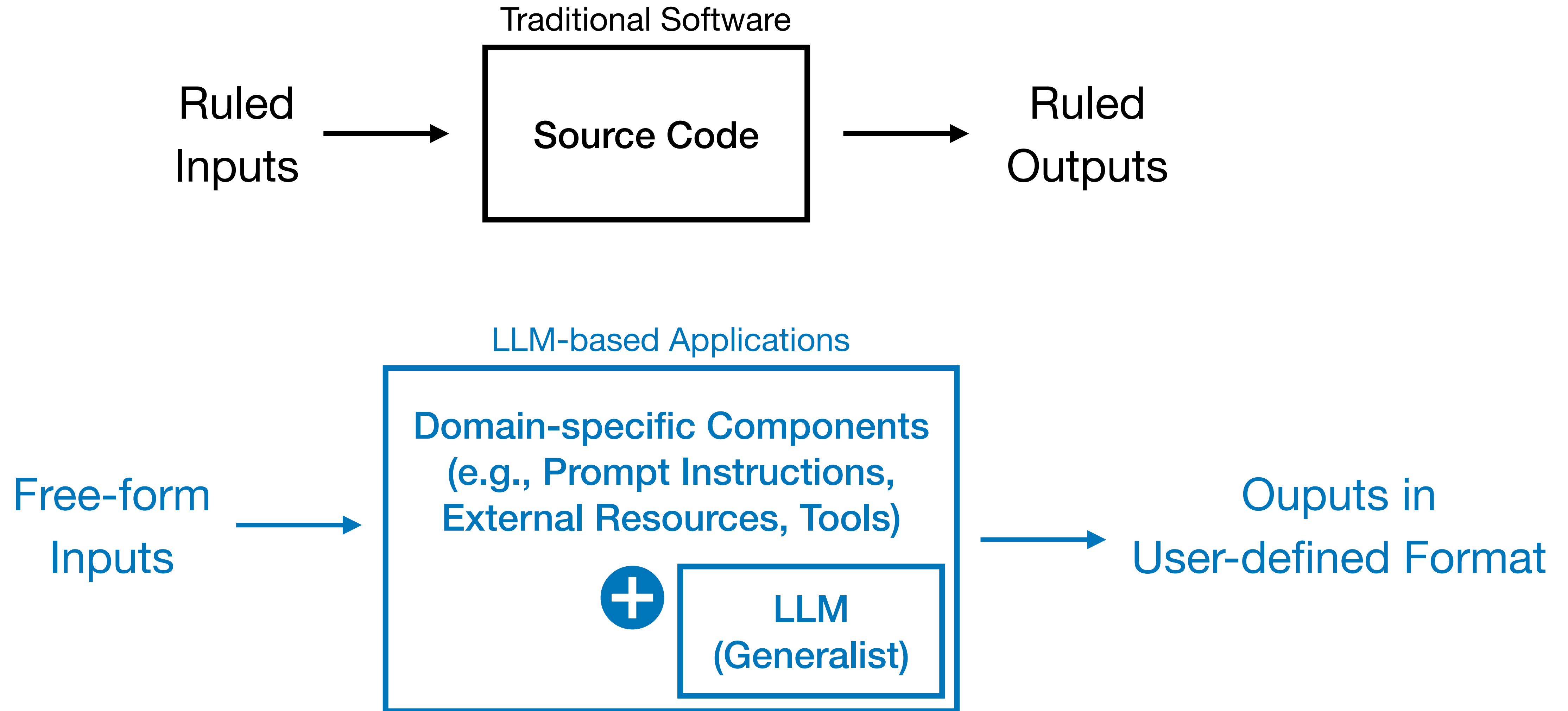


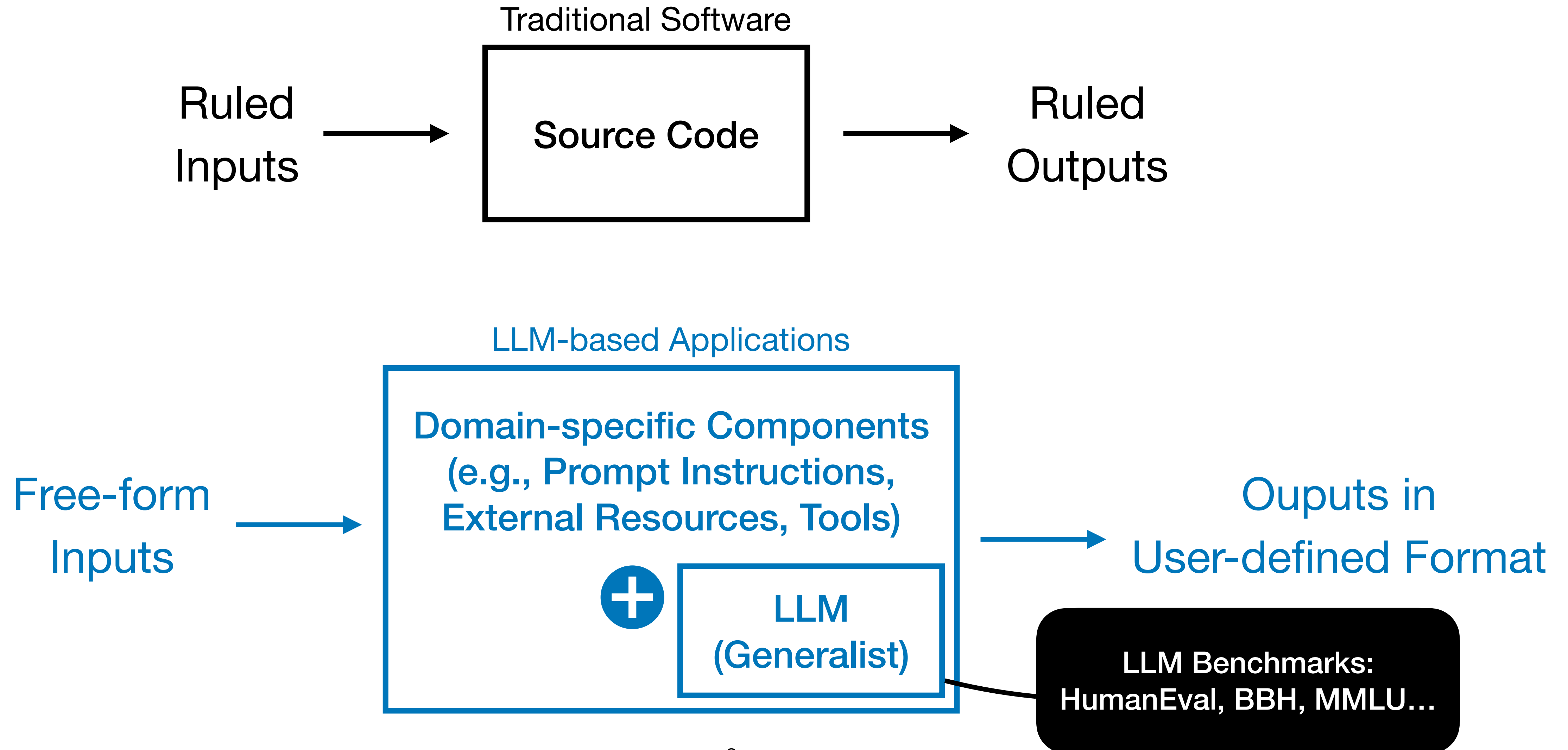
LLM의 효율적인 테스트를 위한, 도메인 특화 입력의 생성 전 적합도 평가

