🔍 Through a Compressed Lens: Investigating The Impact of Quantization on Factual Knowledge Recall

Qianli Wang[1,2] **Mingyang Wang**[4,5,6] **Nils Feldhus**[1,2,8] **Simon Ostermann**[2,3,7]
Yuan Cao[9] **Hinrich Schütze**[4,6] **Sebastian Möller**[1,2] **Vera Schmitt**[1,2]

[1]Quality and Usability Lab, Technische Universität Berlin [2]German Research Center for Artificial Intelligence (DFKI)
[3]Saarland Informatics Campus [4]LMU Munich [5]Bosch Center for Artificial Intelligence (BCAI)
[6]Munich Center for Machine Learning (MCML) [7]Centre for European Research in Trusted AI (CERTAIN)
[8]BIFOLD – Berlin Institute for the Foundations of Learning and Data [9]Technical University of Munich

Correspondence: qianli.wang@tu-berlin.de

Abstract

Quantization methods are widely used to accelerate inference and streamline the deployment of large language models (LLMs). Although quantization's effects on various LLM capabilities have been extensively studied, one critical area remains underexplored: factual knowledge recall (FKR), the process by which LLMs access stored knowledge. To this end, we conduct comprehensive experiments using three common quantization techniques at distinct bit widths, in conjunction with interpretability-driven analyses on two tasks, *knowledge memorization* and *latent multi-hop reasoning*. We show that quantization typically results in information loss within LLMs, consequently diminishing their capacity for FKR. This effect is particularly amplified in smaller models within the same architectural families. However, models quantized at reduced bit precision do not consistently exhibit inferior performance and occasionally quantization may even enhance model FKR. We find that `BitSandBytes` demonstrates highest preservation of the original full-precision model's FKR. Despite variability across models and methods, quantization causes modest performance degradation and remains an effective compression strategy.

1 Introduction

The shift towards LLMs has created strong demand for efficient inference and accessible deployment. In response, numerous quantization techniques have been developed (Dettmers et al., 2022; Frantar et al., 2023; Xiao et al., 2023; Lin et al., 2024). By reducing the precision of a model's parameters, quantization allows us to decrease model size while mostly preserving its performance (Gray and Neuhoff, 1998). Although the effects of quantization in LLMs have been evaluated across various aspects, e.g., multilinguality, bias, fairness, and trustworthiness (Marchisio et al., 2024; Gonçalves and Strubell, 2023; Ramesh et al., 2023; Hong et al., 2024; Liu et al., 2024; Wang et al., 2026), the *degree*, *variability*, and *practical implications* of its impact on LLMs' ability to recall factual knowledge from pretrained memory remains underexplored and have yet to be comprehensively characterized.

To this end, we present a comprehensive study of quantization's impact on factual knowledge recall in LLMs. Instead of solely observing the performance degradation, we perform interpretability-driven analyses on *knowledge memorization* and *latent multi-hop reasoning* tasks (§3.1), to scrutinize behavior across **neuron**, **layer**, and **model** levels and evaluate three LLMs across three widely adopted model quantization techniques at different bit widths (Figure 1). We reveal that quantization typically impairs FKR, which is particularly acute for smaller models within a given model family. LLMs quantized to a lower bit precision do not consistently underperform those with higher precision. In some cases, quantization can even paradoxically enhance factual recall. Our findings indicate that `BitSandBytes` is most effective at preserving the model's FKR. We observed no significant FKR degradation from quantization that would compromise model compression effectiveness, though effects vary by model and technique.

2 Background and Related Work

Quantization. Quantization techniques compress LLMs by converting model weights, activations, or the KV cache into lower-precision data types (Zhu et al., 2024). These techniques can be broadly categorized into two types: quantization-aware training (QAT) and post-training quantization (PTQ). QAT requires retraining to mitigate errors introduced by quantization, whereas PTQ facilitates the direct application of a quantized model during inference. In this paper, we primarily evaluate the impact of weight-only quantization (§3.3)

Proceedings of the 6th Workshop on Trustworthy NLP (TrustNLP 2026), pages 21–39

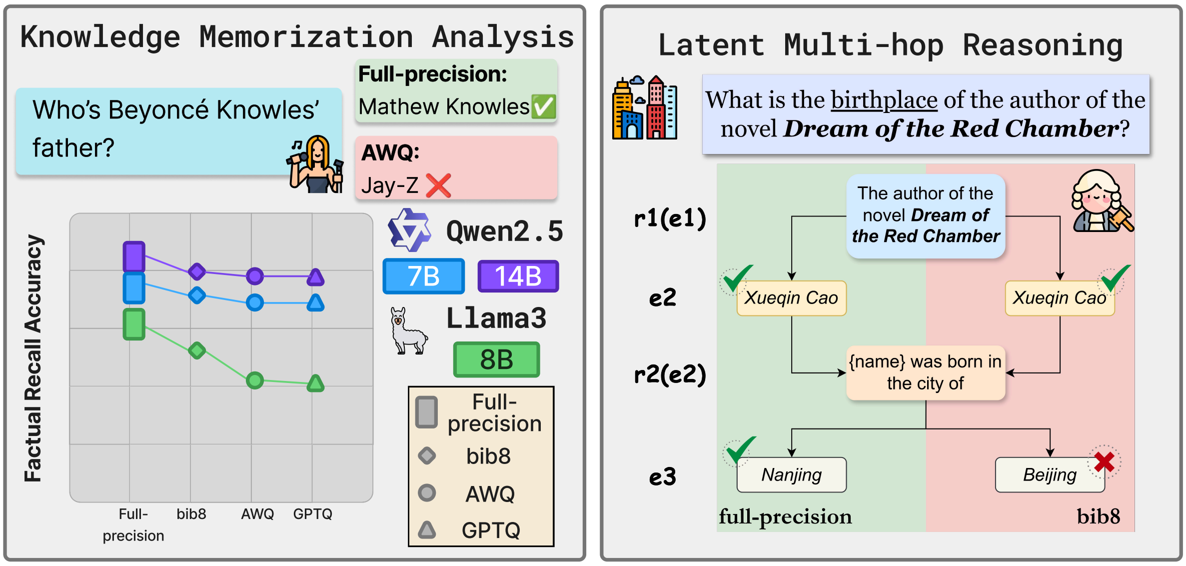


Figure 1: The effect of quantization on factual knowledge recall through *knowledge memorization analysis* and *latent multi-hop reasoning analysis*.

on FKR (§4), which eliminates the need for retraining to address errors resulting from quantization.

Impact of Quantization. Recent work has extensively examined the impact of quantization on various capabilities of LLMs. Marchisio et al. (2024) conduct a thorough analysis of quantized multilingual LLMs, focusing on performance degradation across languages. Gonçalves and Strubell (2023); Kirsten et al. (2025) explore the emergence of bias in the outputs generated by quantized models. Liu et al. (2024) find that in-context learning ability gradually declines in heavily quantized LLMs. Jin et al. (2024) observe that models with 4-bit quantization can still retain the alignment ability. Singh and Sajjad (2025) investigate the impact of quantization on model calibration. Wang et al. (2026) explore the impact of quantization on self-explanation quality and faithfulness. In our work, we explicitly explore how quantization affects FKR.

3 Experimental Setup

We examine two representative methods (§3.1) to evaluate the impact of quantization on FKR. Specifically, we compare full-precision LLMs (§3.4) with LLMs quantized using different techniques and bit configurations (§3.3) across two datasets (§3.2).

