NetConfigQA2.0: 네트워크 설정 해석 능력 평가를 위한 질의응답 벤치마크

박유진[1], 이현정[1], 김기현[2], 박찬진[2], 박진영[3], 김태훈[1], 박천음[1]*

국립한밭대학교[1], 한국과학기술정보연구원[2], 성균관대학교[3]

{20191906, leehj}@edu.hanbat.ac.kr, {kkh1258, pcj0722}@kisti.re.kr, jy.bak@skku.edu, {thkim, parkce}@hanbat.ac.kr

NetConfigQA2.0: A Question-Answering Benchmark for Network Configuration Interpretation

Yujin Park[1], Hyeonjeong Lee[1], Ki-Hyeon Kim[2], Chanjin Park[2], Jinyeong Park[3], Taehoon Kim[1], Cheoneum Park[1]*

Hanbat National Univ.[1], KISTI[2], Sungkyunkwan Univ.[3]

요 약

본 논문은 거대언어모델(LLM)의 네트워크 설정 해석 및 운영 추론 능력을 체계적으로 평가하기 위해, Batfish 기반의 네트워크 질의응답 벤치마크인 NetConfigQA2.0을 제안한다. NetConfigQA2.0은 정적 분석에 국한된 기존 벤치마크와 달리, 실제 네트워크 설정 파일을 바탕으로 패킷 도달성 및 장애 시나리오와 같은 동적 추론 능력을 평가한다. 본 논문에서는 벤치마크 자동 구축 파이프라인과 함께, 네트워크 데이터의 수치적 정밀도를 평가하기 위한 Type-Aware Accuracy (TA-Acc) 지표를 도입한다. 실험 결과, 모델 간 성능 차이가 최대 0.73으로 기록되었으며, 상위 모델인 GPT-OSS-20B는 단순 설정 조회(L1-L2)에서 최대 0.873 정확도를 보였으나, 경로 추론 및 장애 분석(L4-L5) 단계에서는 모든 평가 모델이 0.3 이하의 낮은 성능을 보였다. 이는 NetConfigQA2.0이 LLM의 네트워크 설정 해석 능력을 평가하는 데 변별력이 있는 벤치마크임을 보인다.

I. 서론

네트워크 인프라의 복잡성이 증가함에 따라, 거대언어모델(LLM)을 활용한 네트워크 관리 자동화에 대한 연구가 활발히 진행되고 있다. 그러나 LLM의 네트워킹 응용에 관련된 기존 연구는 주로 의도 해석 및 네트워크 설정 생성 단계에 집중되어 있으며 [1] [2], 생성된 설정의 검증이나 운영 중인 토폴로지의 해석에 관한 연구는 상대적으로 미비하다[3]. 선행 연구인 NetConfigQA[4]는 PnetLab 환경에서 수집한 XML 로그를 기반으로 LLM의 네트워크 설정 해석 능력을 평가했으나, 정적 분석에 한정되어 패킷 도달성이나 장애 시나리오 기반의 추론형 질문 생성에는 한계가 존재한다.

본 연구에서는 이러한 한계를 개선하기 위해, 오픈소스 네트워크 분석 도구인 Batfish를 활용하여 생성한 네트워크 질의응답 데이터셋인 NetConfigQA2.0을 제안한다. 생성 파이프라인은 네트워크 설정 파일만을 입력으로 5단계 난이도의 질의응답 데이터셋을 자동 생성하며, Batfish가 제공하는 정형화된 분석 결과를 정답으로 활용하여 데이터셋의 신뢰성을 확보한다. 특히, 특정 실험 환경에 종속되지 않고 네트워크 설정 파일만을 입력으로 데이터셋을 자동으로 구축할 수 있다는 점에서 기존 연구와 차별성을 가진다.

II. 제안 방법

본 논문은 [그림 1]과 같이 PnetLab 환경에서 MPLS VPN 기반 네트워크(10개 노드: 4 Leaf, 2 PE, 4 P)를 구축하여 실험을 진행한다. PE 라우터는 VRF 기술로 네트워크를 격리하며 (VRF_AI/BIO/HPC), OSPF, MP-BGP, LDP를 활용한 표준 MPLS L3VPN 구조이다.