20260201 ERC 동계 워크숍

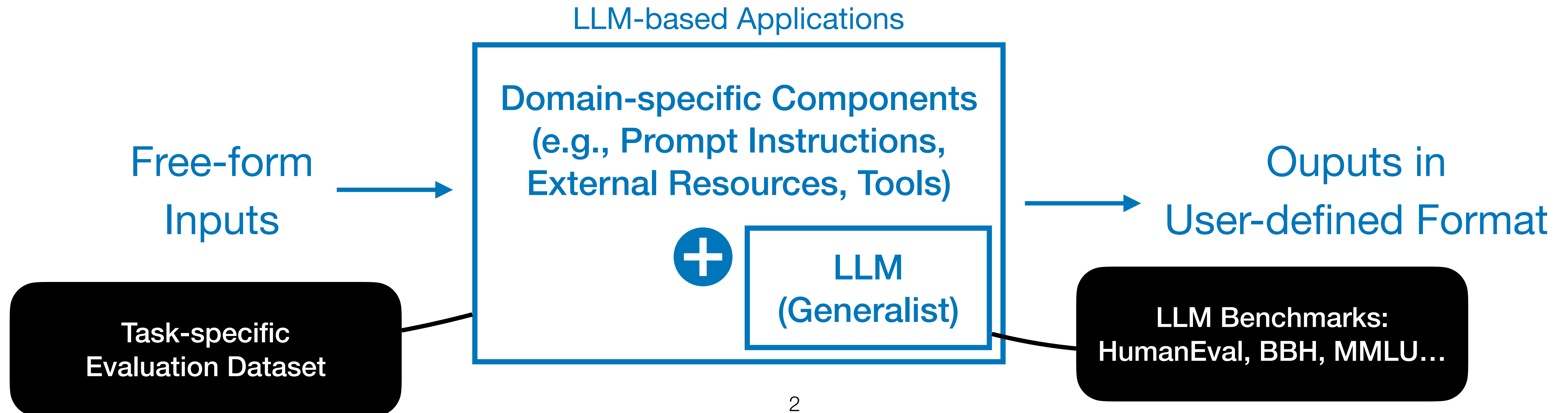
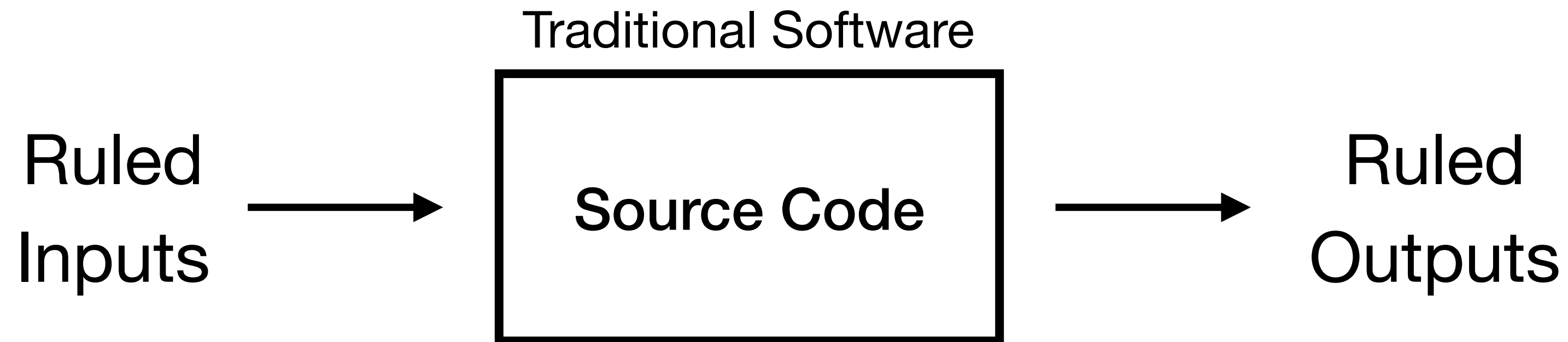
LLM-based Applications & Task-specific Evaluation



LLM-based Applications & Task-specific Evaluation



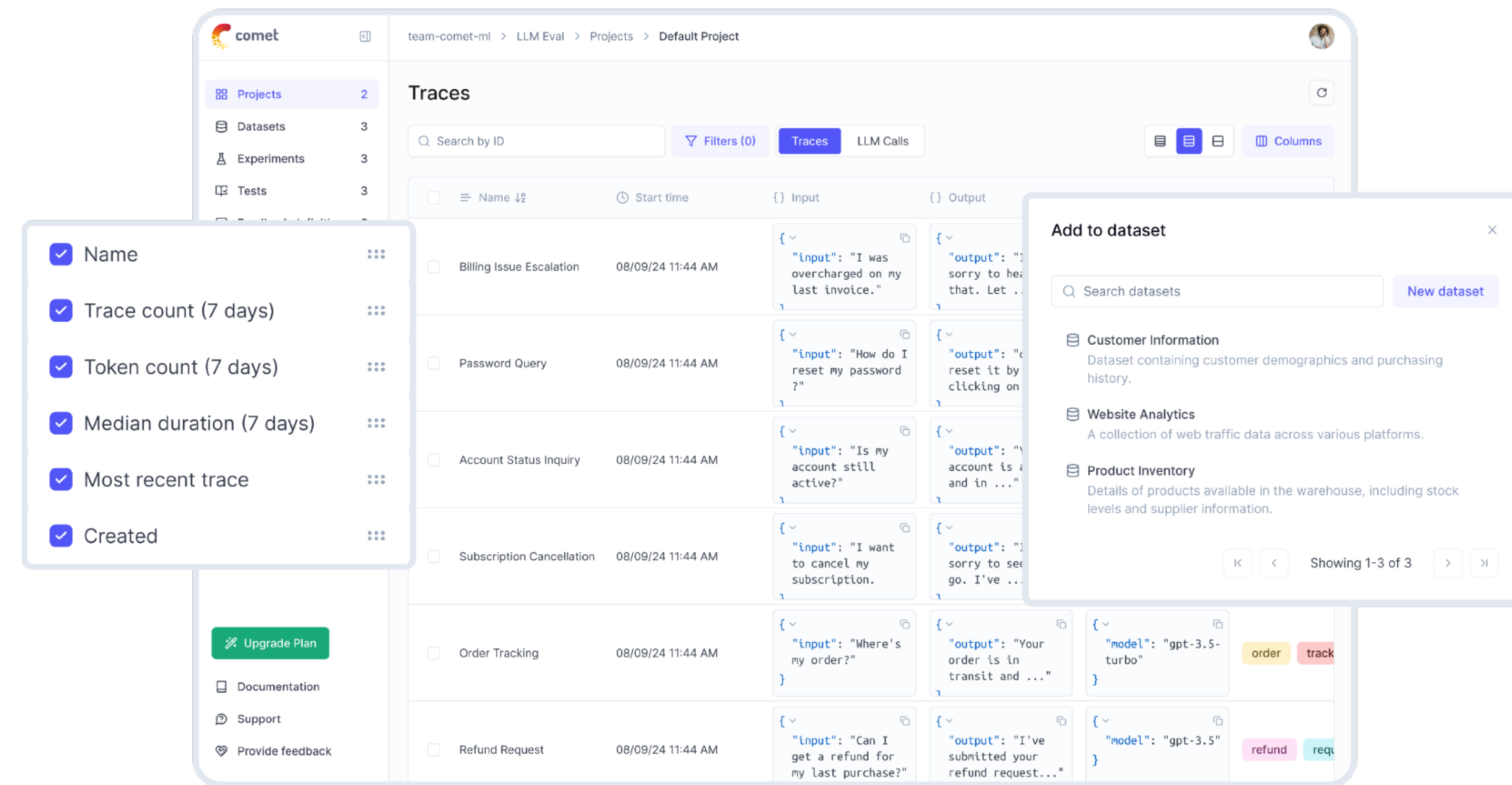
LLM-based Applications & Task-specific Evaluation



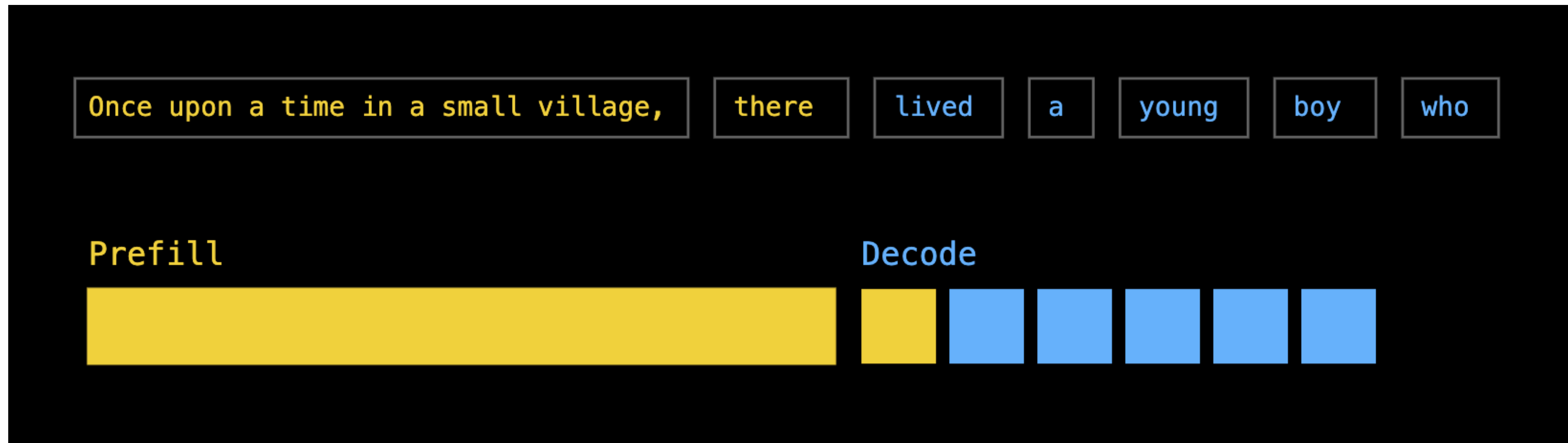
Pain Points?