3.1 Methods

Knowledge Memorization Analysis. We investigate the memorization of a model's factual knowledge, identifying the reasons behind potential factual forgetting (Namburi et al., 2023). By leveraging the theory of knowledge neurons, which suggests that specific neurons in LLMs are responsible for storing specific pieces of knowledge, we explore how quantization alters the storage and retrieval processes within these neurons.

Latent Multi-hop Reasoning Analysis. We adapt the methodology from Yang et al. (2024) for inspecting latent multi-hop reasoning errors. Specifically, the analysis tests whether full-precision and quantized LLMs employ similar latent reasoning pathways and internal knowledge recall mechanisms to answer complex factual queries.

3.2 Datasets

Our study employs two widely recognized datasets[1] for evaluating FKR with selected interpretability methods (§3.1).

LRE (Hernandez et al., 2024) is a knowledge probing dataset consisting of knowledge triplets (s, r, o), structured in a one-hop setting, where s, r, and o represents the subject, relation, and object, respectively.

TwoHop-Fact (Yang et al., 2024) is a dataset consisting of pairs of prompts: two-hop prompts (τ_{2H}) for compositional queries, representing fact composition queries in the form $((e_1, r_1, e_2), (e_2, r_2, e_3))$, where r_1 and r_2 are relations and e_i denotes an entity; and one-hop prompts (τ_{1H}) for subqueries

[1] Dataset examples are detailed in Appendix A.

(Figure 4). e_1 and e_2 (in the second triplet) are subjects, while e_2 (in the first triplet) and e_3 objects. e_2 serves as a **bridge entity**, linking the two triplet to form a coherent two-hop reasoning chain.

3.3 Quantization Techniques

Building on the prior discussion (§2), we identify three commonly used PTQ techniques applied to the selected LLMs (§3.4) in our experiments:

- `GPTQ` (Frantar et al., 2023) uses a second-order, Hessian-based optimization to quantize weights post-training with minimal accuracy loss;
- `AWQ` (Lin et al., 2024) enhances weight quantization by handling activation outliers to preserve model accuracy at low bit-widths;
- Integer quantization (Dettmers et al., 2022) implemented by BITSANDBYTES (`bib4` and `bib8`) enables fast and memory-efficient inference by using optimized low-bit kernels.

3.4 Models

We evaluate `Llama3-8B` (AI@Meta, 2024), `Qwen2.5-7B`, and `Qwen2.5-14B` (Qwen et al., 2024) using two interpretability methods (§3.1). These models are selected due to the availability of quantized versions for each[2], ensuring the reproducibility of our results.

4 Evaluation Setup

Knowledge Memorization Analysis. To assess the effects of quantization on FKR, we first evaluate factual recall accuracy using LRE. A comparison of results before and after quantization reveals the degree of knowledge forgetting. We then employ a knowledge attribution method (Yu and Ananiadou, 2024) to trace the information loss back to specific layers and neurons, following the knowledge neuron theory. This combined analysis uncovers how quantization impacts the internal mechanisms responsible for storing and retrieving information.

Latent Multi-hop Reasoning Analysis (LMHR). Following Yang et al. (2024), we employ three metrics to evaluate the impact of quantization on FKR, i.e., LMHR. The Entity Recall Score (ENTREC) measures the LLM's ability to recall the bridge entity e_2 within a two-hop prompt τ_{2H} (Figure 4). ENTREC is defined with respect to the hidden representation in a specific layer ℓ, at the final position of the bridge entity's descriptive mention in the two-hop prompt. A higher $\text{ENTREC}_\ell(e_2, \tau_{2H})$ indicates stronger internal recall of the bridge entity e_2 at the ℓ-th layer. The Consistency Score (CNSTSCORE) assesses how consistently an LLM responds to both the two-hop and one-hop prompts. CNSTSCORE calculates the similarity between the output probability distributions in response to the τ_{2H} and τ_{1H} prompts to measure the consistency between the two outputs. Additionally, we evaluate FKR accuracy in predicting the target object e_3.

[2]Detailed information for each model, including links to their respective quantized versions, is provided in Table 1.

5 Results

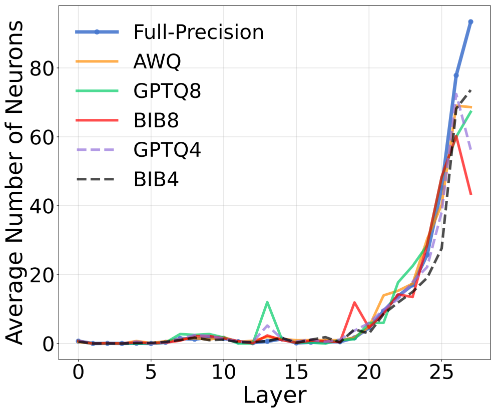
5.1 Knowledge Memorization Analysis

Table 2 shows that quantization introduces varying degrees of accuracy drop depending on both the model size and quantization method. Notably, the accuracy drop is more obvious in smaller models within the same model family. Additionally, the accuracy of `AWQ`, `GPTQ4`, and `bib4` quantized models consistently drop across all model sizes, whereas `GPTQ8` and `bib8` models retain performance comparable to the full-precision models. Furthermore, relations where full-precision performance has not saturated tend to exhibit more severe factual knowledge recall degradation (Table 4), indicating that such knowledge is more fragile under quantization.

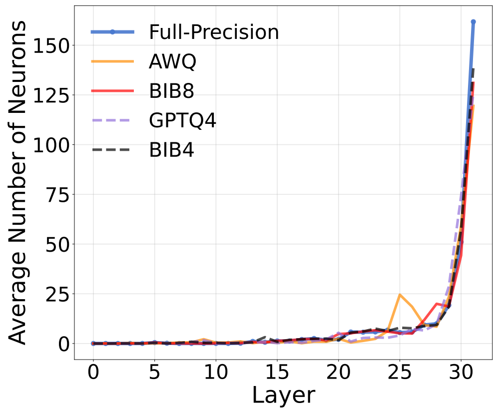
Neuron-level Trends. Following the neuron-level method of Yu and Ananiadou (2024), we assign each neuron in the model a contribution score equal to the increase in log-probability of the correct answer token that the neuron induces. Comparing these scores before and after quantization allows us to quantify how much individual neurons' influence changes. We analyze four relations that suffer a notable accuracy drop: *landmark on continent*, *person father*, *person mother*, and *person sport position*.

We track the top-300 feed-forward neurons in full-precision model, set τ as their minimum contribution score, and count how many neurons in quantized models exceed τ. As shown in Figure 2a and 2b, the per-layer counts decrease after quantization for both models, with the most obvious reductions in the last layers. Complete results across relations are provided in Figures 7 and 8 in App. C.1.

