LongBench Pro

2026-01-24

김기범

문제의식

- **평가 데이터 :** 기존 long-context 벤치마크들의 데이터셋은 1. 합성 데이터 기반으로 현실성이 부족하거나, 2. 수작업이라 길이·다양성·난이도 확장이 불가능했다.

- **평가 축:** long context 정확도만으로는 부족하다. 더 많은 정보량이 필요하다.

LongBench Pro: A More Realistic and Comprehensive Bilingual Long-Context Evaluation Benchmark

---

Ziyang Chen [1 2] **Xing Wu** [1] **Junlong Jia** [3] **Chaochen Gao** [1 2] **Qi Fu** [4] **Debing Zhang** [4] **Songlin Hu** [1 2]

2026년 1월 6일 Arxiv 업로드

1. Institute of Information Engineering, Chinese Academy of Sciences
2. School of Cyber Security, University of Chinese Academy of Sciences
3. School of Artificial Intelligence, Beihang University Xiaohongshu Inc.
4. Xiaohongshu Inc

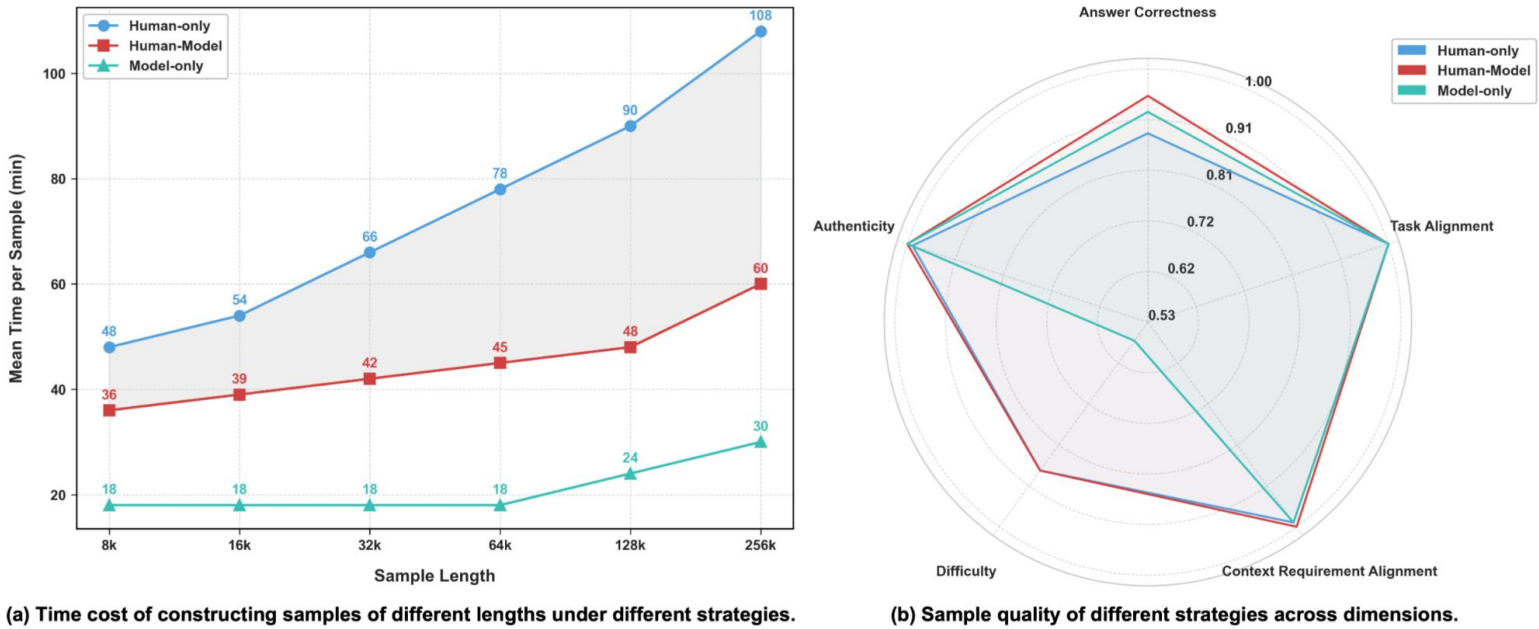
평가 데이터

1. 프런티어 LLM들(Gemini-2.5-Pro, GPT-5, Claude-4-Sonnet 등)이 긴 natural text에 대해 질문, 정답, 설계 의도, 풀이 과정을 먼저 생성.
   (i) **질문**, (ii) **정답**, (iii) **설계 근거 + 풀이 과정**(인간 평가자의 인지부하를 줄이기 위함.)

