비양자화(Non-Quantized) 대규모 언어 모델의 서버 환경별 실행 가능성 및 성능 비교 연구

최유빈○, 유한솔, 하준서, 김주성, 윤태복
서일대학교 AI게임융합학과
e-mail:tbyoon@seoil.ac.kr

A Comparative Study on the Feasibility and Performance of Non-Quantized Large Language Models Across Server Environments

Youbin Choi○, Hansol Yu, Junseo Ha, Jusung Kim and Taebok Yoon
*Dept. of AI Game Convergence, Seoil University

요 약

대규모 언어 모델의 성능 평가 수요가 증가함에 따라 다양한 서버 환경에서의 실행 가능성 검토가 중요한 과제로 부상하고 있다. 본 연구는 비양자화 BF16 조건에서 대규모 언어 모델의 서버 하드웨어별 실행 가능성과 성능 차이를 비교·분석하였다. 이를 위해 세 가지 서버 환경(Srv-A, Srv-B, Srv-C)에서 9개의 기반 밀집 언어 모델을 대상으로 총 27개의 실험 작업을 수행하였으며, 언어 모델 평가 프레임워크와 추론 백엔드를 활용하여 5개의 벤치마크 과제를 동일한 조건 하에 실행하였다. 실험 결과, Srv-A는 30억 매개변수부터 720억 매개변수까지 모든 모델을 안정적으로 평가할 수 있었던 반면, Srv-B와 Srv-C는 120억 매개변수 모델까지만 평가가 가능하였으며 320억 매개변수 이상의 모델에서는 로드 시간 초과 또는 메모리 부족 오류가 발생하였다. 이를 통해 비양자화 BF16 조건에서 RTX 4090 기반 서버의 실용적 평가 가능 범위는 약 120억 매개변수 수준임을 확인하였으며, 서버 하드웨어 사양이 대규모 언어 모델의 실행 가능성과 평가 성능에 직접적인 영향을 미침을 실증적으로 보여주었다.

1. 서론

최근 공개 대규모 언어 모델의 규모가 빠르게 증가하면서, 벤치마크 성능뿐만 아니라 실제 평가 환경에서 모델을 안정적으로 실행할 수 있는지가 중요한 문제로 부각되고 있다. 대형 모델은 높은 성능을 기대할 수 있지만, 방대한 GPU 메모리와 긴 실행 시간을 요구하기 때문에 특정 하드웨어 구성에서 어느 규모의 모델까지 안정적으로 평가할 수 있는지를 사전에 파악하는 것이 필요하다.

본 연구는 비양자화 BF16 조건에서 서버 하드웨어 구성에 따른 공개 대규모 언어 모델의 실행 가능성과 평가 결과를 비교·분석한다. 이를 위해 H200 4장을 탑재한 Srv-A, RTX 4090 3장을 탑재한 Srv-B, RTX 4090 2장을 탑재한 Srv-C의 세 가지 서버 환경에서 9개의 기반 밀집 언어 모델을 동일한 벤치마크 과제, 제한 시간 조건, 평가 프레임워크로 실행하였다. 평가 지표로는 벤치마크 점수와 함께 실행 성공 여부, 로드 시간 초과, 메모리 부족, 실행 시간을 활용하였다. 이를 통해 RTX 4090 기반 서버의 실용적 평가 가능 범위와 H200 기반 서버에서의 대형 모델 평가 가능성을 실증적으로 확인하고자 한다.

2. 관련 연구

공개 대규모 언어 모델은 LLaMA, Mistral, Gemma, Qwen 등 다양한 계열로 확장되고 있으며, 이에 따라 모델의 성능을 정량적으로 비교하려는 연구가 활발히 진행되고 있다. 기존 연구에서는 공개 LLM을 대상으로 문맥 내 학습, 사고의 연쇄, 검색 증강 생성 등의 기법을 적용하거나, 한국어 및 도메인 특화 benchmark를 활용하여 모델별 응답 정확성, 언어 이해 능력, 활용 가능성을 분석하였다[1][2][3]. 이러한 연구들은 공개 LLM의 품질을 비교하고 활용 가능성을 제시했다는 점에서 의의가 있다.