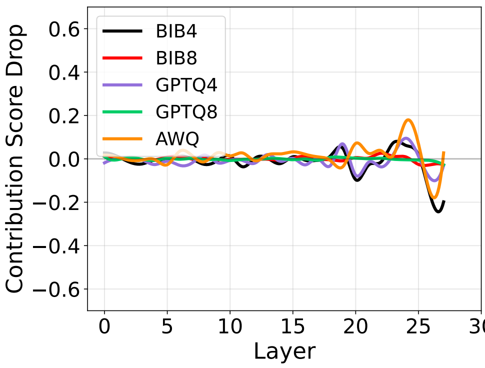
Layer-wise Trends. Figure 2 depicts the drop in aggregate contribution scores for attention (*left*) and feed-forward (*right*) sub-layers. All quantiza-



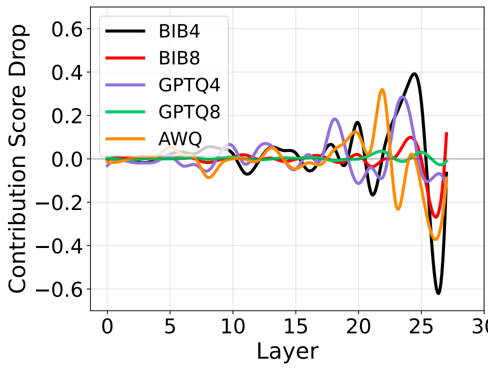
(a) Top neuron distribution (`Qwen2.5-7B`)



(b) Top neuron distribution (`Llama3-8B`)



(c) Attention (`Qwen2.5-7B`): Landmark on continent



(d) FFN (`Qwen2.5-7B`): Landmark on continent

Figure 2: **Top:** Distribution of high-contributing neurons across layers, showing the average number of top-300 neurons per layer for `Qwen2.5-7B` (*left*) and `Llama3-8B` (*right*). **Bottom:** Layer-wise drop in neuron contribution scores across quantization methods for the *landmark on continent* relation in `Qwen2.5-7B`, comparing attention sublayers (*left*) and feed-forward sublayers (*right*).

tion methods exhibit a pronounced decline in the final two layers on `Qwen2.5-7B`, this trend is consistent on all relations we investigate, as shown in Figure 5 in App C.1. While on `Llama3-8B` (see Figure 6), the decline occurs in middle-to-late layers, with an increase in the final layers, indicating that information loss patterns vary across model architectures. This divergence may be ascribable to the different ways in which factual knowledge is stored across model families (Choe et al., 2025).

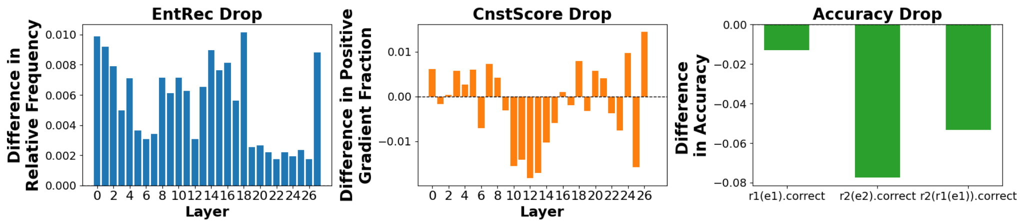
Summary. Collectively, both layer-level and neuron-level analyses reveal that quantization primarily affects the network's last layers. These findings confirm that quantization degrades the decisive information stored in these late layers, accounting for the factual knowledge recall degradation reported in Table 2.[3]

[3]The complete analysis of both models across different relations is provided in Appendix C.1.

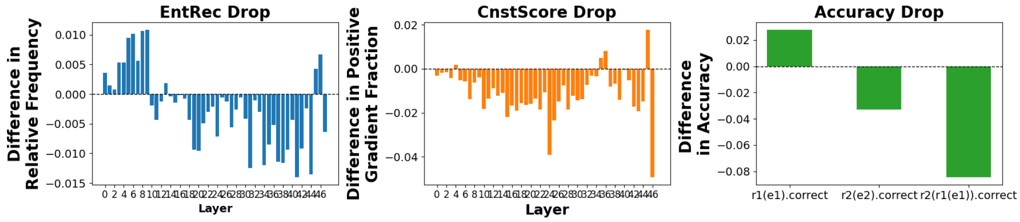
5.2 Latent Multi-hop Reasoning Analysis

Quantization affects the first-hop Reasoning the most. Table 3 reveals that quantization substantially affects the first hop $r_1(e_1)$, by as much as 30.08%, while its impact on the second hop is minimal, with an average of degradation of only 4.25%. The $r_2(r_1(e_1))$ deterioration due to quantization is strongly correlated with the ability to correctly predict the bridge entities $r_1(e_1)$, as indicated by a Spearman's correlation of 0.93. Nevertheless, the FKR deterioration is not dramatic and considerably acceptable (e.g., `Qwen2.5-7B` shows a minor average deterioration of 0.77%).

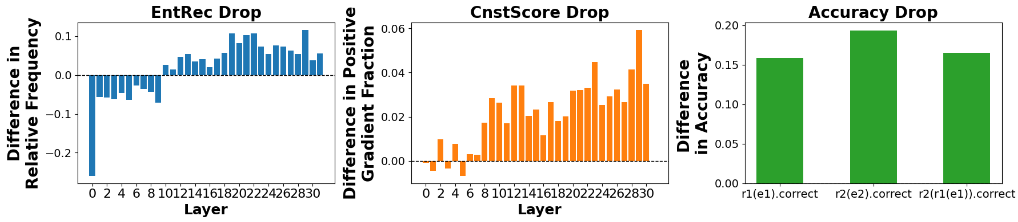
Quantization effects are not consistent. Figure 3 illustrates that quantization effects on FKR are largely unpredictable. The quantization effect is heterogeneous across layers and variable across models, particularly across architectures, given that facts are stored in various ways across `Llama3`



(a) `Qwen2.5-7B`



(b) `Qwen2.5-14B`



(c) `Llama3-8B`

Figure 3: Difference in the *entity recall score* (ENTREC), *consistency score* (CNSTSCORE), and *accuracy* between the `GPTQ8` quantized and full-precision models, evaluated across all layers.

and `Qwen2.5` models (Choe et al., 2025). Besides, different quantization approaches often affect the model FKR in different manners (Figure 11). Nevertheless, for a given quantization method with different bit width, the layer-wise impact remains broadly similar. Surprisingly, quantization can occasionally even improve FKR (Figure 3). This effect may be attributed to the regularization effect (Park et al., 2022) or quantization-induced noise (Li et al., 2024), which can inadvertently enhance the model's capability to recall factual knowledge.

`bib` largely preserve the factual knowledge recall capability and cross-model comparison. For `Qwen2.5-7B`, quantization consistently reduce knowledge recall, with the largest degradations under `GPTQ` and `AWQ` (Figure 11). By contrast, `bib4/8` degrades one-hop reasoning $r_1(e_1)$ and $r_2(e_2)$, as reflected by modest shifts in ENTREC and CNSTSCORE relative to the full-precision model, but leaves two-hop reasoning largely unaffected. For `Qwen2.5-14B`, `bib4/8` likewise outperforms other quantization methods, best preserving and occasionally improving FKR (Figure 12). Moreover, despite its greater layer depth than `Qwen2.5-7B`, `Qwen2.5-14B` exhibits a similar layer-wise pattern of `bib4/8` quantization effects (Figure 11e, 12e), whereas other quantization methods do not. For `Llama3-8B`, the effect of different quantization methods are alike, i.e., ENTREC and CNSTSCORE of quantized model are lower than full-precision model in shallow layers, but becomes much higher in deeper layers (Figure 13). Among the three LLMs evaluated, `Llama3-8B` is least affected by quantization in terms of FKR.

6 Conclusion

In this work, we examined the impact of quantization on factual knowledge recall. In general, quantization leads to information loss within the model, which in turn degrades factual knowledge recall; this trend becomes more pronounced for smaller models within each model family. Moreover, LLMs quantized with lower bit precision do *not* invariably perform worse than those with higher bit precision. Quantization can occasionally even improve factual knowledge recall. While quantization effects vary by model and technique, we observed no performance degradation severe enough to compromise its viability as a compression strategy.