LLM 기반 어플리케이션의 검증/평가를 위한 오픈소스 툴 탐색

- **SUT**
 - Single prompt Template
 - Basic RAG/Tool Usage
- 기본적으로 **Black-box** 검증만 지원
- **Cost Efficiency**
 - LLM Execution + Oracle Labelling
 - 생성한 이메일이 사용자의 요구 조건을 모두 포함하는지?
 - 잘못된 정보가 들어있지는 않은지?
 - 심지어 Stochasticity 때문에 여러 출력을 고려
 - 중요한 데이터부터 먼저 - Prioritisation/Selection



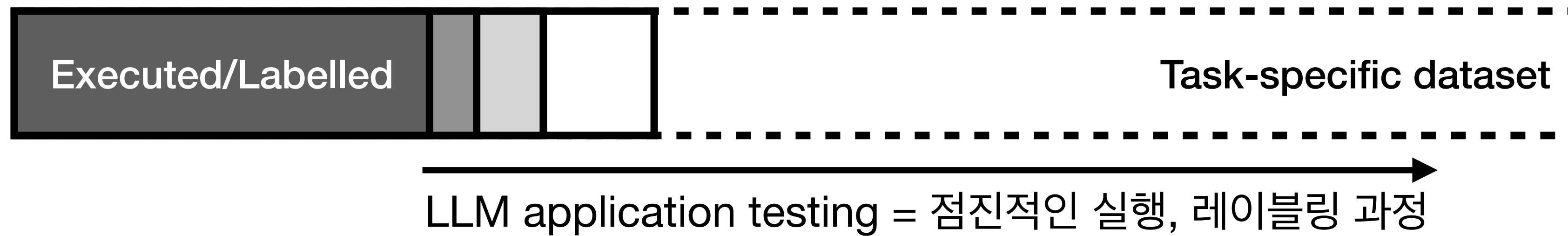
Decomposition of Testing Costs



<https://blog.exolabs.net/nvidia-dgx-spark/>

- LLM의 실행
 - Prefill + **Decode** (출력의 길이가 길수록 ⬆)
- LLM이 생성한 출력의 **레이블링** (정확성 판별)
 - 보통 길이가 긴 답변 생성, 다양한 기준 적용 필요, 한 입력에 대해 여러 출력 확인

Adaptive Testing



- 높은 실행/오라클 판별 비용 → Test Selection/Prioritisation이 중요
- 다음에 실행할 테스트 입력을 전략적으로 고르기
 - 지금까지랑 비슷하지 않은 것
 - 어려운 것!

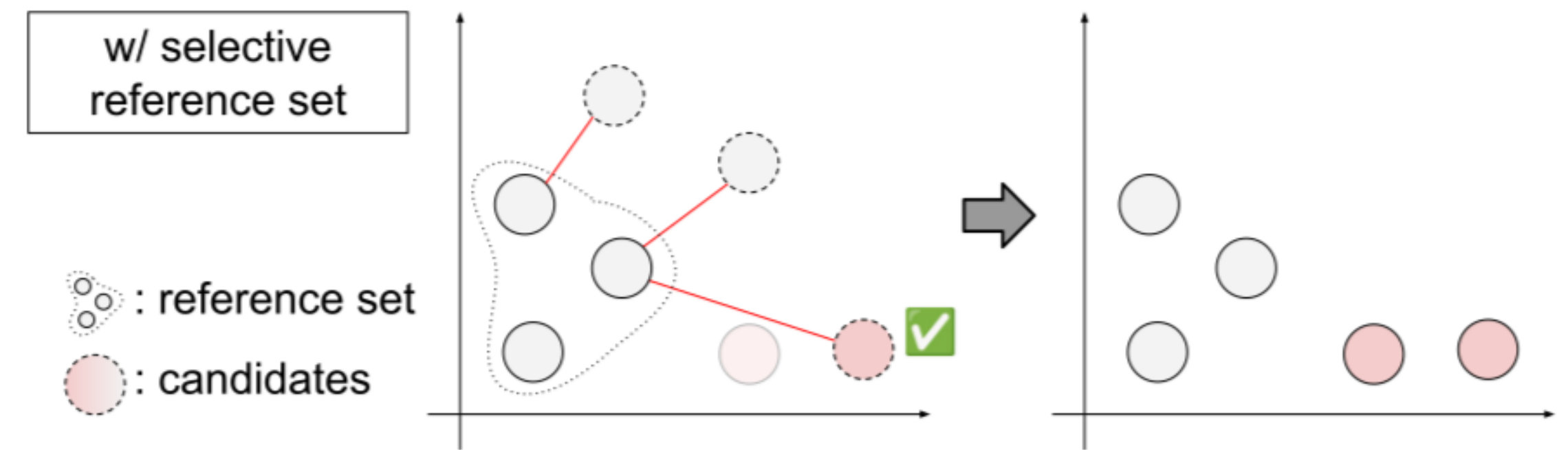
Adaptive Testing

- **External Representation**

- 외부 모델을 활용한 입력 임베딩 (ex. sBERT)
- 이미 확인한 입력과 겹치지 않도록, 다양하게

- **Internal Representation**

- 쉬운 입력 vs. 어려운 입력의 구분
- 트랜스포머의 Activation Vectors = 현재 입력에 대한 모델 내부 상태 → “인식” 반영



도메인 특화 공간에서 입력의 난도 예측

Prompt (prompt template + input)

Identify the topic of the following text:

Input

LLM

Generated tokens

token 1

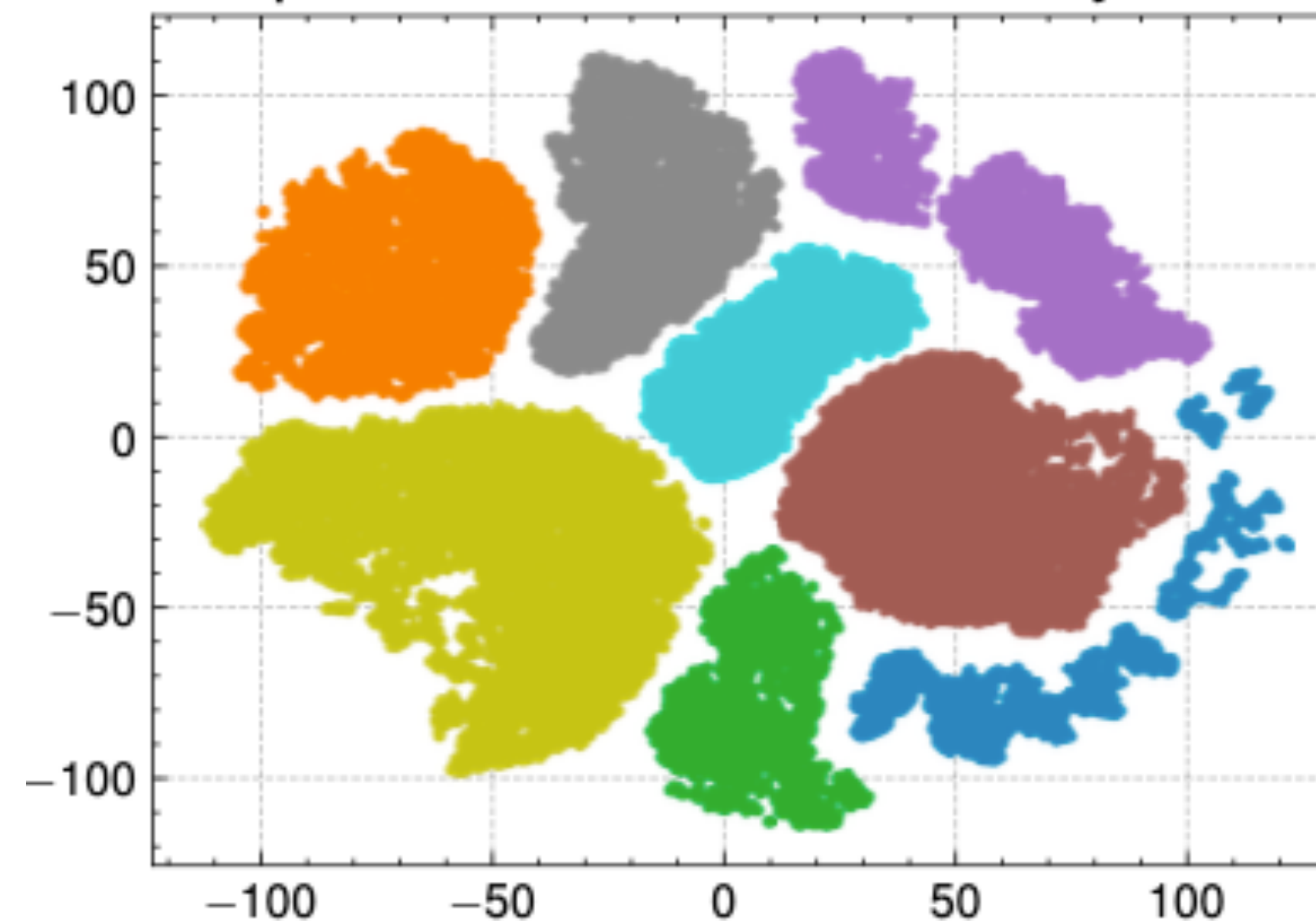
token 2

token 3

token 4 ...