LLM 평가의 재현성을 높이기 위해 lm-evaluation-harness와 같은 평가 프레임워크도 널리 활용되고 있다[4]. 또한 대규모 모델 실행 과정에서는 GPU 메모리와 KV cache 관리가 중요한 병목으로 작용하며, 이를 완화하기 위해 PagedAttention 기반의 vLLM과 같은 추론 backend가 제안되었다[5]. 그러나 기존 연구들은 주로 benchmark score, 프롬프트 조건, 언어 또는 도메인별 성능에 초점을 두는 경우가 많으며, 동일한 모델을 서로 다른 서버 하드웨어에서 실행했을 때의 OOM, timeout, runtime과 같은 실행 가능성 지표를 함께 비교한 연구는 상대적으로 제한적이다.

이에 본 연구는 기존 LLM 평가 연구와 추론 backend 연구를 바탕으로, non-quantized BF16 조건에서 서버 하드웨어 구성에 따른 공개 LLM의 실행 가능성과 평가 결과를 함께 분석한다는 점에서 차별성을 가진다.

3. 본론

본 연구는 non-quantized BF16 조건에서 서버 하드웨어 구성에 따른 대규모 언어 모델의 실행 가능성과 평가 결과를 분석하였다. 세 개의 서버와 9개의 base dense LLM을 대상으로 총 27개의 평가 job을 수행하였으며, benchmark score뿐만 아니라 모델 로딩 성공 여부, timeout, CUDA out-of-memory 발생 여부를 함께 비교하였다.

3.1 실험 환경

실험에는 세 가지 서버 구성을 사용하였다. Srv-A는 NVIDIA H200 4장을 탑재한 고성능 GPU 서버이며, Srv-B와 Srv-C는 각각 RTX 4090 3장 및 2장을 탑재한 서버이다. 서버별 주요 하드웨어 구성은 [표 1]과 같다.

[표 1] 서버 하드웨어 스펙

Server ID	CPU	RAM	GPU	GPU Memory	OS
Srv-A	Intel Xeon Platinum 8558	1.0 TiB	NVIDIA H200 × 4	140 GB × 4	Ubuntu 24.04 LTS
Srv-B	Intel Xeon Silver 4310	125 GiB	NVIDIA GeForce RTX 4090 × 3	24 GB × 3	Ubuntu 22.04 LTS
Srv-C	AMD Ryzen Threadripper PRO 5995WX	62 GiB	NVIDIA GeForce RTX 4090 × 2	24 GB × 2	Ubuntu 22.04 LTS

평가 대상 모델은 파라미터 수에 따라 Small, Medium, Large 세 그룹으로 구분하였다. Small 그룹은 10B 미만, Medium 그룹은 10B 이상 70B 미만, Large 그룹은 70B 이상 모델로 정의하였다. 모든 모델은 양자화를 적용하지 않은 BF16 조건에서 평가하였다. 평가에는 lm-evaluation-harness와 vLLM backend를 사용하였으며, benchmark task는 mmlu, arc_challenge, hellaswag, truthfulqa_mc2, gsm8k의 다섯 가지로 구성하였다.

[표 2] 실험 평가 모델

Group	Model	Params
Small	meta-llama/Llama-3.2-3B	3B
Small	Qwen/Qwen3-4B-Base	4B
Small	google/gemma-2-9b	9B
Medium	mistralai/Mistral-Nemo-Base-2407	12B
Medium	Qwen/Qwen3-32B	32B
Medium	01-ai/Yi-1.5-34B	34B
Large	meta-llama/Llama-3.1-70B	70B
Large	Qwen/Qwen2.5-72B	72B
Large	swiss-ai/Apertus-70B-2509	70B

각 task는 num_fewshot=0, limit=50인 zero-shot 조건에서 실행하였고, 각 job에는 3600초의 timeout을 적용하였다. 모델의 최대 입력 길이는 max_model_len=2048로 제한하였으며, GPU memory utilization은 모델 크기에 따라 0.85 또는 0.90으로 설정하였다.