Limitations

Our experimental work is confined to English-language datasets. Consequently, the effectiveness of our experiments in other languages may not be comparable and multilingual factual knowledge recall may be simultaneously affected by the degradation of multilingual capabilities due to quantization (Marchisio et al., 2024). Extending experiments to the multilingual settings is considered for future work.

We restrict our experiments to `Llama3-8B`, `Qwen2.5-7B` and `Qwen2.5-14B` for computational feasibility: our gradient-based analyses exceed the available GPU memory for larger models. Between the model families (Qwen, Llama, Mistral, DeepSeek), there are lots of architectural equivalences and similarities, e.g., the same attention (grouped-query attention), position embeddings (RoPE), normalization (RMSNorm) or FFN activation (SwiGLU). We argue that, based on our comprehensive experiments, our results are considerably robust and generalizable given the similar architectures compared to other model families.

In our experiments, we extensively compare full-precision models with different quantized versions in 4-bit and 8-bit formats. Lower-bit quantization, such as 1-bit or 2-bit, is not included in our study.

Moreover, the scope of our experiments is limited to post-training quantization techniques. Investigating the impact of weight-activation quantization, KV cache compression, or quantization-aware training techniques on factual knowledge recall is counted as future work.

Author Contributions

Author contributions are listed according to the CRediT taxonomy as follows:

- QW: Writing, idea conceptualization, experiments and evaluations, analysis, visualization.
- MW: Writing, experiments and evaluations, and analysis.
- NF: Writing – review & editing, experiments and evaluations, and supervision.
- SO: Writing – review & editing.
- YC: Experiments and evaluations.
- HS: Supervision.
- SM: Supervision and funding acquisition.
- VS: Funding acquisition and proof reading.

Acknowledgment

We are indebted to the anonymous reviewers of TrustNLP workshop co-located at ACL 2026 for their helpful and rigorous feedback. This work has been supported by the Federal Ministry of Research, Technology and Space (BMFTR) as part of the project VERANDA (16KIS2047).

References

AI@Meta. 2024. Llama 3 model card.

Minyeong Choe, Haehyun Cho, Changho Seo, and Hyunil Kim. 2025. Do all autoregressive transformers remember facts the same way? a cross-architecture analysis of recall mechanisms. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 28494–28513, Suzhou, China. Association for Computational Linguistics.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. 2022. Llm.int8(): 8-bit matrix multiplication for transformers at scale. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS ’22, Red Hook, NY, USA. Curran Associates Inc.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. 2023. OPTQ: Accurate quantization for generative pre-trained transformers. In *The Eleventh International Conference on Learning Representations*.

Gustavo Gonçalves and Emma Strubell. 2023. Understanding the effect of model compression on social bias in large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2663–2675, Singapore. Association for Computational Linguistics.

R.M. Gray and D.L. Neuhoff. 1998. Quantization. *IEEE Transactions on Information Theory*, 44(6):2325–2383.

Evan Hernandez, Arnab Sen Sharma, Tal Haklay, Kevin Meng, Martin Wattenberg, Jacob Andreas, Yonatan Belinkov, and David Bau. 2024. Linearity of relation decoding in transformer language models. In *The Twelfth International Conference on Learning Representations*.

Junyuan Hong, Jinhao Duan, Chenhui Zhang, Zhangheng Li, Chulin Xie, Kelsey Lieberman, James Diffenderfer, Brian R. Bartoldson, Ajay Kumar Jaiswal, Kaidi Xu, Bhavya Kailkhura, Dan Hendrycks, Dawn Song, Zhangyang Wang, and Bo Li. 2024. Decoding compressed trust: Scrutinizing the trustworthiness of efficient LLMs under compression. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 18611–18633. PMLR.

Renren Jin, Jiangcun Du, Wuwei Huang, Wei Liu, Jian Luan, Bin Wang, and Deyi Xiong. 2024. A comprehensive evaluation of quantization strategies for large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12186–12215, Bangkok, Thailand. Association for Computational Linguistics.

Elisabeth Kirsten, Ivan Habernal, Vedant Nanda, and Muhammad Bilal Zafar. 2025. The impact of inference acceleration on bias of LLMs. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1834–1853, Albuquerque, New Mexico. Association for Computational Linguistics.

Qun Li, Yuan Meng, Chen Tang, Jiacheng Jiang, and Zhi Wang. 2024. Investigating the impact of quantization on adversarial robustness. In *5th Workshop on practical ML for limited/low resource settings*.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. 2024. Awq: Activation-aware weight quantization for on-device llm compression and acceleration. In *Proceedings of Machine Learning and Systems*, volume 6, pages 87–100.

Peiyu Liu, Zikang Liu, Ze-Feng Gao, Dawei Gao, Wayne Xin Zhao, Yaliang Li, Bolin Ding, and Ji-Rong Wen. 2024. Do emergent abilities exist in quantized large language models: An empirical study. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5174–5190, Torino, Italia. ELRA and ICCL.

Kelly Marchisio, Saurabh Dash, Hongyu Chen, Dennis Aumiller, Ahmet Üstün, Sara Hooker, and Sebastian Ruder. 2024. How does quantization affect multilingual LLMs? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15928–15947, Miami, Florida, USA. Association for Computational Linguistics.

Satya Sai Srinath Namburi, Makesh Sreedhar, Srinath Srinivasan, and Frederic Sala. 2023. The cost of compression: Investigating the impact of compression on parametric knowledge in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5255–5273, Singapore. Association for Computational Linguistics.

Sein Park, Yeongsang Jang, and Eunhyeok Park. 2022. Symmetry regularization and saturating nonlinearity for robust quantization. In *Computer Vision – ECCV 2022*, pages 206–222, Cham. Springer Nature Switzerland.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2024. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Krithika Ramesh, Arnav Chavan, Shrey Pandit, and Sunayana Sitaram. 2023. A comparative study on the impact of model compression techniques on fairness in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15762–15782, Toronto, Canada. Association for Computational Linguistics.

Manpreet Singh and Hassan Sajjad. 2025. Interpreting the effects of quantization on LLMs. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 2267–2281, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Qianli Wang, Nils Feldhus, Pepa Atanasova, Fedor Splitt, Simon Ostermann, Sebastian Möller, and Vera Schmitt. 2026. Can large language models still explain themselves? investigating the impact of quantization on self-explanations. *Preprint*, arXiv:2601.00282.

Guangxuan Xiao, Ji Lin, Mickael Seznec, Hao Wu, Julien Demouth, and Song Han. 2023. Smoothquant: accurate and efficient post-training quantization for large language models. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Sohee Yang, Elena Gribovskaya, Nora Kassner, Mor Geva, and Sebastian Riedel. 2024. Do large language models latently perform multi-hop reasoning? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10210–10229, Bangkok, Thailand. Association for Computational Linguistics.

Zeping Yu and Sophia Ananiadou. 2024. Neuron-level knowledge attribution in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3267–3280, Miami, Florida, USA. Association for Computational Linguistics.

Xunyu Zhu, Jian Li, Yong Liu, Can Ma, and Weiping Wang. 2024. A survey on model compression for large language models. *Transactions of the Association for Computational Linguistics*, 12:1556–1577.

A Dataset Information

LRE An exemplary query consists of $s \leftarrow$ "*the company that created Visual Basic*" and $o \leftarrow$

TwoHop-Fact (Multi-hop Reasoning)

One-hop prompts: (p_1): The mother of Stevie Wonder is Lula. (p_2): The singer of 'Superstition' is Stevie Wonder.

Two-hop prompt: (p): The mother of the singer of 'Superstition' is

Answer: Lura

Figure 4: An example from the TwoHop-Fact dataset.