Hidden states (Last layer, Last prompt token)

Input Hidden States Colored by Task



LIHS: Last Input Hidden States

도메인 특화 공간에서 입력의 난도 예측

Prompt (prompt template + input)

Identify the topic of the following text:

Input

LLM

Generated tokens

token 1

token 2

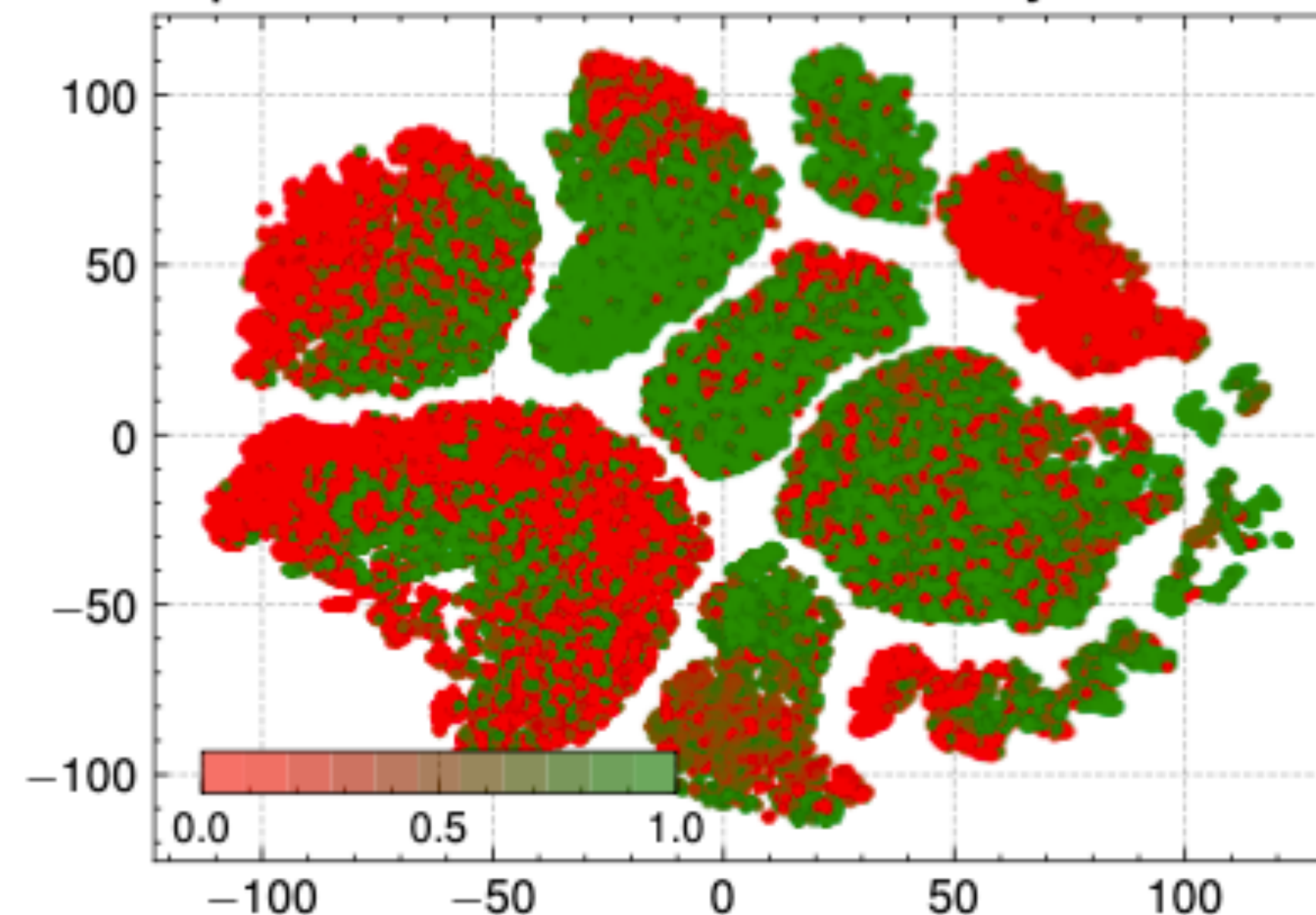
token 3

token 4

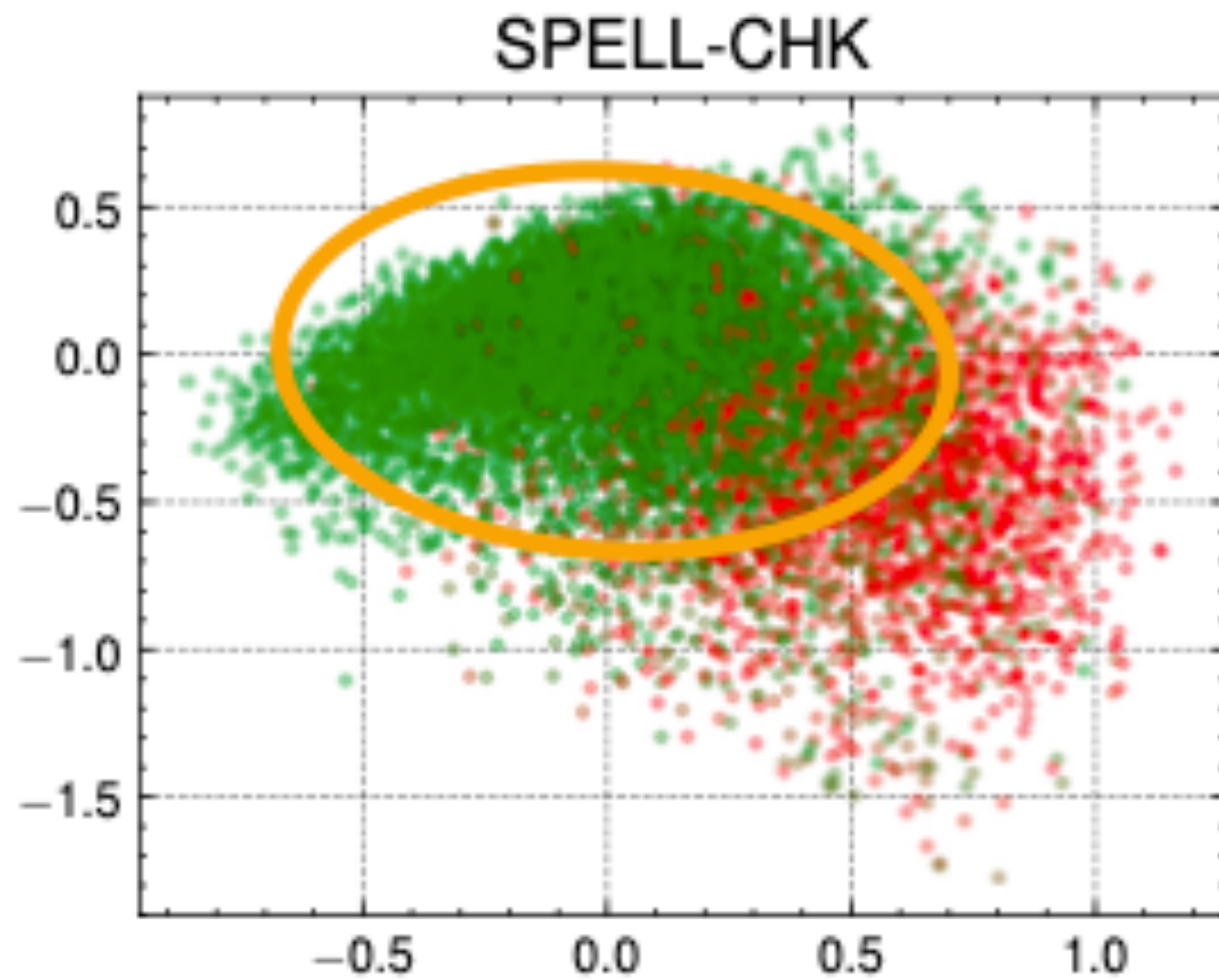
...

Hidden states (Last layer, Last prompt token)

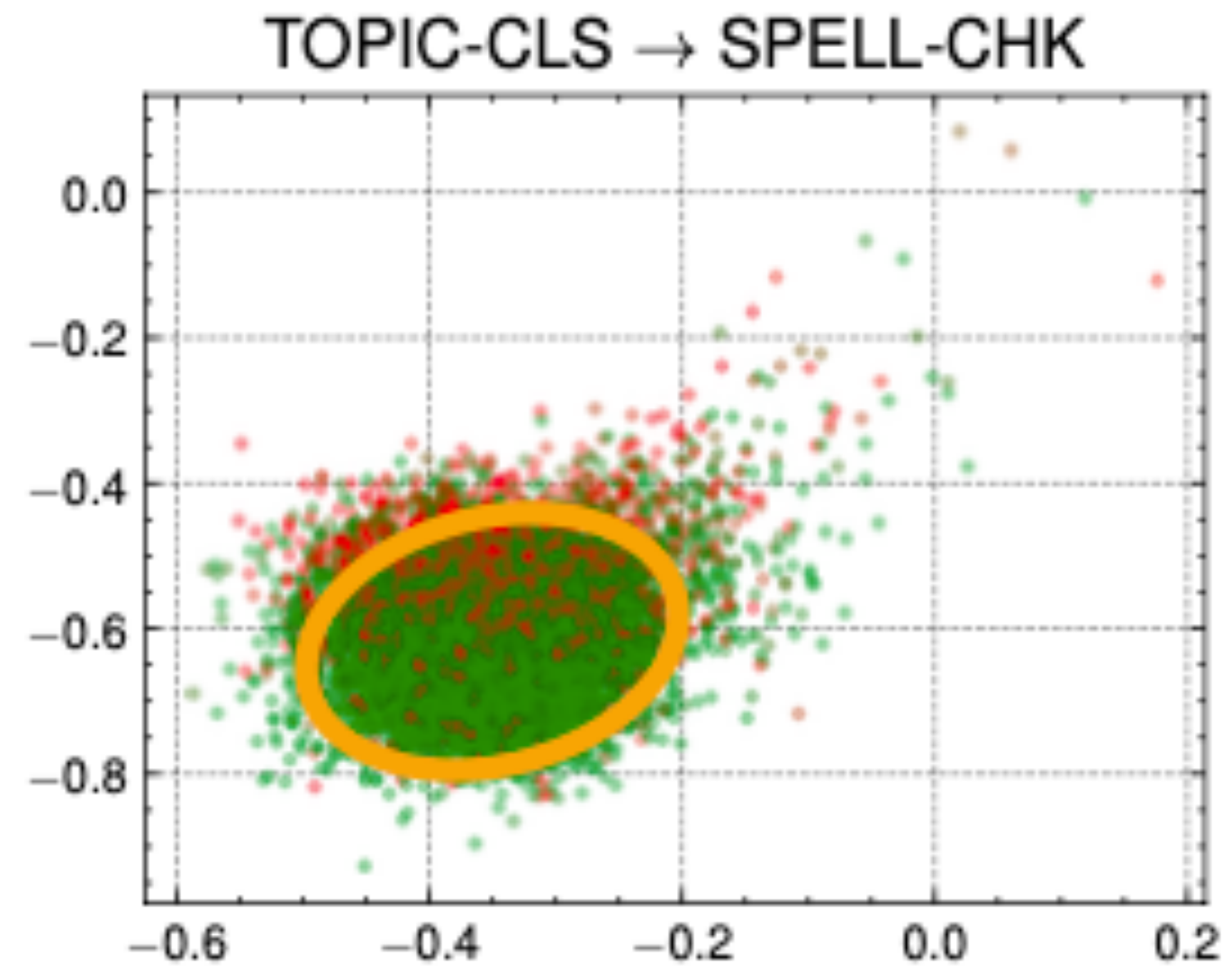
Input Hidden States Colored by Pass/Fail



