# 양자화된 다국어 LLM 의 한국어 성능 개선을 위한 LAPE 기반 AWQ 연구

이재영 [1], 최진영 [2]
[1] 아주대학교 산업공학과 석사과정
[2] 아주대학교 산업공학과 교수

ljae1129@ajou.ac.kr, choijy@ajou.ac.kr

# A Study on LAPE-Based AWQ for Improving Korean Performance of Quantized Multilingual LLM

Jaeyoung Lee[1], Jin Young Choi[2]
[1]Dept. of Industrial Engineering, Ajou University
[2]Dept. of Industrial Engineering, Ajou University

**요 약**

대규모 언어 모델(LLM)은 높은 계산 비용과 메모리 요구량으로 인해 실용적인 배포와 활용에 어려움이 있다. 이를 해결하기 위해 모델의 크기를 줄이고 연산 효율성을 향상하는 다양한 양자화 기법이 개발되었다. 그러나 현재까지의 연구는 주로 영어 기반 모델과 데이터에 초점을 맞추고 있으며, 비영어권 언어에서의 성능 저하 문제는 거의 다루어지지 않고 있다. 최근 연구에 따르면, 양자화를 다국어 대규모 언어 모델에 적용할 경우 성능 하락이 영어보다 비영어권 언어에서 훨씬 더 크게 나타나는 경향이 있으며 한국어 또한 그에 해당한다. 본 연구에서는 이러한 문제를 해결하기 위해 한국어에 최적화되도록 개량된 양자화 기법을 제안한다. 개선된 양자화 방법은 한국어 태스크에 대해 기존 양자화 방법에 비해 개선된 성능을 보였다.

## 1. 서론

대규모 언어 모델(Large Language Models, LLM)은 자연어 처리(Natural Language Processing, NLP)의 성능을 획기적으로 향상시켰다. 그러나 모델의 규모가 크기 때문에 실제 환경에서의 배포와 활용에는 여러 제약이 따른다. 특히 고성능 하드웨어가 요구되므로, 실시간 응용이나 자원이 제한된 환경에서는 사용이 어렵다. 이를 해결하기 위해 다양한 양자화(quantization) 기법이 개발되었으며, 이는 모델의 메모리 사용량을 줄이고 연산 효율성을 향상시키는 데 효과적인 방법으로 활용되고 있다.

그러나 기존의 LLM 및 양자화 연구는 대부분 영어를 중심으로 진행되었으며, 비영어권 언어에 대한 연구는 상대적으로 부족한 실정이다. 최근 연구에 따르면, 기존 양자화 기법을 비영어권 모델에 적용할 경우 성능 저하가 영어보다 더욱 두드러지며, 특히 한국어에서 그 영향이 크게 나타난다[1].

본 연구에서는 다국어 LLM 을 양자화할 때 발생하는 한국어 성능 저하 문제를 해결하기 위해, 한국어에 최적화된 양자화 기법을 개발하고 이를 통해 성능 저하를 최소화하는 방법을 제안한다. 이를 위해 다양한 실험을 수행하고, 기존 기법과 비교하여 제안 기법의 효과를 분석한다.