본 논문에서 제안하는 네트워크 QA 데이터셋 생성 파이프라인은

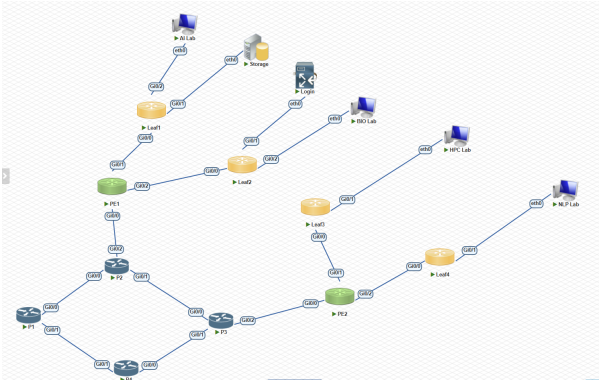

그림 1: PnetLab MPLS VPN 실험 환경 (3계층 구조)

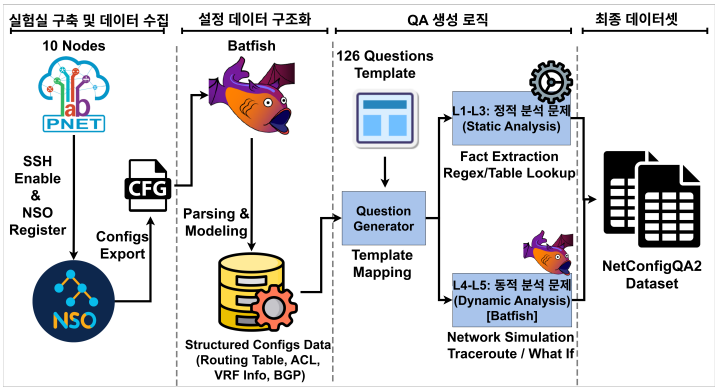

그림 2: NetConfigQA2 데이터셋 자동 생성 파이프라인

*Corresponding author

[그림 2]와 같이 구성된다. 파이프라인은 Cisco NSO를 활용하여 수집된 네트워크 설정(Config) 파일을 Batfish로 분석하여 인터페이스, 라우팅, VRF 등 네트워크 구성 정보를 추출한다. 이후 사전에 정의한 126개의 문제 템플릿을 기반으로 질문을 자동 생성한다. 데이터셋은 각 질문의 내용 기반의 난이도에 따라 Level 1–5로 나눈다. L1–L3 질문은 추출된 네트워크 구성 정보를 기반으로 질문의 정답을 생성하고, 경로 추론이나 장애 상황 분석이 요구되는 L4–L5 질문은 Batfish의 시뮬레이션 기능(Reachability, Link Failure 등)을 이용해 도달성과 경로 분석 결과를 정답으로 정의한다. 이와 같은 방식으로, 네트워크 설정 파일만 주어지면 다양한 네트워크 토폴로지와 설정 환경에 대한 벤치마크 데이터셋을 자동으로 구축할 수 있다. NetConfigQA2.0의 세부 구성은 [표 1]과 같다.

표 1: NetConfigQA2.0 메트릭 난이도 및 데이터셋(10 Nodes) 구성

Level	Description	Count
L1	단일 장비 설정 조회 (Hostname, IP, VRF 등)	364
L2	복수 장비 설정 집계 (SSH 활성화, OSPF Area 등)	21
L3	장비 간 설정 정합성 검증 (iBGP Mesh, L2VPN 등)	127
L4	동적 패킷 도달성 분석 (Trace, Reachability)	149
L5	장애 시나리오 기반 추론 (Link Failure Impact)	101
	Total Questions	**762**

III. 실험

본 실험에서는 제안하는 NetConfigQA2.0 데이터셋의 유효성을 평가하기 위해, 통신 도메인 벤치마크 데이터셋 3종(TeleQnA [5], TeleQuAD [6], NetBench [7])과 비교 분석을 수행한다. TeleQnA는 통신 표준 및 연구 논문 기반의 객관식 데이터셋으로 Accuracy를, TeleQuAD와 NetBench는 각각 추출형 및 서술형 QA 데이터셋으로 정답과의 의미 유사도를 평가하는 BERTScore를 활용한다. 반면, NetConfigQA2.0의 정답은 IP/JSON 등 구조화 값이므로 텍스트 유사도 기반 지표가 부적절하다. 본 논문에서는 답변 타입 t에 따라 적절한 비교 함수 $s_t(\cdot)$ (norm EM 또는 set/map F1)를 적용하고, 전체 성능을 평균으로 계산하는 Type-Aware Accuracy(TA-Acc)[수식 1]를 제안한다.

$$\text{TA-Acc} = \frac{1}{N}\sum_{i=1}^{N} s(p_i, g_i \mid t_i)$$
$$s(\cdot) = \begin{cases} \text{F1}_{\text{set/map}} & (t_i \in \text{Struct}) \\ \text{EM}_{\text{norm}} & (\text{otherwise}) \end{cases} \quad (1)$$

표 2: 통신 도메인 벤치마크에서의 LLM 성능 비교

Model	TeleQnA (Accuracy)	TeleQuAD (BERTScore)	NetBench (BERTScore)	Average (Norm.)
GPT-4o-mini	0.743	0.824	0.720	0.762
Llama3.1-8B	0.664	0.851	0.798	0.771
Mistral3-8B	0.706	0.866	0.790	0.787
Qwen3-8B	0.733	0.867	0.799	0.800
GPT-OSS-20B	**0.757**	**0.876**	**0.808**	**0.814**