"*The current CEO of*". A correct answer (in this example: *Satya Nadella*) by the explained model is the criterion by which the LRE data is filtered.

TwoHop-Fact Figure 4 illustrates an example from the TwoHop-Fact dataset. Based on the input text, (e_1, r_1, e_2) corresponds (Superstition, singer, Stevie Wonder), while (e_2, r_2, e_3) represents (Stevie Wonder, mother, Lula).

B Models & Inference Time

Table 1 presents details of the all LLMs used in our experiments (§3.4), including model sizes, quantization approaches and corresponding URLs from the Hugging Face Hub. All models were directly obtained from the Hugging Face repository. All experiments were conducted using A100 or H100 GPUs. Neuron-level and layer-level attribution can be completed within 10 hours, while LMHR experiments take 30 hours averagely.

C Additional Experiments

C.1 Knowledge Memorization Analysis

Table 2 illustrates knowledge recall accuracy results on the LRE dataset for `Qwen2.5-{7B,14B}` and `Llama3-8B` models across different quantization methods.

Knowledge Recall Accuracy. In Table 4, and 5, we provide the per-relation factual recall accuracy on `Qwen2.5-7B`, `Qwen2.5-14B`, and `Llama3-8B` models, respectively. As discussion in Section 5.1, we observe that the accuracy drop is more severe in relations where the original performance has not yet saturated.

Neuron Attribution Analysis. Here we present the neuron-level knowledge attribution results on relations *person father*, *person mother*, *person sport position*, complementing our results in Section 5.1. Analysis on layer-wise neuron contribution scores are shown in Figure 5 and Figure 6, the analysis on top-300 neurons are given in Figure 9 and Figure 10.

C.2 Latent Multi-hop Reasoning

Table 3 shows the latent multi-hop reasoning accuracy comparison between full-precision models and quantized models. Additionally, Figure 11, Figure 12, and Figure 13 display the differences in the *entity recall score*, *consistency score*, and *accuracy* between the `AWQ`, `GPTQ8`, `GPTQ4`, `bib8`, `bib4` quantized and full-precision models of `Qwen2.5-7B`, `Qwen2.5-14B`, and `Llama3-8B`, evaluated across all layers.

Name	Citation	Size	Precision	Link
Llama3	AI@Meta (2024)	8B	full	https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct
Llama3	AI@Meta (2024)	8B	GPTQ4	https://huggingface.co/TechxGenus/Meta-Llama-3-8B-Instruct-GPTQ
Llama3	AI@Meta (2024)	8B	AWQ	https://huggingface.co/TechxGenus/Meta-Llama-3-8B-Instruct-AWQ
Qwen2.5	Qwen et al. (2024)	7B	Full	https://huggingface.co/Qwen/Qwen2.5-7B-Instruct
Qwen2.5	Qwen et al. (2024)	7B	AWQ	https://huggingface.co/Qwen/Qwen2.5-7B-Instruct-AWQ
Qwen2.5	Qwen et al. (2024)	7B	GPTQ4	https://huggingface.co/Qwen/Qwen2.5-7B-Instruct-GPTQ-Int4
Qwen2.5	Qwen et al. (2024)	7B	GPTQ8	https://huggingface.co/Qwen/Qwen2.5-7B-Instruct-GPTQ-Int8
Qwen2.5	Qwen et al. (2024)	14B	Full	https://huggingface.co/Qwen/Qwen2.5-14B-Instruct
Qwen2.5	Qwen et al. (2024)	14B	AWQ	https://huggingface.co/Qwen/Qwen2.5-14B-Instruct-AWQ
Qwen2.5	Qwen et al. (2024)	14B	GPTQ4	https://huggingface.co/Qwen/Qwen2.5-14B-Instruct-GPTQ-Int4
Qwen2.5	Qwen et al. (2024)	14B	GPTQ8	https://huggingface.co/Qwen/Qwen2.5-14B-Instruct-GPTQ-Int8

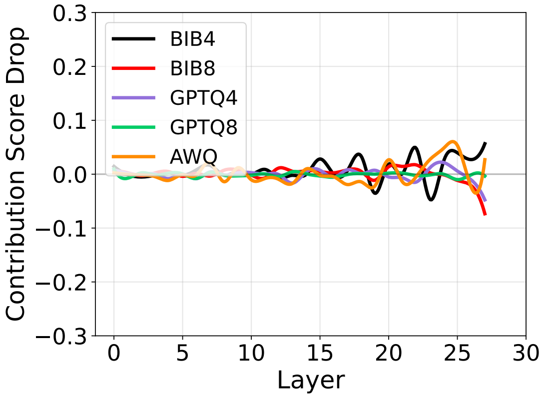
Table 1: Detailed information about used LLMs in our experiments.

Model	Method	Accuracy % ↑	# Correct
Qwen2.5-7B	full	**63.25**	**6133**
	bib4	60.72	5887
	bib8	63.01	6109
	gptq4	60.10	5827
	gptq8	63.22	6130
	awq	60.60	5876
Qwen2.5-14B	full	**73.08**	**7086**
	bib4	70.33	6819
	bib8	73.06	7084
	gptq4	25.20	2443
	gptq8	73.03	7081
	awq	70.61	6846
Llama3-8B	full	**77.62**	**7526**
	bib4	72.19	7000
	bib8	76.95	7461
	gptq8	71.39	6922
	awq	71.83	6965

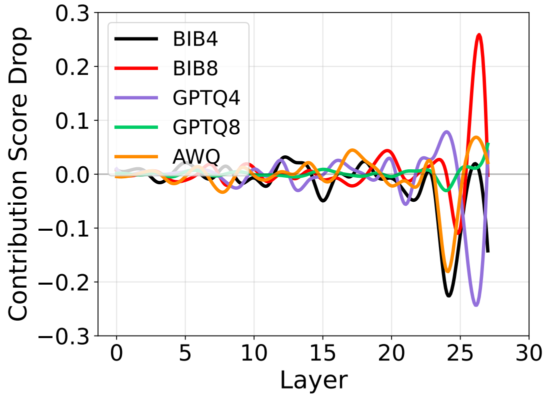
Table 2: Knowledge recall accuracy results (%) and number of correct predictions (out of 9696 queries) on the LRE dataset for Qwen2.5-{7B,14B} and Llama3-8B models across different quantization methods.

Model	Method	$r_1(e_1)$ ↑	$r_2(e_2)$ ↑	$r_2(r_1(e_1))$ ↑
Qwen2.5-7B	full	**25.03**	39.07	**20.61**
	bib4	25.01	39.02	20.61
	bib8	25.01	39.02	20.61
	gptq4	22.29	37.44	19.59
	gptq8	24.76	**39.15**	20.55
	awq	22.07	38.89	18.05
Qwen2.5-14B	full	**35.23**	40.45	24.72
	bib4	35.16	40.56	24.76
	bib8	35.16	40.56	**24.76**
	gptq4	32.22	**45.19**	23.29
	gptq8	35.16	39.55	24.61
	awq	24.69	35.61	21.80
Llama3-8B	full	7.79	21.39	4.45
	bib4	4.94	27.44	3.14
	bib8	5.79	18.27	2.90
	gptq8	**23.62**	**40.73**	**20.94**
	awq	22.35	37.79	19.56

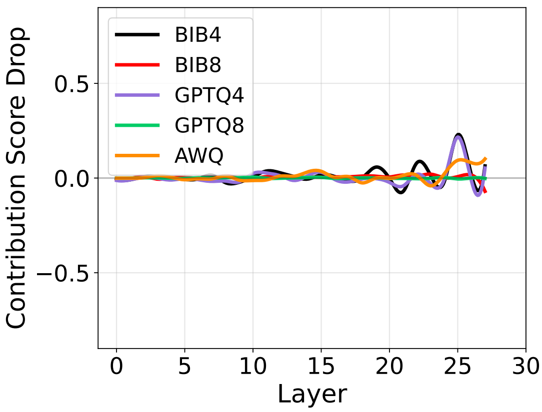
Table 3: Accuracy (in %) of different models in connecting and traversing implicit knowledge to successfully answer latent multi-hop queries on TwoHop-Fact.