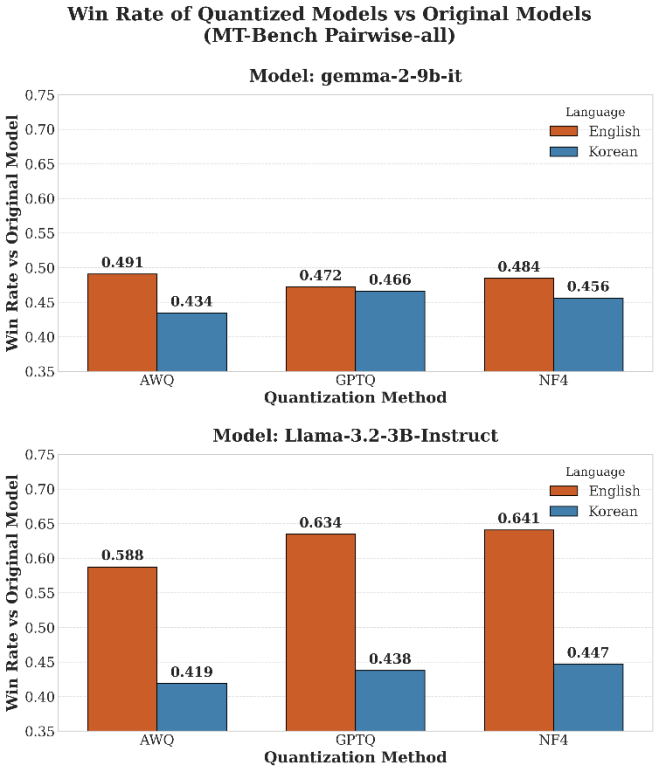
## 2. 기존 양자화 기법의 한국어 성능 저하 문제

LLM 의 양자화는 모델 크기를 줄이고 연산 효율성을 높이기 위한 핵심 기술이다. 가중치를 단순히 낮은 비트(bit)로 변환하는 방식은 모델 성능의 저하를 초래할 수 있다. 이를 해결하기 위해 성능을 보존하면서 양자화를 하는 다양한 방법이 개발되었다.

현재 다양한 양자화 방법 중 활발하게 사용되는 대표적인 양자화 기법으로는 GPTQ, AWQ, NF4 등이 있다[2-4]. 이들은 모두 가중치만을 양자화 하는 기법으로, 가중치를 4 비트 혹은 그 이하의 정밀도로 변환하여 모델의 파라미터 크기를 줄이고 추론 효율성을 높이는 데 초점을 맞춘다. 이러한 기법들은 추가적인 재학습 없이 빠르게 적용 가능하다는 장점이 있어 최근 다양한 환경에서 널리 활용되고 있다.

하지만 이들 대부분의 연구는 영어 기반의 모델과 데이터를 중심으로 이루어졌으며, 비영어권 언어에 대해서는 충분한 고려가 이루어지지 않았다. 최근 연구에 따르면, 양자화된 LLM 은 비영어권 언어에서 성능 저하가 두드러지게 나타나고 특히 한국어와 같은 비라틴계 언어는 그 영향이 큰 것으로 보고되었다 [1].

이를 검증하기 위해 본 연구에서는 최근 한국어를 포함한 다국어 태스크에서 뛰어난 성능을 보이고 있는 다국어 LLM 중 Google 의 gemma-2-9b-it[5]와 Meta AI 의 Llama-3.2-3B-Instruct [6], 두 개의 LLM 을 선정하고 해당 두 모델을 다양한 방법으로 양자화를 수행한 후, 각 언어에 대한 벤치마크를 통해 성능 비교를 수행하였다. 기존 양자화 연구는 영어 중심으로 이루어졌기 때문에, 양자화 성능 평가 또한 대부분 영어 데이터를 기준으로 수행되어 왔다. [1] 연구에 따르면, 언어 모델을 양자화하면 비영어권 언어에서의 성능 저하가 영어보다 훨씬 더 크게 나타나며, 특히 사람이 직접 평가하는(human evaluation) 경우 그 영향이 더욱 두드러지게 나타나고, 자동화된 벤치마크는 성능 하락의 영향을 과소평가하는 것으로 보고되었다. 따라서 본 연구에서는 한국어 성능 저하를 명확하게 파악하기 위해 인간의 평가와 유사한 MT-Bench [7]를 사용하여 영어 성능을 평가하고, 한국어 성능은 MT-Bench 를 한국어에 맞게 개량한 KoMT-Bench [8]를 활용하여 평가하였다.




(그림 1) 양자화된 다국어 LLM 의 영어와 한국어에 대한 성능 하락 비교 그래프

(그림 1)은 선정한 각 다국어 LLM 을 AWQ, GPTQ, NF4, 세 개의 방법으로 양자화를 수행하고 양자화 하지 않은 원본 모델과 MT-Bench 의 Pairwise-all 방법으로 벤치마크를 수행하고 성능을 비교한 그래프이다. 해당 벤치마크는 양자화된 모델과 원본 모델의 160 개 질문에 대해 답변을 각각 생성하고 두 답변을 비교하여 승/무/패를 판정하는 벤치마크이다.

