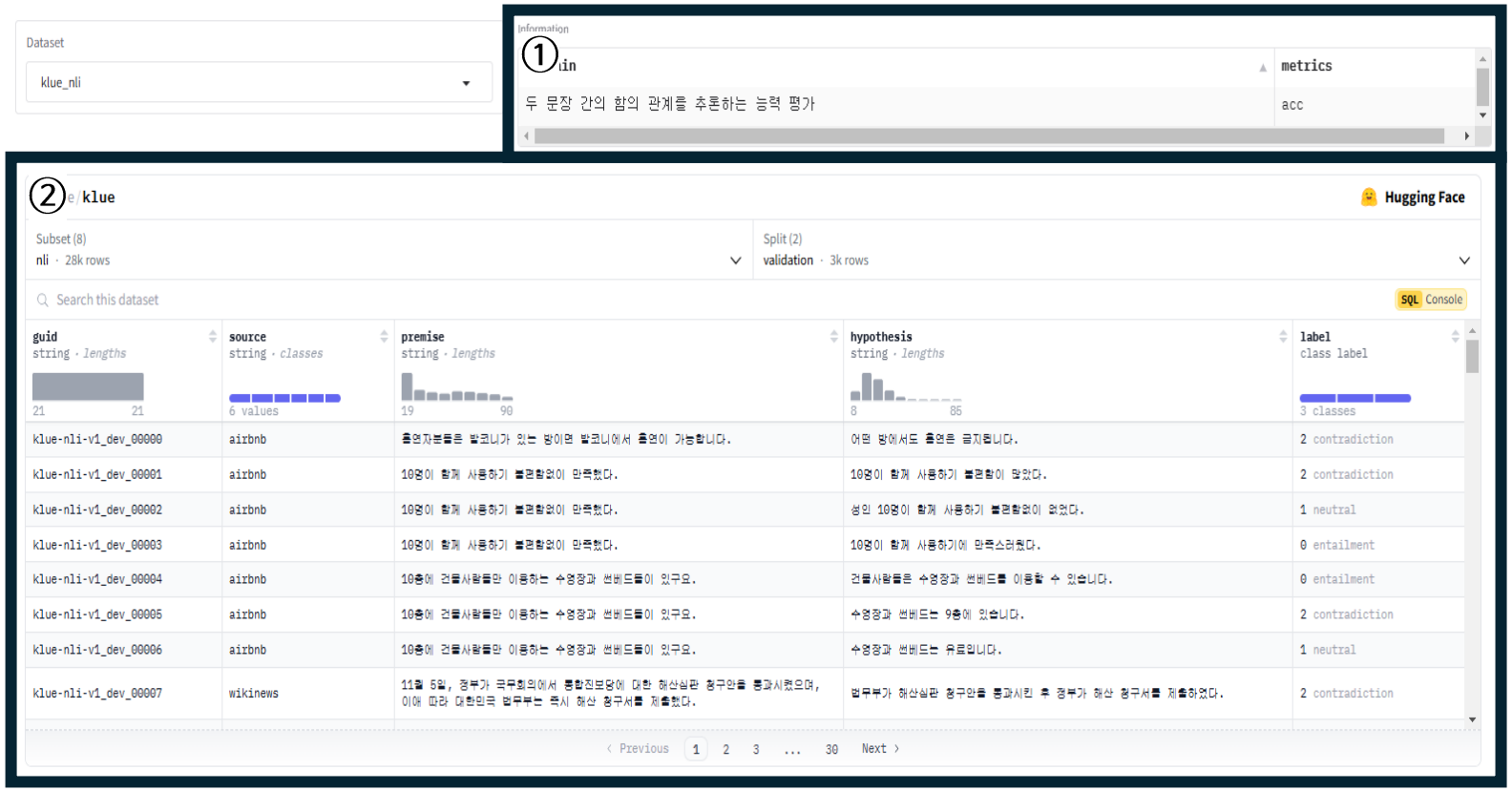
	Average	#Params(B)	kobest_boolq	kobest_copa	kobest_sentineg	kobest_wic	kobest_hellaswag
Qwen/Qwen2-72B-Instruct-GPTQ-Int8	79.5	21.5	0.9687	0.842	0.9673	0.8246	0.62
Qwen/Qwen2-72B-Instruct-GPTQ-Int4	79.1	11.9	0.9665	0.834	0.9723	0.8238	0.618
google/gemma-2-27b-it	77.26	27.2	0.9608	0.83	0.9698	0.8151	0.568
llm-train-model/Qwen2.5-14B-LoRA_Merge/checkpoint-150	77.07	13.76	0.7051	0.792	0.9748	0.7857	0.596
Qwen/Qwen2.5-14B-Instruct	76.94	14.8	0.9202	0.816	0.9748	0.8119	0.596
Qwen/Qwen2.5-72B-Instruct-GPTQ-Int8	75.41	20.8	0.6652	0.846	0.9798	0.8127	0.608
LGAI-EXAONE/EXAONE-3.0-7.8B-Instruct	74.89	7.82	0.911	0.835	0.9798	0.7413	0.602
google/gemma-2-9b-it	73.13	9.24	0.9473	0.774	0.9572	0.7119	0.546
Qwen/Qwen2.5-14B	72.94	14.8	0.7479	0.792	0.9748	0.8	0.598
llm-train-model/gemma-2-9b-LoRA_Merge/checkpoint-150	72.93	8.61	0.7913	0.749	0.9723	0.5381	0.596
Qwen/Qwen2-7B-Instruct	71.9	7.62	0.9081	0.696	0.9295	0.6738	0.548