도메인 특화 공간에서 입력의 난도 예측



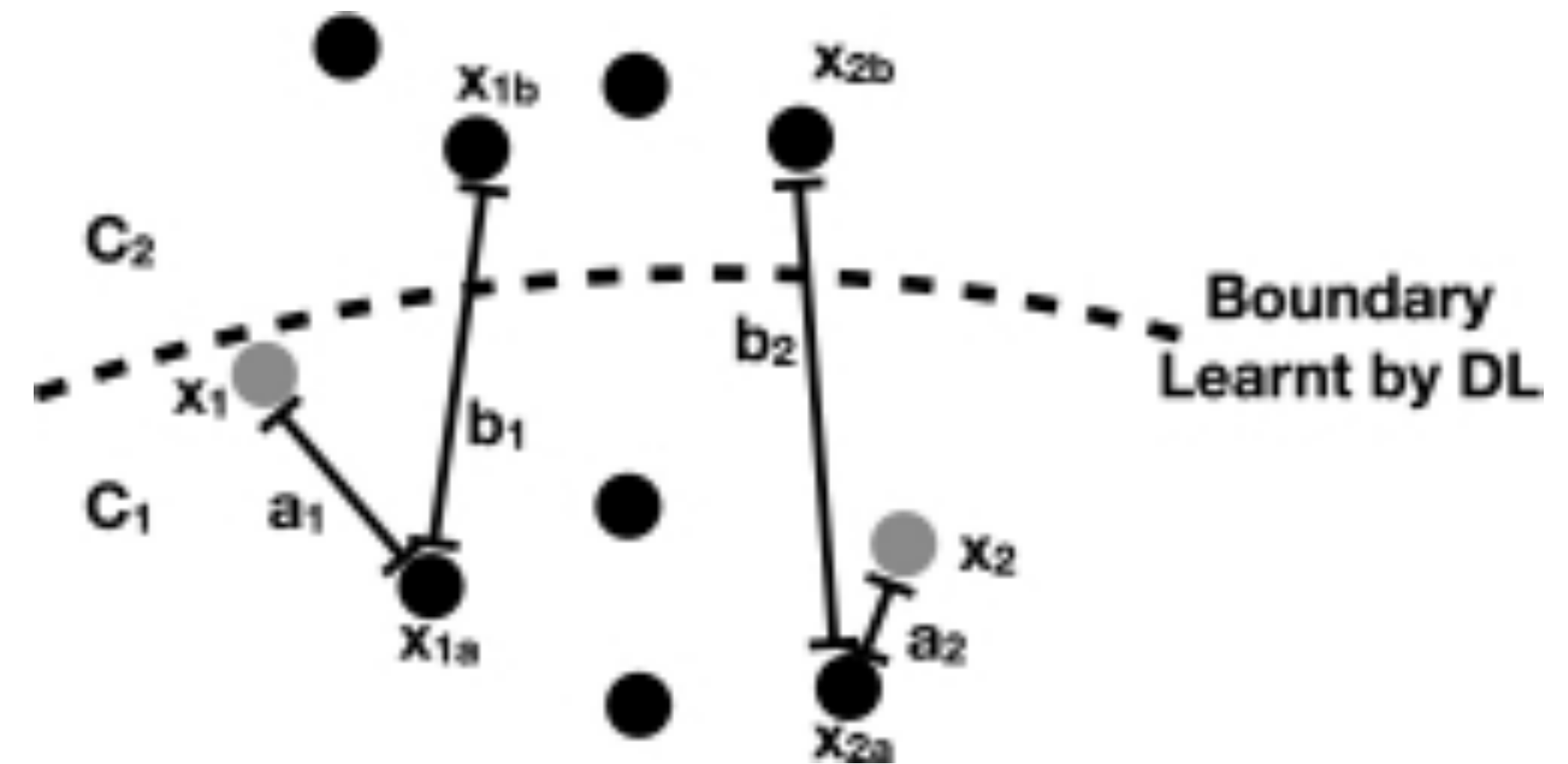
(a) Within-Task



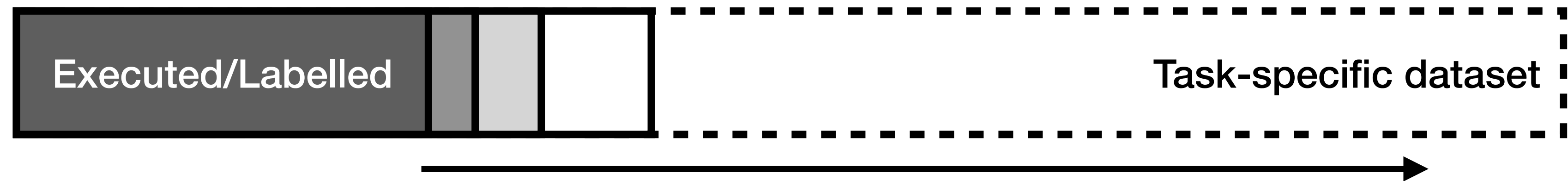
(b) Cross-Task

Test Adequacy as Distributional Deviation

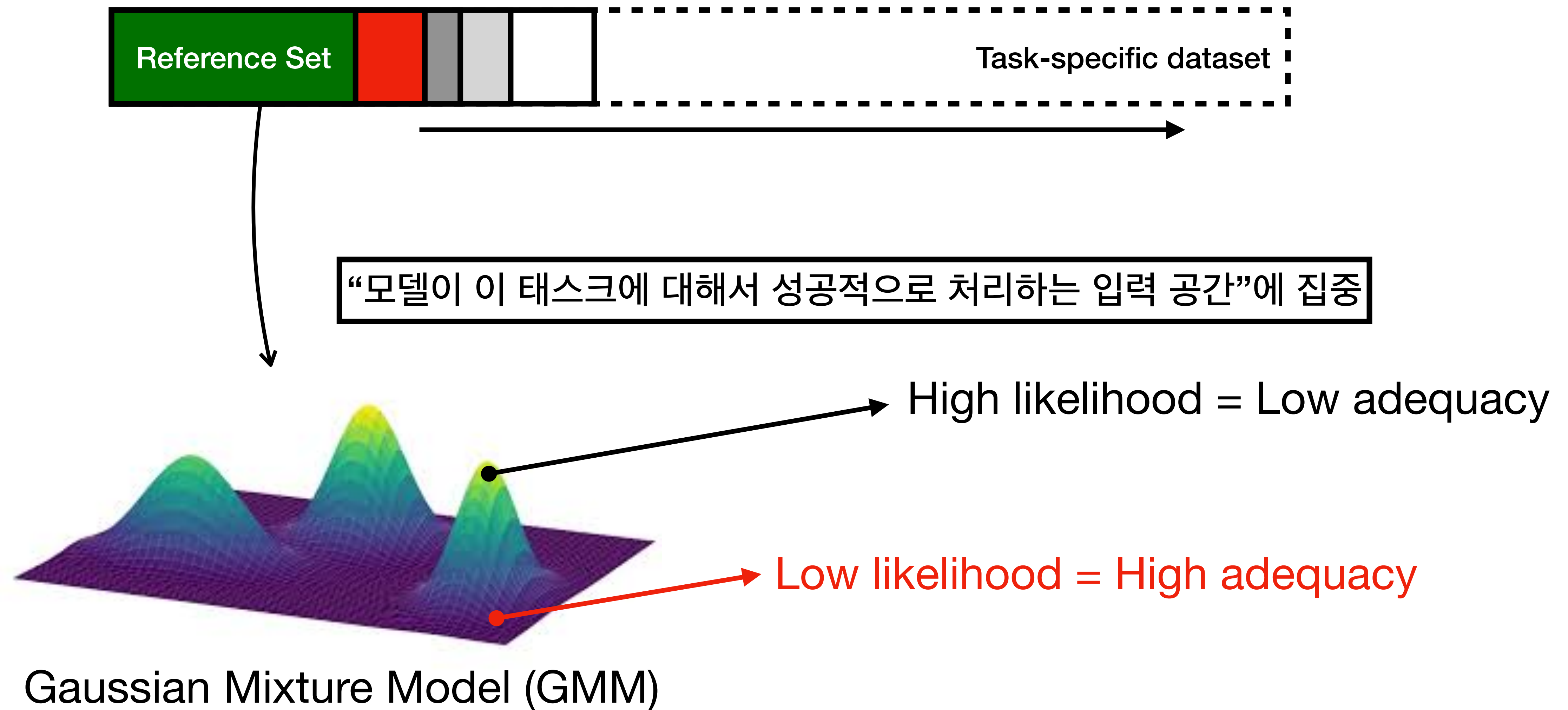
- Surprise Adequacy (2018, Kim et al.)
- 더 의미있는 입력 = “비전형적인” 입력
- 학습 데이터 분포에서 “더 많이 벗어난” 입력을 중요하게 고려
- Challenge: LLM의 거대한 학습 데이터를 단일 분포로 모델링..?
잘 학습된 모델이라면 **쉽게** 정답을 맞출 수 있는 입력들