각 막대 그래프는 양자화된 모델의 원본 모델에 대한 승률을 의미한다. 두 LLM 모두 값의 차이는 있지만 세 가지 양자화 방법에서 모두 양자화된 모델의 한국어 태스크에 대한 승률이 영어 태스크에 대한 승률보다 전체적으로 더 낮은 것을 확인할 수 있다. 이를 통해 양자화된 LLM 에서 한국어 성능의 하락이 영어에 비해 더 두드러지는 경향이 있다는 것을 알 수 있다. 따라서 이와 같이 양자화된 LLM 에서 한국어의 성능이 하락하는 문제를 해결할 필요성이 있다.

## 3. 한국어 성능이 개선된 LAPE 기반 AWQ

현재까지 다양한 양자화 기법들이 제안되어 왔으나, 각 방법마다 양자화를 수행하는 방식과 기준이 다르기 때문에 하나의 공통된 방향으로 모든 기법을 동시에 개선하기는 어렵다. 따라서 본 연구에서는 양자화에 활성화(activation) 값을 활용하는 AWQ 를 기반 방법으로 선정하고, 이를 개선하는 방식으로 접근한다. 본 연구에서는 양자화된 LLM 에서 한국어 성능이 저하되는 문제를 해결하기 위해 LAPE 기법을 기반으로 한국어에서 중요한 뉴런을 식별하고, 이를 보존하는 방식으로 AWQ 를 개선하는 방법을 제안한다.

AWQ 는 모델의 활성화 값의 분포를 고려하여 중요한 가중치는 유지하면서 나머지 가중치를 양자화하는 방법이다. 구체적으로, AWQ 는 소량의 교정 데이터(calibration data)를 활용하여 모델의 활성화 분포를 측정한 후, 뉴런이 입력 데이터에 대해 어떻게 활성화되는지를 분석한다. 그 결과를 바탕으로 활성화 분포 값이 높은 가중치는 중요도가 높은 것으로 판단하여 최대한 보존하고, 나머지 가중치를 양자화하는 방식을 적용한다. 이를 통해 AWQ 는 성능 저하를 최소화하면서도 높은 연산 효율성을 유지할 수 있다. 그러나 이 방식은 입력 데이터에 대한 뉴런의 활성화 값만으로 가중치의 중요도를 판단하기 때문에, 특정 언어에서 중요한 역할을 하는 뉴런이 간과될 수 있다. 특히 다국어 모델의 경우, 언어별로 활성화되는 뉴런 분포가 다르기 때문에, 영어 중심의 교정 데이터를 사용할 경우 (그림 1)의 결과와 같이 비영어권 언어에서 성능 저하가 나타날 수 있다.

이러한 문제를 해결하기 위해 본 연구에서는 Tang et al.(2023)에서 제안된 LAPE 를 활용한다[9]. LLM 의

다국어 처리 능력을 분석하기 위해 [9] 연구는 언어 활성화 확률 엔트로피 (Language Activation Probability Entropy, LAPE) 라는 개념을 제안했다. LAPE 는 언어 모델 내부에서 특정 언어에 특화된 뉴런(language-specific neurons)을 식별하는 방법이다. LAPE 는 다양한 언어의 텍스트를 입력했을 때, 특정 뉴런의 활성화 확률을 측정하는 방식으로 작동한다. 구체적으로, LLM 의 $i$번째 층(layer)의 $j$번째 뉴런이 특정 언어 $k$를 처리할 때의 활성화 확률은 수식 (1)과 같이 계산된다.

$$p_{i,j}^{k} = \mathbb{E}\left(\mathbb{I}(\boldsymbol{act\_fn}(\tilde{\boldsymbol{h}}^{\boldsymbol{i}}\boldsymbol{W}_{\mathbf{1}}^{\boldsymbol{i}})_j > 0) \mid language\ k\right) \quad (1)$$

$$LAPE_{i,j} = -\sum_{k=1}^{l} p_{i,j}^{\prime k}\, log\left(p_{i,j}^{\prime k}\right) \quad (2)$$