표 3: NetConfigQA2.0에서의 LLM 성능 비교

Model	Rouge-L	BERTScore	EM	F1(Token)	TA-Acc
GPT-4o-mini	0.155	0.942	0.398	**0.539**	0.515
Llama-3.1-8B	0.314	0.897	0.176	0.302	0.291
Mistral3-8B	0.279	0.875	0.201	0.304	0.416
Qwen3-8B	0.414	0.932	0.339	0.472	0.465
GPT-OSS-20B	**0.439**	**0.942**	**0.437**	0.529	**0.612**

[표 2]는 통신 도메인 벤치마크에서의 LLM의 성능을 비교하며, 실험 결과 `GPT-OSS-20B`가 모든 벤치마크에서 가장 우수한 성능을 보이지만, 전반적으로 모델 간 성능 차이는 크지 않다.

[표 3]에서는 NetConfigQA2.0에서 LLM 간 성능을 비교한다. 실험 결과, NetConfigQA2.0이 기존 벤치마크에 비해 뚜렷한 성능 편차를 보인다. 이에 따라, 본 논문에서 제안한 구축 방법이 네트워크 설정 기반 추론 능력을 효과적으로 분석할 수 있음을 알 수 있다.

또한, Rouge-1, EM, F1에서는 큰 성능 차이를 보이지만 BERTScore에서는 유사한 성능을 보이는 모델들에 대해, TA-Acc는 이들 간의 성능 차이를 명확히 구분한다. 이는 본 논문에서 제안한 metric이 유형별로 적절한 평가를 수행함을 의미한다.

표 4: NetConfigQA2.0 난이도별(L1–L5) 성능 비교 (TA-Acc)

Model	L1	L2	L3	L4	L5
GPT-4o-mini	0.765	0.541	0.369	**0.267**	0.159
Llama-3.1-8B	0.368	0.371	0.305	0.184	0.138
Mistral3-8B	0.572	0.143	0.500	0.158	0.183
Qwen3-8B	0.639	0.294	0.431	0.256	**0.225**
GPT-OSS-20B	**0.873**	**0.873**	**0.605**	0.266	0.134

[표 4]는 문제 난이도별 실험 결과로, `GPT-OSS-20B`와 같은 큰 모델은 L1, L2 에서 최대 0.873의 성능을 보인다. 그러나 경량 모델들은 기초적인 L1 단계에서도 성능 편차가 크게 나타나며, L4, L5 에서는 모든 모델은 0.3 이하의 성능을 기록한다. 이는 LLM이 개별 설정 요소에 대한 이해도는 갖추었으나, 네트워크의 전체적인 상태를 논리적으로 추론하는 데는 명확한 한계가 있음을 보인다.

IV. 결론

본 논문에서 제안한 NetConfigQA2.0은 LLM의 네트워크 지식을 평가하는 변별력 있는 벤치마크임을 확인하였다. 향후 연구에서는 finetuning과 Batfish, Cisco NSO를 결합한 Multi-Agent system을 도입하여 Pnetlab 기반 실험실을 LLM이 직접 제어하도록 확장할 예정이다. 또한, 제안 방법을 네트워크 운영 태스크 중심 벤치마크로 발전시키고, 실제 네트워크 관리 성능을 평가할 계획이다.

사 사 문 구

본 과제(결과물)는 2025년도 교육부 및 대전광역시의 재원으로 대전 RISE센터의 지원을 받아 수행된 지역혁신중심 결과입니다. (2025-RISE-06-002)

참 고 문 헌

[1] C. Wang *et al*., "Netconfeval: Can llms facilitate network configuration?," *Proc. ACM Netw*., vol. 2, June 2024.

[2] S. K. Mani *et al*., "Enhancing network management using code generated by large language models," in *ACM Workshop on Hot Topics in Networks* (*HotNets*), November 2023.

[3] Boateng *et al*., "A survey on large language models for communication, network, and service management: Application insights, challenges, and future directions," *IEEE Communications Surveys & Tutorials*, 2025.

[4] Y. Park *et al*., "Netconfigqa: A question-answering dataset for network configuration interpretation," in *Proceedings of the 37th Annual Conference on Human and Language Technology*, (Daejeon, Korea), pp. 343–347, Korea Science, 2025.

[5] A. Maatouk *et al*., "Teleqna: A benchmark dataset to assess large language models telecommunications knowledge," 2023.

[6] F. Gebre *et al*., "Telequad: A suite of question answering datasets for the telecom domain," 2025.

[7] [NetoAi], "Netbench dataset," 2025.