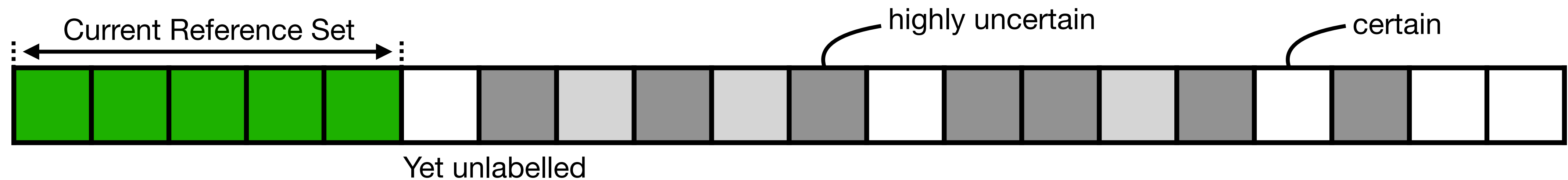
Adaptive Modelling of Reference Set



Adaptive Modelling of Reference Set

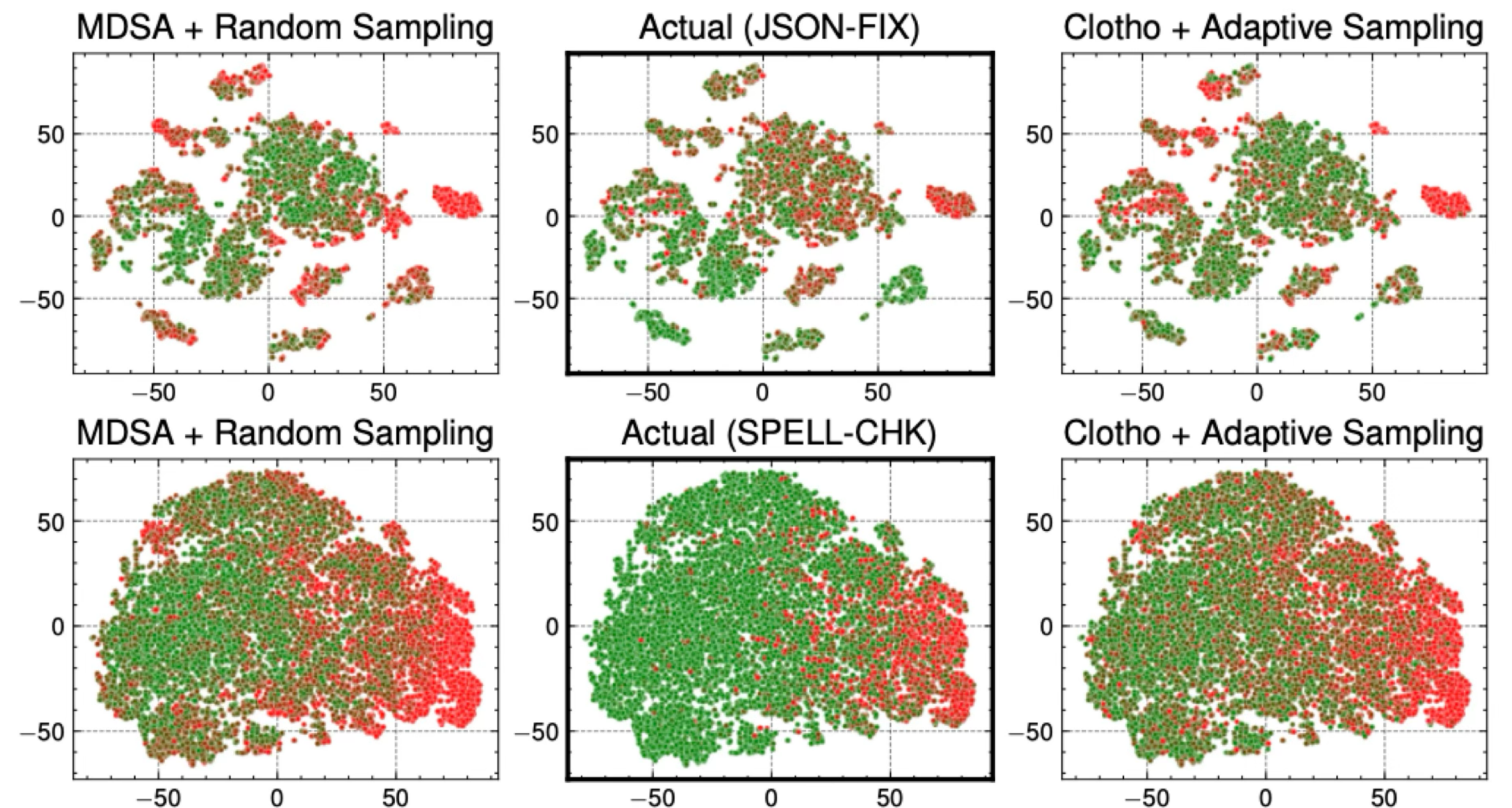


Reference Set Expansion



- **Active Learning**

- 모델이 레이블할 학습 데이터를 특정한 전략에 의거하여 선택
- Query Strategy: 모델에 추가로 학습시켰을 때 더 높은 정보량을 도입할 수 있는 샘플 선택



Reference Set Expansion

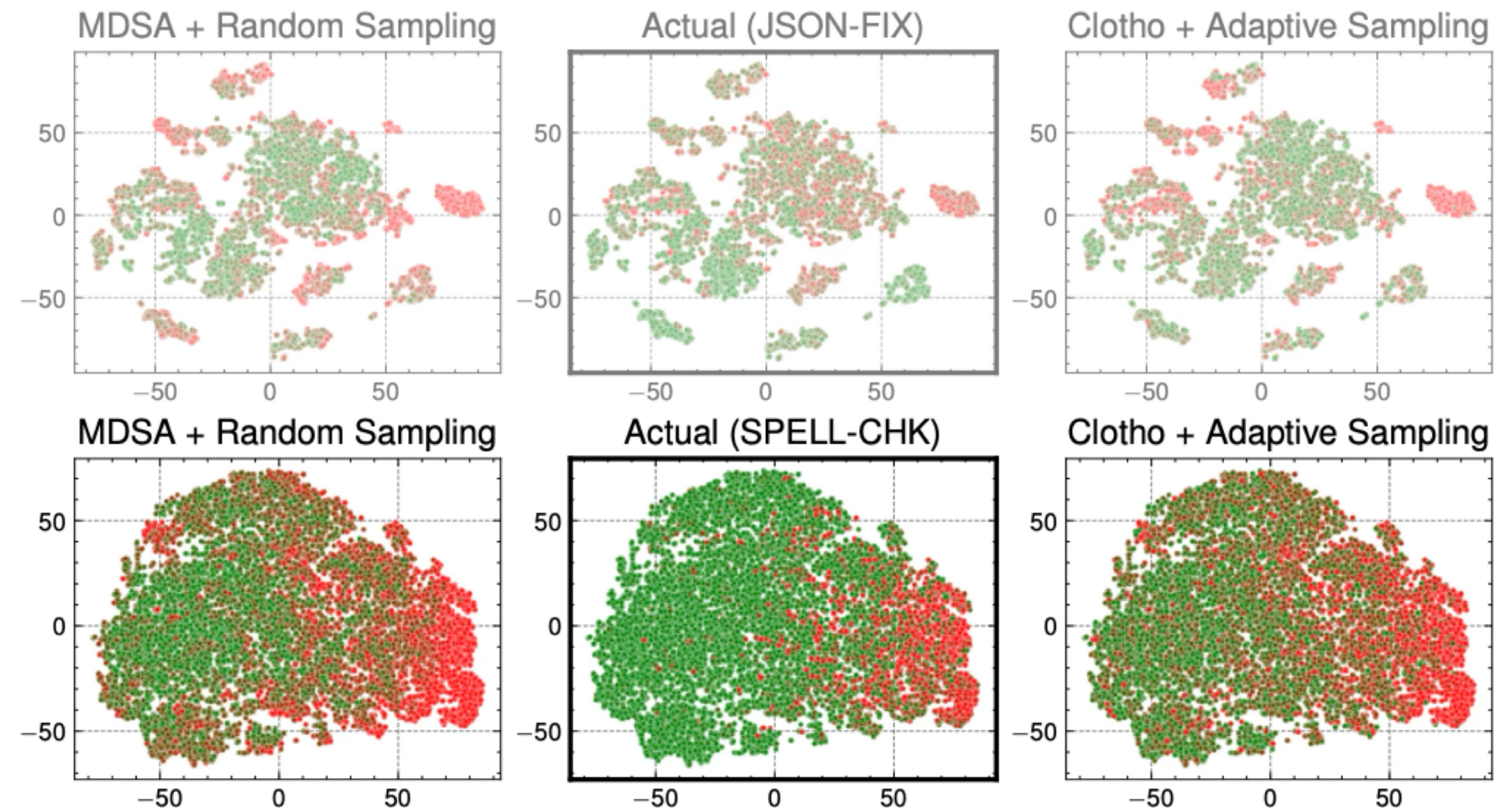
- **Exploration (Diversity)**
 - 기존의 reference set에 대해 minimum euclidean distance 기준 가장 먼 샘플
 - 입력 공간을 더 넓게 탐색하도록 유도
- **Exploitation (Entropy)**
 - GMM의 component assignment probability entropy
 - 현재 Mixture Component의 바운더리를 탐색함으로써 Passing Input Space를 더 정교하게 모델링