2. 인간 전문가는 정합성 검증, hallucination 제거, 난이도 보정, 최종 승인
   Task 정합성, 정답의 정확성 등.
   각 natural text에 대해 문제 후보 중 최선을 선택.

LLM-generated only 보다 정확하고, Human-only보다 저렴하다.

평가 데이터



(a) Time cost of constructing samples of different lengths under different strategies.

(b) Sample quality of different strategies across dimensions.

Figure 10. Comparison of sample construction strategies.

LLM-generated only 보다 정확하고, Human-only보다 저렴하다.

평가 축

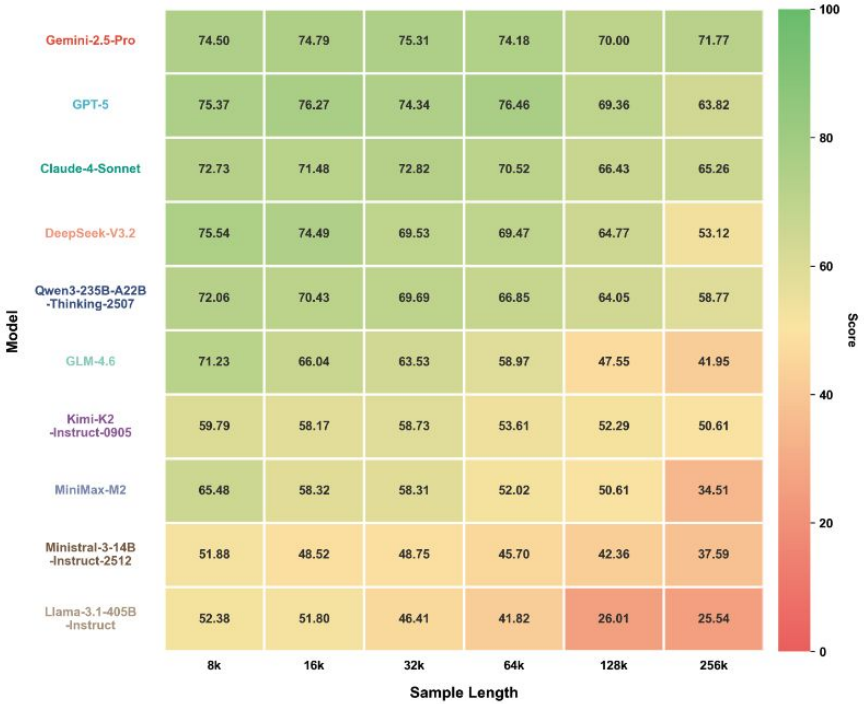
- **Context Requirement:**
  - Partial: 국소적 검색, 회수 중심(~needle in a haystack)
  - Full: 문서 전반에 흩어진 정보의 통합·추론 필요

- **Length**: 6단계 (8k / 16k / 32k / 64k / 128k / 256k)

- **Difficulty**: Easy / Moderate / Hard / Extreme, 사람의 주관이 아니라 모델 성능 분포로 정의.
  - Extreme: high-performing 티어에서 정답을 맞히는 모델이 최대 1개인 샘플
  - Hard: Extreme을 제외하고, mid-performing 티어에서 정답을 맞히는 모델이 최대 1개인 샘플
  - Moderate: Hard까지 제외하고, low-performing 티어에서 정답을 맞히는 모델이 최대 1개인 샘플
  - Easy: 나머지 전부