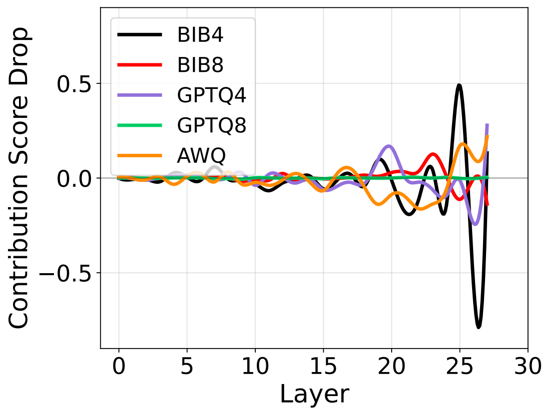
(a) Attention: Person father



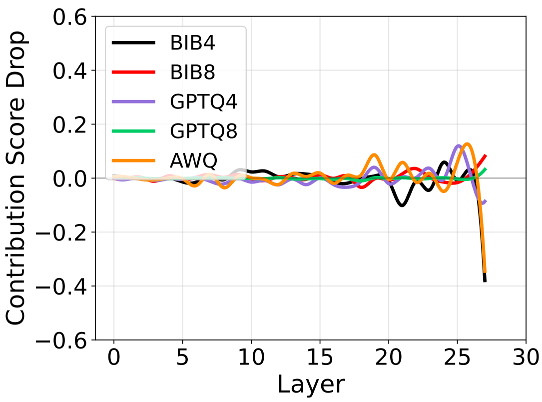
(b) FFN: Person father



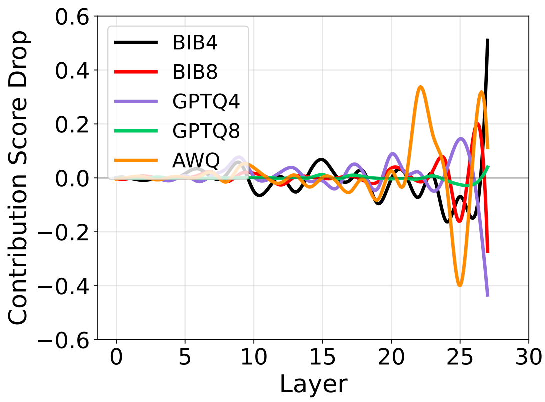
(c) Attention: Person mother



(d) FFN: Person mother

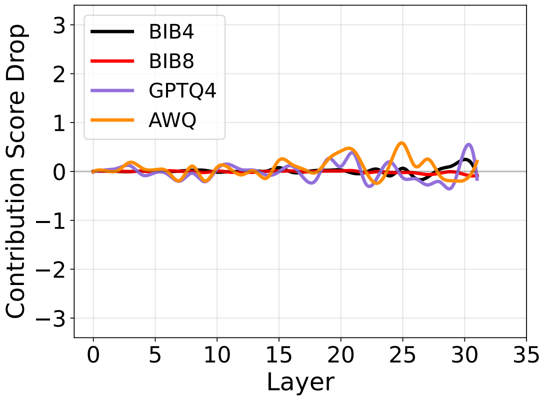


(e) Attention: Person sport position

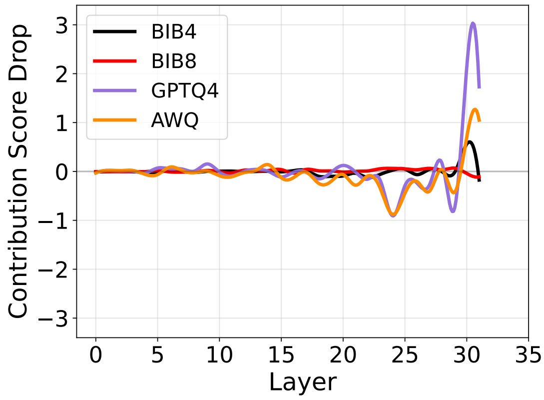


(f) FFN: Person sport position

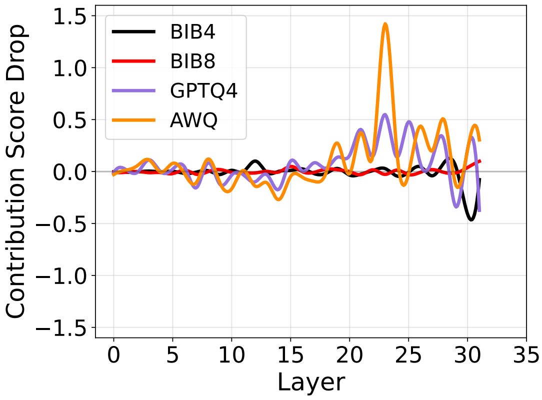
Figure 5: Contribution score drop across quantization methods on `Qwen2.5-7B` for other relationship types, comparing attention sublayers (left) and feed-forward sublayers (right). Rows show different relationships: (a-b) person father, (c-d) person mother, and (e-f) person sport position. The plots reveal how different quantization methods affect knowledge representation across model layers.