3.2 실행 결과 분류 기준

각 job의 실행 결과는 benchmark 결과 생성 여부와 실행 로그를 기준으로 분류하였다. benchmark 결과가 정상적으로 생성된 경우는 SUCCESS로 분류하였다. 모델 로딩 단계에서 3600초 제한 시간을 초과한 경우는 TIMEOUT_LOAD, CUDA 메모리 부족으로 모델 로딩에 실패한 경우는 OOM_LOAD로 분류하였다. 원시 결과에서 PARTIAL_OR_CHECK_REQUIRED로 기록된 job은 job_summary.json과 실행 로그를 확인하여 최종 상태를 재분류하였다.

3.3 서버별 실행 가능성 분석

전체 27개 job 중 17개 job이 정상적으로 bench mark 결과를 산출하였다. Srv-A는 9개 모델 전체에서 평가에 성공하여 100%의 실행 성공률을 보였다. 반면 Srv-B와 Srv-C는 각각 4개 모델에서만 평가에 성공하여 44.4%의 성공률을 보였다.

[표 3] 서버별 실행 결과

Server ID	SUCCESS	TIMEOUT_LOAD	OOM_LOAD	Total Jobs	Success Rate
Srv-A	9	0	0	9	100.0%
Srv-B	4	2	3	9	44.4%
Srv-C	4	0	5	9	44.4%
Total	17	2	8	27	63.0%

Srv-A는 Small, Medium, Large 그룹의 모든 모델을 제한 시간 내 평가할 수 있었다. 이는 H200의 대용량 GPU 메모리가 non-quantized BF16 조건의 대형 모델 로딩에 충분한 여유를 제공했기 때문으로 해석된다. 반면 RTX 4090 기반의 Srv-B와 Srv-C는 12B 모델까지는 평가가 가능했으나, 32B 이상 모델에서는 timeout 또는 OOM이 발생하였다.

특히 Srv-B에서는 32B 및 34B 모델이 모델 로딩 단계에서 제한 시간을 초과하였고, 70B급 모델은 모두 OOM으로 종료되었

다. Srv-C에서는 32B 이상 모델이 모두 OOM으로 종료되었다. 이를 통해 RTX 4090 기반 서버의 non-quantized BF16 평가 가능 범위는 본 실험 조건에서 약 12B 수준으로 관찰되었다.

3.4 모델 크기별 실행 가능성 분석

모델 크기 그룹별로 보면 Small 모델은 모든 서버에서 정상적으로 실행되었다. Medium 모델은 12B 모델까지는 모든 서버에서 평가가 가능했으나, 32B 및 34B 모델에서는 서버 구성에 따라 timeout 또는 OOM이 발생하였다. Large 모델은 Srv-A에서만 평가에 성공하였으며, Srv-B와 Srv-C에서는 모두 모델 로딩에 실패하였다.

[표 4] 모델 그룹별 실행 결과

Model Group	SUCCESS	TIMEOUT_LOAD	OOM_LOAD	Total Jobs	Success Rate
Small	9	0	0	9	100.0%
Medium	5	2	2	9	55.6%
Large	3	0	6	9	33.3%

이 결과는 모델 파라미터 수가 증가할수록 GPU 메모리 요구량과 로딩 시간이 급격히 증가하며, 특히 양자화를 적용하지 않은 BF16 조건에서는 24GB GPU 기반 서버에서 32B 이상 모델 평가가 제한적임을 보여준다.