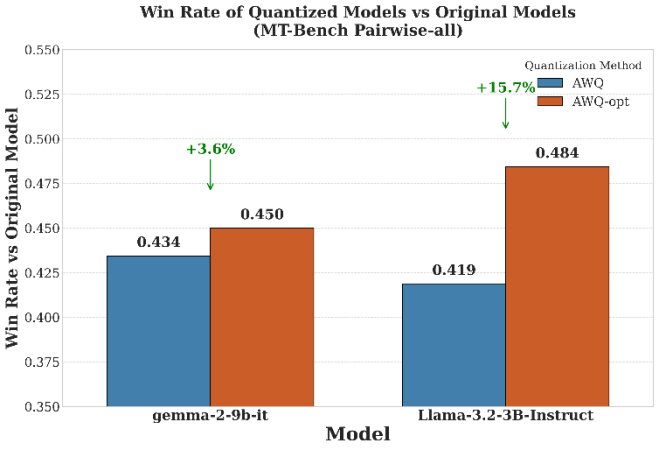
그리고 수식 (1)을 통해 산출된 $p_{i,j}^{k}$ 값을 유효한 확률 분포로 바꾸기 위해 L1 normalization 을 적용하여 $p_{i,j}^{\prime}$를 산출하고 수식 (2)와 같이 LAPE Score 를 계산하여 값이 낮은 뉴런은 언어 특화 뉴런으로 지정된다. LAPE Score 가 낮다는 것은 해당 뉴런이 특정 언어에 대해 높은 활성화 확률을 보이고, 다른 언어에 대해서는 낮은 활성화 확률을 보인다는 것이기 때문이다.

[9] 연구는 영어, 중국어, 일본어, 프랑스어, 스페인어, 베트남어, 인도네시아어 등 총 7 개 언어를 사용하여 언어 특화 뉴런을 분석하였다. 본 연구에서는 [9] 연구에 포함되지 않은 한국어를 추가하여, 총 8 개 언어의 데이터를 기반으로 뉴런의 활성화 특성을 분석하였다. 이 과정을 통해 본 연구는 한국어에서 주로 활성화되는 뉴런을 식별하고, 해당 뉴런 정보를 양자화 과정에 반영하여 보존한다. 식별된 한국어 특화 뉴런은 전체 모델 파라미터 중 1% 미만으로 극히 일부에 해당하기 때문에, 이들의 가중치를 보존하더라도 전체 모델의 크기에는 큰 영향을 미치지 않는다.

즉, 기존의 AWQ 는 영어 기반 교정 데이터를 바탕으로 활성화 분포를 분석하여 가중치의 중요도를 평가했지만, 본 연구는 여기에 한국어 특화 뉴런 정보를 추가적으로 반영함으로써, 한국어 성능 저하를 최소화하는 양자화 기법을 제안한다.

## 4. 실험

본 연구에서 제안한 LAPE 기반 한국어 최적화 AWQ 의 성능 검증을 위해 앞서 수행한 양자화된 LLM 에서 성능 하락 검증과 마찬가지로 Google 의 gemma-2-9b-it 와 Meta AI 의 Llama-3.2-3B-Instruct 의 두 가지 LLM 으로 실험을 수행한다. LAPE score 를 구하기 위한 데이터 셋은 [9] 연구와 마찬가지로 각 언어의 위키백과(Wikipedia) 덤프에서 문서를 무작위로 추출하고 토큰화(tokenization)하여 한국어를 포함한 8 개 언어에 대해 각각 100 만 개의 토큰으로 구성되었다. 해당 데이터를 통해 각 모델에서 LAPE Score 를 계산하여 한국어에 특화된 뉴런을 찾고, 해당 뉴런의 가중치를 별도로 보존하면서 AWQ 를 통해 양자화를 진행하였다. (그림 2)는 gemma-2-9b-it 와 Llama-3.2-3B-Instruct 모델에 대해 기존 AWQ 와 본 연구에서 제안한 한국어에 최적화된 양자화 방법(AWQ-opt)의 성능을 비교한 결과이다. 두 양자화 방법 모두 원본 모델과 KoMT-Bench 의 Pairwise-all 방식으로 한국어 태스크에 대해 평가되었다. 실험 결과, 한국어에 최적화된 AWQ-opt 는 gemma-2-9b-it 에서 기존 AWQ 보다 3.6%, Llama-3.2-3B-Instruct 에서는 15.7% 높아진 성능을 보였다. 이러한 성능 향상은 소량의 뉴런 가중치만을 보존함으로써 달성된 것으로, gemma-2-9b-it 모델에서는 한국어 특화 뉴런의 가중치를 7.86MB 만큼 추가로 보존해서 양자화된 모델의 크기인 6.16GB 대비 약 0.13% 증가하였고, Llama-3.2-3B-Instruct 모델에서는 7.72MB 만큼 추가로 보존해서 3.04GB 대비 약 0.25% 증가하여 모델의 크기에는 거의 영향을 미치지 않으면서 유의미한 성능 향상을 달성하였음을 알 수 있다. 이러한 결과는 언어별 뉴런 활성화 특성을 반영한 양자화 전략이 다국어 LLM 의 성능 저하를 효과적으로 완화할 수 있음을 실증적으로 보여주며, 자원 효율성을 유지하면서 비영어권 언어의 활용 가능성을 높이는 실용적인 대안이 될 수 있음을 시사한다.