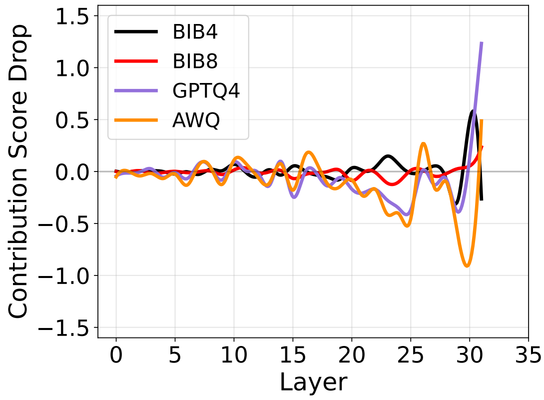
(a) Attention: Person father



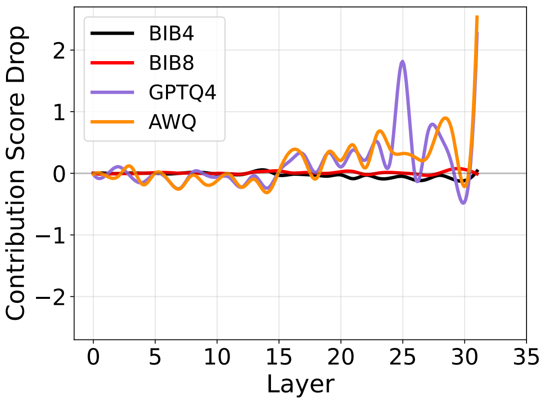
(b) FFN: Person father



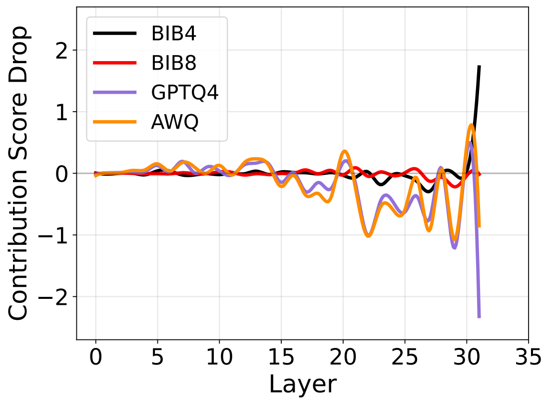
(c) Attention: Person mother



(d) FFN: Person mother

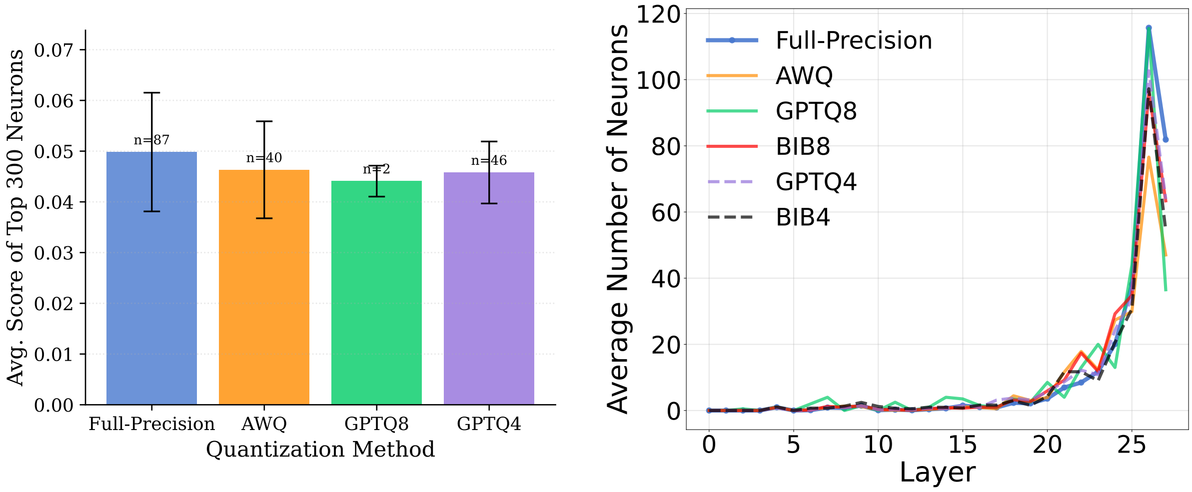


(e) Attention: Person sport position



(f) FFN: Person sport position

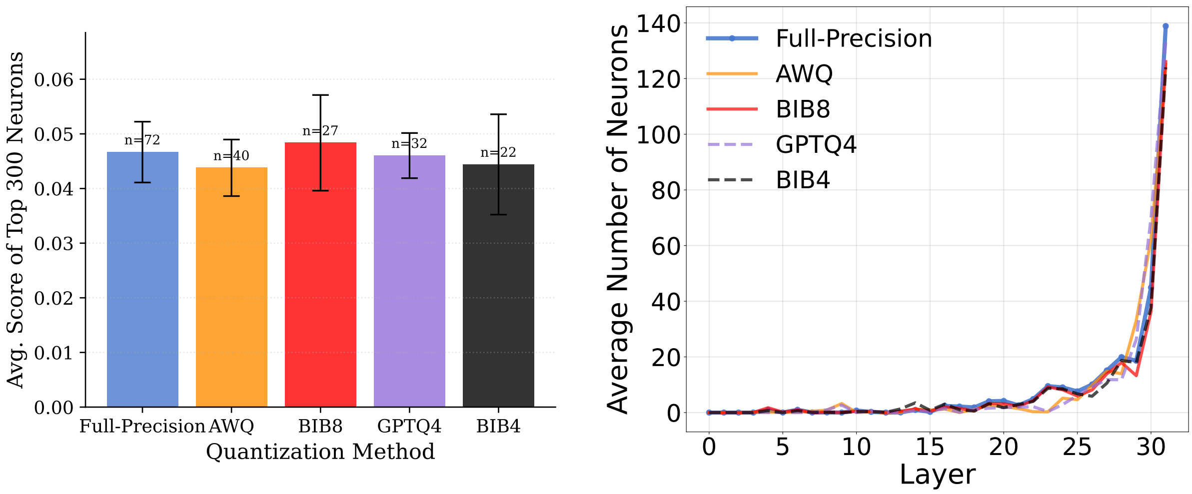
Figure 6: Contribution score drop across quantization methods on `Llama3-8B` for other relationship types, comparing attention sublayers (left) and feed-forward sublayers (right). Rows show different relationships: (a-b) person father, (c-d) person mother, and (e-f) person sport position. The plots reveal how different quantization methods affect knowledge representation across model layers.



(a) Contribution score comparison

(b) Layer distribution

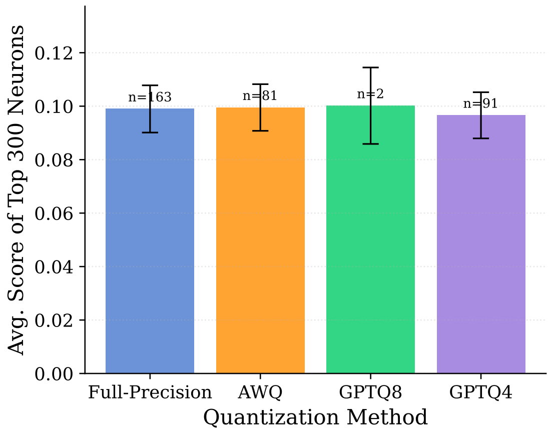
Figure 7: Analysis of the top-300 neurons with highest contribution scores for the *landmark on continent* relation under different quantization methods applied to `Qwen2.5-7B`. (a) Average contribution scores of top 300 feed-forward neurons across different quantization methods, showing how each method affects neuron activation patterns. (b) Distribution of high-scoring neurons across layers, showing the number of neurons exceeding the full-precision model's 300th neuron score threshold in each layer.



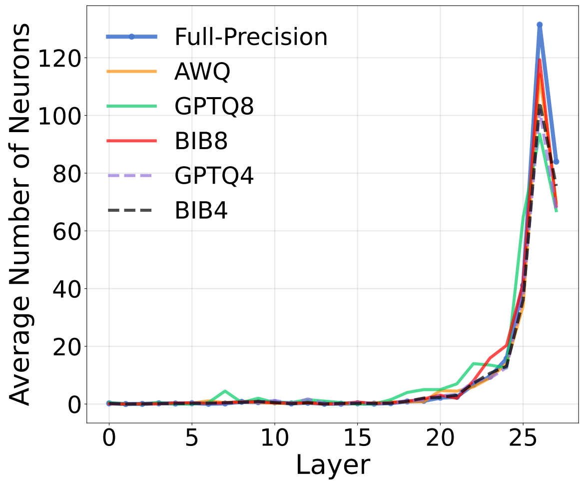
(a) Contribution score comparison

(b) Layer distribution

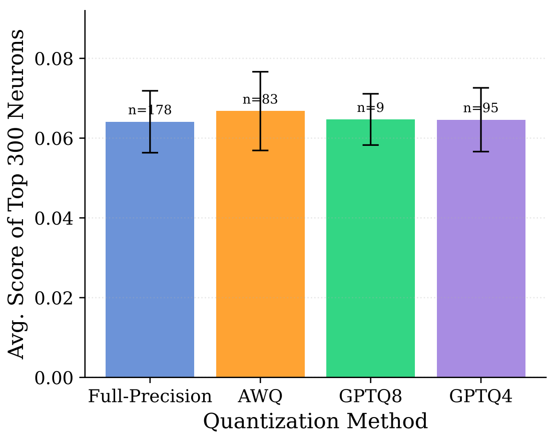
Figure 8: Analysis of the top-300 neurons with highest contribution scores for the *landmark on continent* relation under different quantization methods applied to `Llama3-8B`. (a) Average contribution scores of top 300 feed-forward neurons across different quantization methods, showing how each method affects neuron activation patterns. (b) Distribution of high-scoring neurons across layers, showing the number of neurons exceeding the full-precision model's 300th neuron score threshold in each layer.