SURROMIND

LLM 평가 시스템: 평가 데이터셋 조회

기능 4: 지원 데이터셋 정보 확인

지원 데이터셋의 평가 지표, 능력 평가 범위, 예시를 제공



① 벤치마크 데이터셋별
평가 능력 및 평가 지표 확인

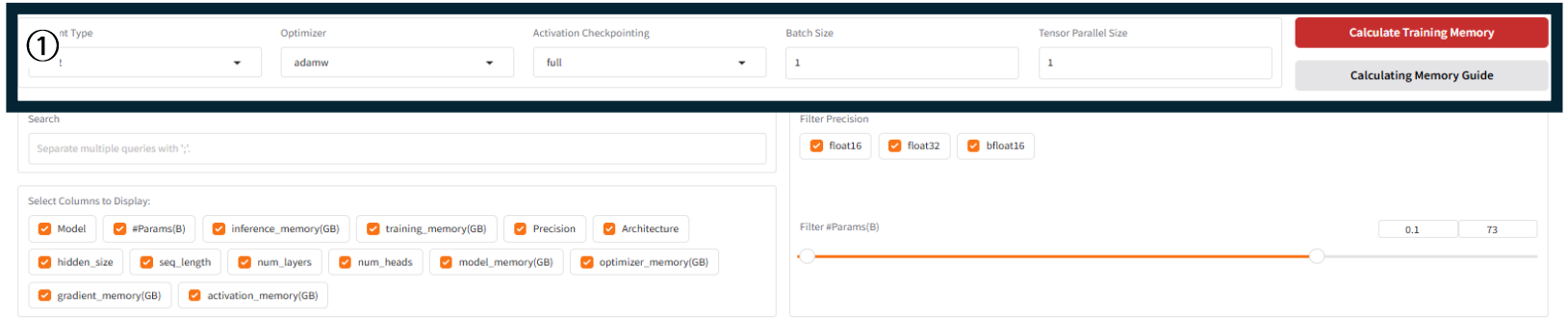
② 벤치마크 데이터셋별
예시 확인

SURROMIND

LLM 평가 시스템: 모델 정보 확인

기능 5: 모델별 GPU 메모리 소요량 추정

모델별 학습 또는 추론에 소요되는 GPU 메모리 추정량 제공



① Gradient Type, Optimizer, Activation Checkpointing, Batch Size, Tensor Parallel Size를 고려하여 LLM을 학습하는 데에 소요되는 GPU 메모리 추정

②

	#Params(B)	inference_memory(GB)	training_memory(GB)	Precision	Architecture	hidden_size	seq_length	num_layers	num_heads	model_memory(GB)	optimize
CohereForAI/c4ai-command-r-08-2024	32.3	77.52	661.4	float16	CohereForCausalLM	8192	131072	40	64	64.6	387.6
CohereForAI/c4ai-command-r-plus-08-2024	104	249.6	2064	float16	CohereForCausalLM	12288	131072	64	96	208	1248
LGAI-EXAONE/EXAONE-3.0-7.8B-Instruct	7.82	18.77	157.4	float32	ExaoneForCausalLM	4096	4096	32	32	31.28	93.84
Qwen/Qwen2-0.5B	0.49	1.18	14.07	bfloat16	Qwen2ForCausalLM	896	131072	24	14	0.98	5.88
Qwen/Qwen2-0.5B-Instruct	0.49	1.18	10.13	bfloat16	Qwen2ForCausalLM	896	32768	24	14	0.98	5.88
Qwen/Qwen2-1.5B	1.54	3.7	38.22	bfloat16	Qwen2ForCausalLM	1536	131072	28	12	3.08	18.48
Qwen/Qwen2-1.5B-Instruct	1.54	3.7	30.35	bfloat16	Qwen2ForCausalLM	1536	32768	28	12	3.08	18.48
Qwen/Qwen2-72B-Instruct-GPTQ-Int4	11.9	28.56	254.2	float16	Qwen2ForCausalLM	8192	32768	80	64	23.8	142.8
Qwen/Qwen2-72B-Instruct-GPTQ-Int8	21.5	51.6	427	float16	Qwen2ForCausalLM	8192	32768	80	64	43	258
Qwen/Qwen2-7B	7.62	18.29	161.66	bfloat16	Qwen2ForCausalLM	3584	131072	28	28	15.24	91.44
Qwen/Qwen2-7B-Instruct	7.62	18.29	143.28	bfloat16	Qwen2ForCausalLM	3584	32768	28	28	15.24	91.44
Qwen/Qwen2.5-0.5B	0.49	1.18	10.13	bfloat16	Qwen2ForCausalLM	896	32768	24	14	0.98	5.88

② 모델 config에서 확인할 수 있는 정보 및 추론 및 학습에 소요되는 GPU 메모리 추정량 확인 (아래 수식 고려하여 계산)

$$\text{Total Memory}_{\text{Inference}} \approx 1.2 \times \text{Model Memory}.$$

$$\text{Total Memory}_{\text{Training}} = \text{Model Memory} + \text{Optimizer Memory} + \text{Activation Memory} + \text{Gradient Memory}$$

SURROMIND

감사합니다

문의사항이 있으시다면
아래 주소로 연락 부탁드립니다.

sychung@surromind.ai