(그림 2) 기존 AWQ 와 한국어에 최적화된 AWQ 의 한국어 성능 비교 그래프

## 5. 결론

본 연구에서는 LLM 의 양자화 과정에서 발생하는 한국어 성능 저하 문제를 해결하기 위해 언어 활성화 확률 엔트로피(LAPE)를 기반으로 한 한국어 최적화 AWQ 방법을 제안하였다. 기존의 양자화 연구들이 주로 영어 중심으로 이루어져 왔고, 비영어권 언어, 특

히 한국어에서는 상대적으로 큰 성능 하락이 발생한다는 문제에 초점을 맞추었다.

실험 결과, 제안된 방법은 gemma-2-9b-it 와 Llama-3.2-3B-Instruct 모델에서 기존 AWQ 대비 한국어 성능을 상당히 개선시켰다. 특히 한국어에 특화된 뉴런을 식별하고 이를 양자화 과정에서 보존함으로써, 양자화된 모델의 크기를 거의 그대로 유지하면서 한국어 성능의 저하를 최소화할 수 있었다. KoMT-Bench 를 통한 평가에서 한국어에 최적화하여 개선한 AWQ 로 양자화한 다국어 LLM은 기존 AWQ로 양자화한 다국어 LLM 보다 높은 승률을 보였으며, 이는 제안된 방법론의 효과를 입증한다.

본 연구 결과는 양자화된 다국어 LLM 을 한국어와 같은 비영어권 언어 환경에서도 효율적으로 활용할 수 있는 가능성을 보여준다. 특히 한국어 특화 뉴런의 선택적 보존을 통해 손실되는 한국어 성능을 최소화할 수 있음을 실증함으로써, 제한된 컴퓨팅 자원 하에서도 다국어 LLM 의 실용성과 확장성을 높일 수 있는 현실적인 해법을 제시한다.

## 참고문헌

[1] Marchisio, Kelly, et al. "How Does Quantization Affect Multilingual LLMs?.", Findings of the Association for Computational Linguistics: EMNLP 2024, pages 15928–15947, 2024.

[2] Frantar, Elias, et al. "GPTQ: Accurate Post-Training Quantization for Generative Pre-trained Transformers", arXiv preprint arXiv:2210.17323, 2022.

[3] Lin, Ji, et al. "AWQ: Activation-aware Weight Quantization for On-Device LLM Compression and Acceleration.", Proceedings of Machine Learning and Systems, 6, 87-100, 2024.

[4] Dettmers, Tim, et al. "QLoRA: Efficient Finetuning of Quantized LLMs.", Advances in neural information processing systems 36, 10088-10115, 2023.

[5] Gemma Team. (2024). Gemma. Kaggle. https://www.kaggle.com/m/3301

[6] Meta. (2024, September 25). meta-llama/Llama-3.2-3B-Instruct. Hugging Face. https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct

[7] Zheng, Lianmin, et al. "Judging llm-as-a-judge with mt-bench and chatbot arena.", Advances in Neural Information Processing Systems 36: 46595-46623, 2023.

[8] LG AI Research. (2024). KoMT-Bench [Data set]. Hugging Face. https://huggingface.co/datasets/LGAI-EXAONE/KoMT-Bench

[9] Tang, Tianyi, et al. " Language-Specific Neurons: The Key to Multilingual Capabilities in Large Language Models.", Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 5701–5715, 2024.