* 난이도 평가시 '정답'은 평가자들의 exact match, 실제 benchmark시에는 부분점수 있음(e.g. 큰 순서대로 보여줘)

결과



Model	Model Type	Context Length	Overall		Language				Difficulty							
					English		Chinese		Extreme		Hard		Moderate		Easy	
Gemini-2.5-Pro	Thinking	1M	-	73.42	-	72.35	-	74.49	-	50.77	-	81.03	-	81.98	-	84.40
Gemini-2.5-Flash	Mixed	1M	55.92	67.41	55.29	67.22	56.54	67.59	44.26	47.39	57.87	72.19	53.99	72.39	66.55	79.82
Gemma-3-27B-It	Instruct	128k	36.14	37.34	37.46	40.89	34.81	33.78	30.22	27.78	33.20	30.56	25.04	24.53	49.96	57.81
Gemma-3-12B-It	Instruct	128k	32.16	31.92	33.03	34.43	31.28	29.41	26.44	25.74	30.43	28.02	23.39	22.61	43.66	45.48
Gemma-3-4B-It	Instruct	128k	21.76	21.20	22.63	23.28	20.89	19.12	19.31	18.72	20.70	19.87	15.82	13.85	28.18	28.66
GPT-5	Thinking	272k	-	72.61	-	73.24	-	71.97	-	48.37	-	78.74	-	82.31	-	85.23
GPT-4o	Instruct	128k	46.67	49.44	47.67	52.61	45.66	46.26	36.30	34.39	44.88	41.35	43.03	43.07	59.38	71.84
GPT-OSS-120B	Thinking	128k	-	52.61	-	54.67	-	50.54	-	35.4	-	44.97	-	50.66	-	74.06
GPT-OSS-20B	Thinking	128k	-	44.66	-	47.83	-	41.49	-	31.59	-	35.89	-	39.33	-	65.05
Claude-4-Sonnet	Mixed	1M	56.07	69.87	57.14	71.09	54.99	68.65	42.92	47.05	57.57	74.72	53.96	76.58	68.42	83.78

의문

- 모델 평가와 데이터 라벨링(난이도)이 뒤섞여 있음. 평가 대상 모델의 성능이 문제의 난이도를 결정. 최신 모델들(gemini 3 pro 등)로 검증하면 각 문제의 난이도가 바뀔 수 있음. 즉, 난이도 label이 시간에 대한 일관성이 없는 것 같다.(이건 elo score도 마찬가지)
  논문의 주장: 난이도를 human 주관 대신 model-centric으로 정의하는 방식이 벤치마크와 모델 능력의 co-evolution을 위한 더 자연스러운 기반이라 주장. (???)

- 문제 생성 모델과 검증 모델이 같으면 cheating임. 실제로 성능이 올라감.
  다양한 모델이 문제를 생성하여 인간-기반 최종 수정&선택이니 괜찮지 않을까?(논문에 따로 언급은 없음)

다른 생각

1.

Benchmark상으로 드러난 특정 LLM의 ‘effective’ context length 정보가 있다고 가정.

Multi-agent system에서 orchestration 역할을 수행하는 agent가 실제 task를 수행하는 agent의 context 사이즈를 재고, 만약 effective max token length를 초과하는 경우 task를 쪼개거나 추가적인 context 압축을 수행하도록 지시할 수 있지 않을까? 다시말해 agent가 부여하는 task의 context 복잡성을 ‘감'이 아니라 effective context length 기준으로 판단할 수 있지 않을까?

2.

중국발 eval 논문들은 중국어/영어 데이터셋인 경우가 많다. 간접적으로 LLM 모델들로 하여금 중국어 성능을 높이도록 압박하는 셈.