실험 환경

Task ID	# Inputs	Template	Input	Aim
ODD-ADD	6,000	P.E. Guide [3]	Randomly Sampled Numbers	To find sum of only odd numbers
GH-TYPO	10,000	Ours	GitHub Typo Corpus [19]	To fix typos extracted from GitHub commits
JSON-FIX	6,563	Ours	Synthetic JSON + Bugs [10]	To repair invalid JSON
MODEL-EX	9,810	P.E. Guide [3]	ML-ArXiv [9], Synthetic Abstracts	To extract ML model names from paper abstracts
POS-TAG	15,359	PromptPex [45, 46]	UD_English-EWT [47]	To detect Part-of-Speech tags
SPELL-CHK	10,000	Ours	WordNet sentences [16] + Misspellings [39]	To fix misspelt words
SYN-BUG	20,518	BICS-Dataset [34]	Python Syntactic Bug [34]	To find code lines with syntactic bugs
TOPIC-CLS	7,600	PromptPex [46]	AG News [56]	To classify topics of news articles

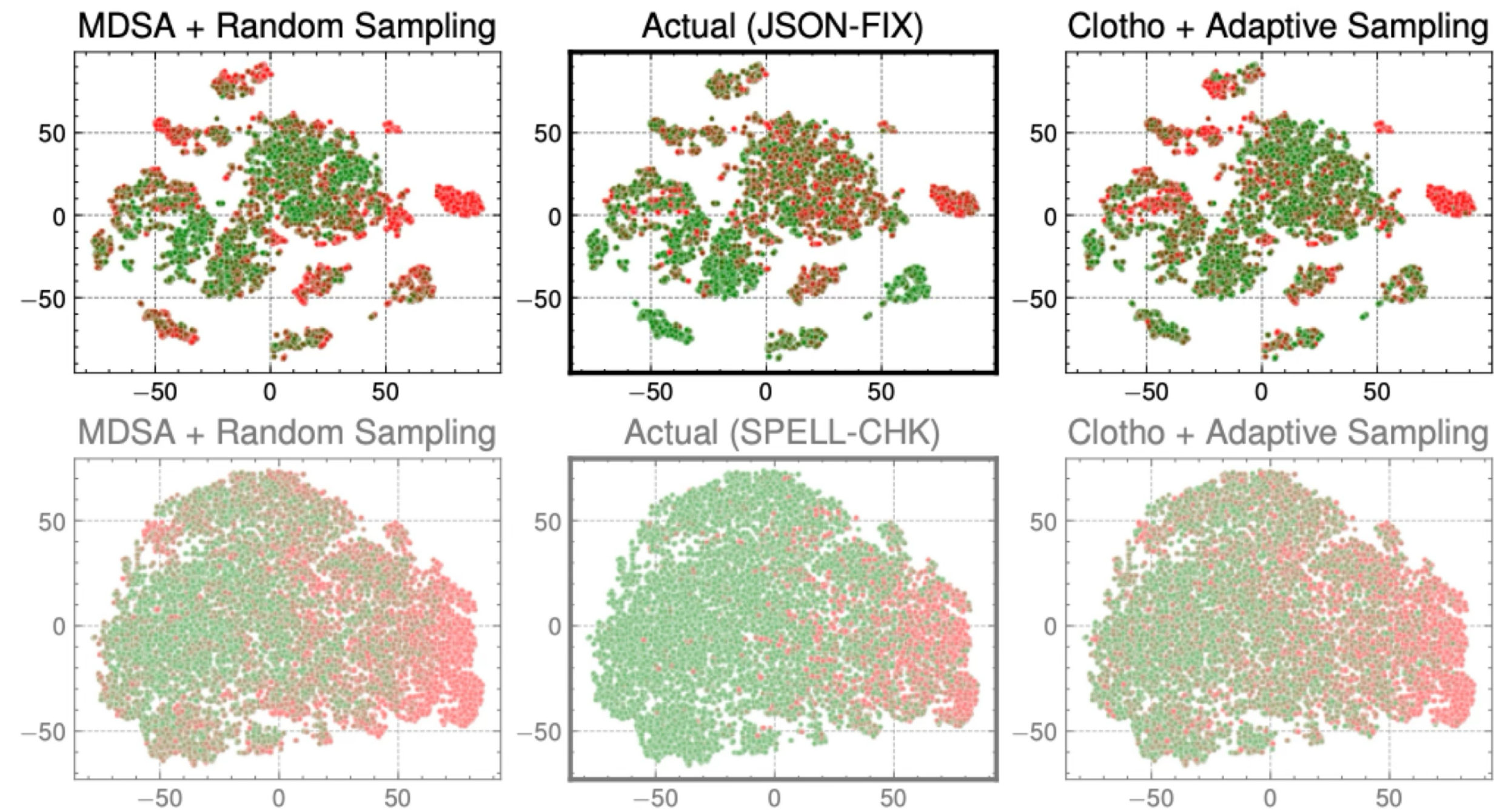
- Studied LLMs: Gemma-2 (9B), Llama-3.1 (8B), Mistral (7B)

Failure Prediction



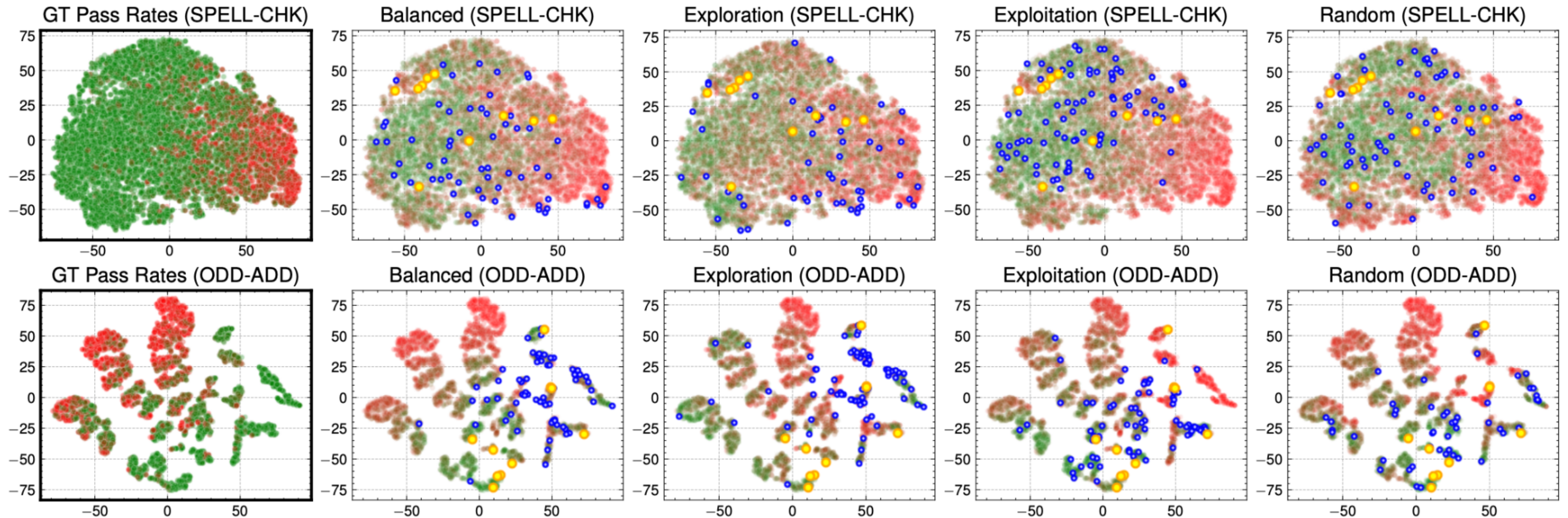
- *Multi-modal* Distribution: GMM이 여러 입력 공간 클러스터를 더 잘 포착
- *Uni-modal* distribution: MDSA와 같은 단순 거리 기반 deviation도 어느 정도 잘 작동