3.5 Benchmark score 및 실행 시간 비교

정상적으로 평가가 완료된 job을 기준으로 benchmark score와 runtime을 비교하였다. 모든 서버에서 공통으로 성공한 3B, 4B, 9B, 12B 모델의 평균 overall score는 Srv-A 0.5329, Srv-B 0.5283, Srv-C 0.5282로 서버 간 차이가 크지 않았다. 반면 평균 runtime은 Srv-A 439.4초, Srv-B 1565.0초, Srv-C 1849.2초로 나타나 실행 시간에서는 뚜렷한 차이를 보였다.

[표 5] 공통 성공 모델 기준 overall score 및 runtime 비교

Server ID	Avg Overall Score	Avg Runtime
Srv-A	0.5329	439.4 sec
Srv-B	0.5283	1565.0 sec
Srv-C	0.5282	1849.2 sec

이 결과는 동일 모델이 정상적으로 평가를 완료한 경우, 서버 하드웨어 차이가 benchmark score보다는 runtime과 실행 가능성에 더 크게 반영됨을 보여준다.

4. 결론 및 향후 연구

본 연구는 non-quantized BF16 조건에서 서버 하드웨어별 대규모 언어 모델의 실행 가능성과 평가 결과를 비교하였다. 실험 결과, H200 4장을 탑재한 Srv-A는 3B부터 72B까지 모든 모델을 평가할 수 있었던 반면, RTX 4090 기반의 Srv-B와 Srv-C는 12B 모델까지는 평가가 가능했으나 32B 이상 모델에서는 timeout 또는 OOM이 발생하였다.

공통으로 평가가 완료된 모델에서는 서버 간 overall score 차이가 크지 않았지만, runtime과 실행 성공 여부에서는 큰 차이가 나타났다. 이는 서버 하드웨어 구성이 모델의 정답률보다는 모델 로딩 가능성, 메모리 부족 발생 여부, 평가 완료 시간에 더 직접적인 영향을 미친다는 것을 의미한다.

향후 연구에서는 양자화 적용, 병렬화 설정 변경, task limit 확대, 반복 실험을 통해 서버별 추론 효율성과 비용 대비 성능을 보다 정밀하게 분석할 필요가 있다.

참고문헌

[1] 정승호, 김도훈, 박진수, "대규모 언어 모델(LLM)의 포괄적 성능 비교 평가를 위한 평가 지표 및 데이터셋 개발 : 폐쇄형 LLM과 공개형 LLM의 비교를 중심으로", 경영정보학연구, Vol.26, No.3, pp.163-185, 2024.

[2] C. Park, H. Kim, D. Kim, S. Cho, S. Kim, S. Lee, Y. Kim, and H. Lee, "Open Ko-LLM Leaderboard: Evaluating Large Language Models in Korean with Ko-H5 Benchmark", Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, Vol.1, pp.3220-3234, 2024.

[3] G. Son, H. Lee, S. Kim, S. Kim, N. Muennighoff, T. Choi, C. Park, K. M. Yoo, and S. Biderman, "KMMLU: Measuring Massive Multitask Language Understanding in Korean", Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol.1, pp.4076-4104, 2025.

[4] S. Biderman, H. Schoelkopf, L. Sutawika, L. Gao, J. Tow, B. Abbasi, A. F. Aji, P. S. Ammanamanchi, S. Black, J. Clive, A. DiPofi, J. Etxaniz, B. Fattori, J. Z. Forde, C. Foster, J. Hsu, M. Jaiswal, W. Y. Lee, H. Li, C. Lovering, N. Muennighoff, E. Pavlick, J. Phang, A. Skowron, S. Tan, X. Tang, K. A. Wang, G. I. Winata, F. Yvon, and A. Zou, "Lessons from the Trenches on Reproducible Evaluation of Language Models", arXiv preprint arXiv:2405.14782, 2024.

[5] W. Kwon, Z. Li, S. Zhuang, Y. Sheng, L. Zheng, C. H. Yu, J. E. Gonzalez, H. Zhang, and I. Stoica, "Efficient Memory Management for Large Language Model Serving with PagedAttention", Proceedings of the 29th ACM Symposium on Operating Systems Principles, pp.611-626, 2023.