Failure Prediction



- *Multi-modal* Distribution: GMM이 여러 입력 공간 클러스터를 더 잘 포착
- *Uni-modal* distribution: MDSA와 같은 단순 거리 기반 deviation도 어느 정도 잘 작동

Expansion Strategies



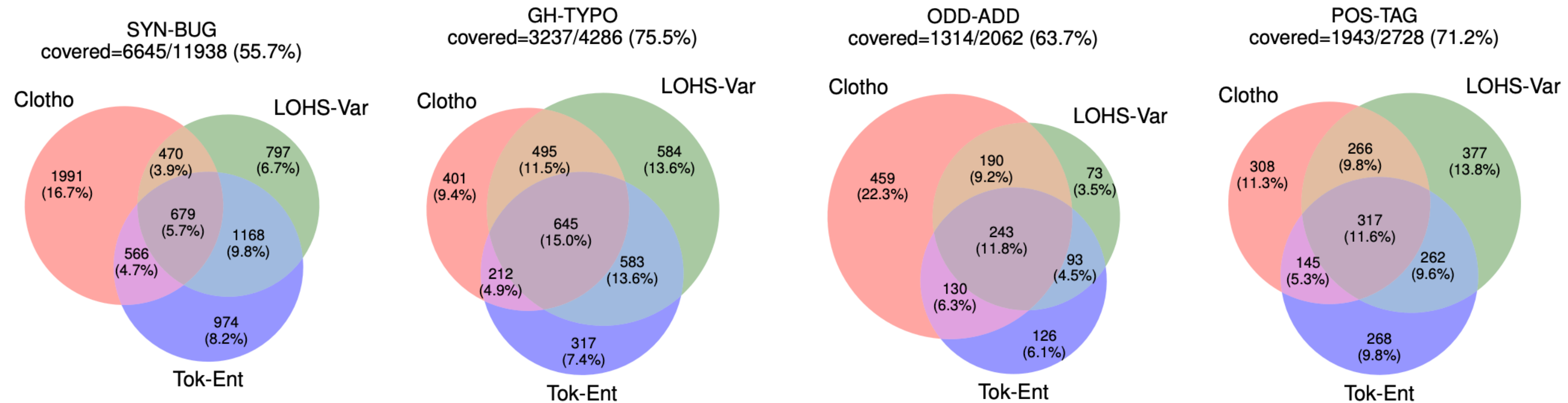
- **Yellow:** initial reference set, **Blue:** expanded samples
- **Exploration:** 더 넓은 입력 공간 커버, **Exploitation:** 성공하는 입력 공간을 더 조밀하게

실제 테스트팅 과정에서의 비용 절약: 테스트 우선순위화

- 500개 입력을 Balanced Strategy로 레이블한 상황 가정 (전체 입력 데이터의 ~5.4%)
- **Failure@N**
 - 높은 적합도의 Top N개 Input을 실제로 레이블했을 때, 그 중 실제로 failure가 발생한 횟수
 - top-100 적합도 순위 중 80.4%가 실제로 failure-inducing
 - ~30% for random selection
 - ~73% for MDSA

Gemma								
		CLOTHO		MDSA		Random		
Task ID	N	μ	σ	μ	σ	μ	σ	
ODD-ADD	100	100.0	0.00	100.0	0.00	40.7	4.06	9
	300	299.5	0.71	300.0	0.00	124.7	9.08	28
	500	455.4	13.31	477.3	6.72	200.8	9.27	45
GH-TYPO	100	86.7	5.01	77.5	7.75	34.6	3.86	9
	300	247.6	17.15	246.3	15.37	106.8	5.18	27
	500	402.2	33.27	408.3	23.98	179.9	6.06	45
JSON-FIX	100	76.7	7.09	67.5	9.03	16.2	3.77	9
	300	220.7	24.33	133.9	7.99	51.4	5.76	23
	500	326.3	39.50	195.0	13.68	85.9	4.07	31
MODEL-EX	100	45.3	17.64	27.0	2.45	33.9	3.84	6
	300	143.2	41.57	88.0	7.21	101.6	7.18	20
	500	240.0	65.77	162.9	13.80	167.1	10.65	33
POS-TAG	100	76.0	13.60	78.2	1.75	22.3	3.80	6
	300	199.6	39.97	190.2	12.17	59.1	4.12	15
	500	282.6	53.80	283.4	19.02	99.7	5.76	23
SPELL-CHK	100	86.9	2.60	82.0	0.82	12.3	4.32	7
	300	238.9	6.08	223.0	12.25	37.5	7.14	20
	500	362.2	14.73	331.8	19.56	61.8	7.04	31
SYN-BUG	100	94.4	3.17	96.2	2.86	58.9	4.68	8
	300	285.1	6.67	287.8	7.84	173.0	8.25	25
	500	477.7	8.60	480.5	9.78	286.6	8.66	42
TOPIC-CLS	100	52.0	10.41	59.9	3.51	15.5	3.31	5
	300	129.3	22.70	154.0	8.38	50.8	5.73	12
	500	193.6	29.28	217.2	15.52	88.4	11.21	18

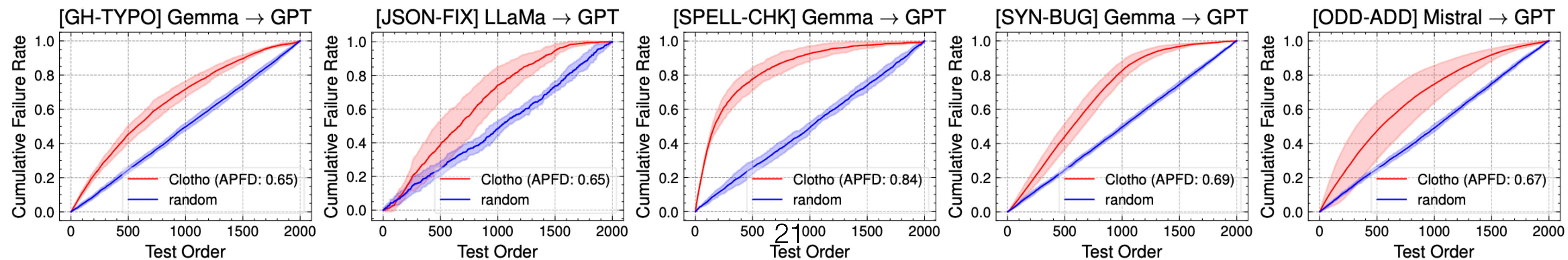
“생성 후” 적합도 평가 메트릭과의 상호보완성



- LOHS-Var: 한 입력에 대한 여러 생성 결과들 사이의 내부 상태 분산값
- Tok-Ent: 디코딩 과정에서 각 토큰 분포의 엔트로피 값 평균

오픈소스 LLM 활용 결과를 상용 모델로 전이

Task ID	Gemma				LLaMA				Mistral			
	SELF	Claude	GPT	Gemini	SELF	Claude	GPT	Gemini	SELF	Claude	GPT	Gemini
ODD-ADD	0.444	<u>0.447</u>	0.094	0.064	0.549	0.411	<u>0.439</u>	0.235	0.434	0.231	<u>0.437</u>	0.182
GH-TYPO	0.438	0.383	0.373	<u>0.442</u>	0.470	0.380	0.365	<u>0.413</u>	0.197	0.106	<u>0.140</u>	0.125
JSON-FIX	0.394	<u>0.128</u>	0.108	0.111	0.360	0.098	<u>0.167</u>	0.160	0.568	<u>0.203</u>	-0.010	0.006
MODEL-EX	0.386	0.248	0.347	<u>0.402</u>	0.512	0.248	0.341	<u>0.427</u>	0.555	0.288	0.380	<u>0.456</u>
POS-TAG	0.289	0.267	<u>0.272</u>	0.253	0.221	0.210	<u>0.235</u>	0.184	0.138	0.190	<u>0.233</u>	0.173
SPELL-CHK	0.452	<u>0.465</u>	0.443	0.419	0.374	<u>0.349</u>	0.340	0.321	0.247	<u>0.292</u>	0.263	0.239
SYN-BUG	0.707	0.435	<u>0.543</u>	0.532	0.345	0.190	0.178	<u>0.224</u>	0.396	0.325	<u>0.379</u>	0.349
TOPIC-CLS	0.252	0.215	<u>0.259</u>	0.229	0.256	0.188	0.201	<u>0.209</u>	0.181	0.139	0.160	<u>0.161</u>



요약

Preprint



Clotho: Measuring Task-Specific Pre-Generation Test Adequacy for LLM Inputs

- “생성 전” 적합도 예측
 - LLM 실행과 레이블링 비용을 모두 고려한 Test Prioritisation/Selection
 - 상용 모델로의 적용 가능성 제시
 - 더 알아봐야 할 것: 모델간의 특성, 오픈소스 모델 앙상블?
- “도메인 특화”
 - LLM 자체가 아닌 LLM의 실제 소프트웨어 응용에 집중

Testing for AI-based Software: 2030

- Bayesian Optimization (더 정확한 불확실성 예측)
- Different Failure Modes (Correct/Incorrect → Scores w.r.t. Different Properties)
- Beyond Single Prompt Template: Agentic AI systems
- Advanced Transformers (or Else?) Architecture (ex. Mixture-of-Experts)