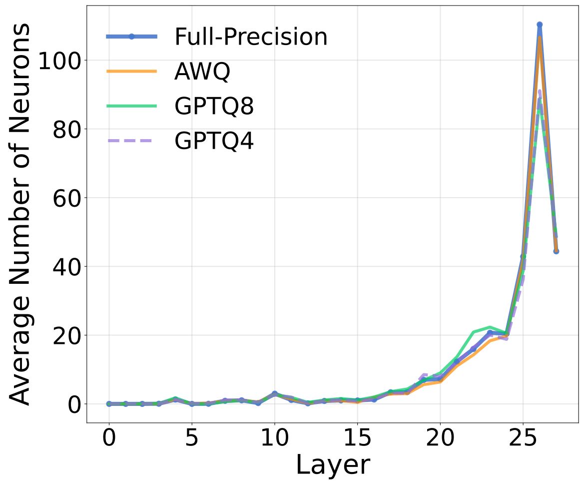
(a) Person father: Contribution scores



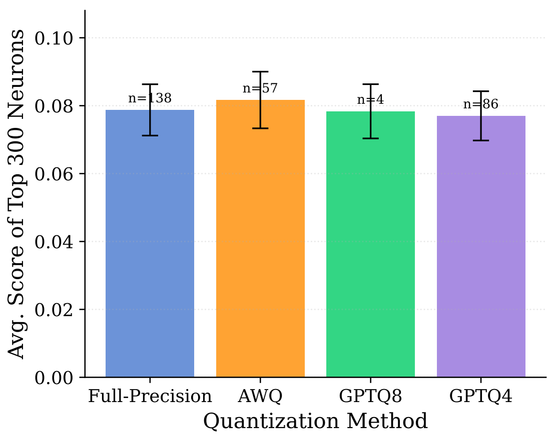
(b) Person father: Layer distribution



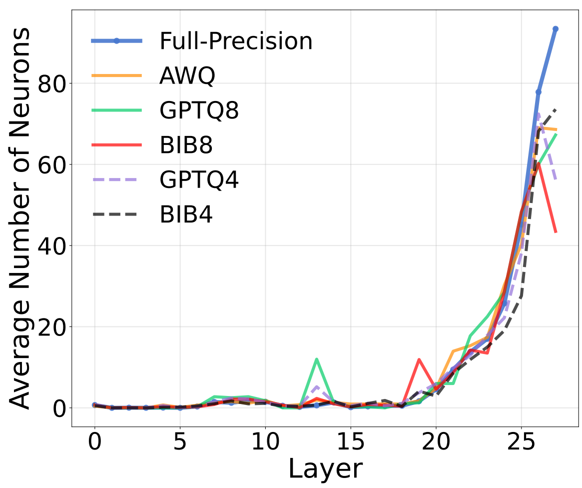
(c) Person mother: Contribution scores



(d) Person mother: Layer distribution

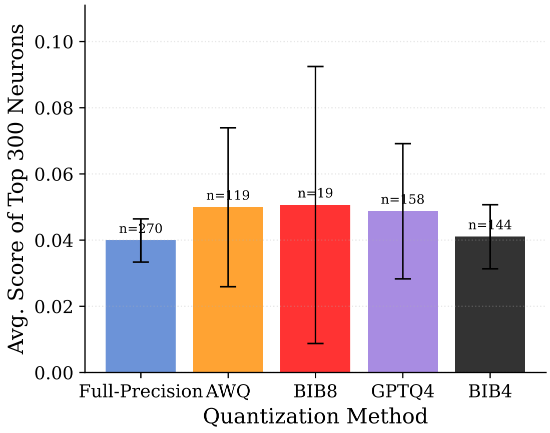


(e) Person sport position: Contribution scores

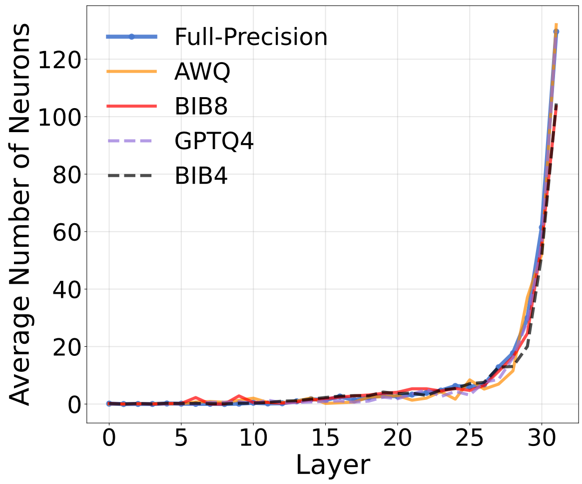


(f) Person sport position: Layer distribution

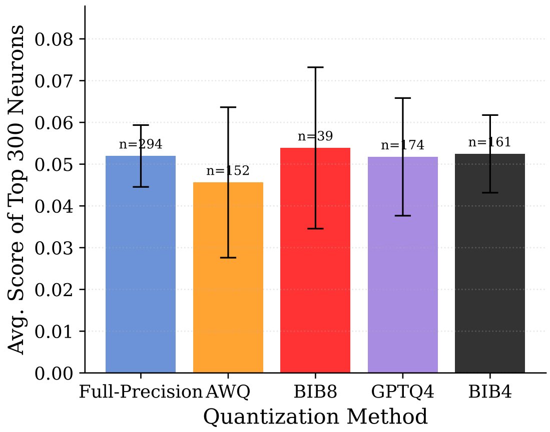
Figure 9: Analysis of other relationship types under different quantization methods applied to `Qwen2.5-7B`. Left column: Average contribution scores of top 300 feed-forward neurons across quantization methods. Right column: Distribution of high-scoring neurons across model layers. Each row represents a different relationship: (a-b) person father, (c-d) person mother, and (e-f) person sport position. These visualizations reveal both the magnitude of contribution score changes and their distribution across the model architecture when applying different quantization techniques to various types of factual knowledge.



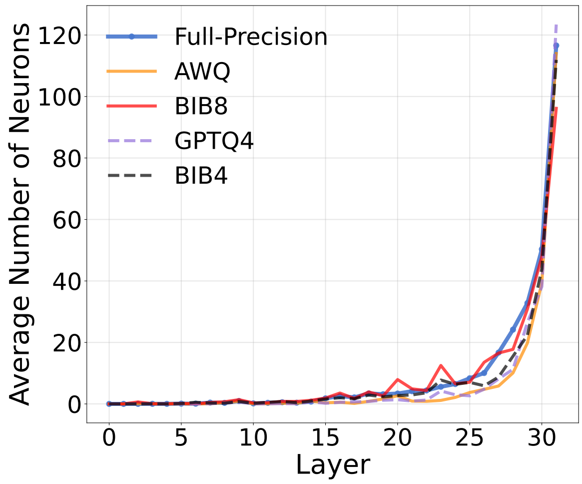
(a) Person father: Contribution scores



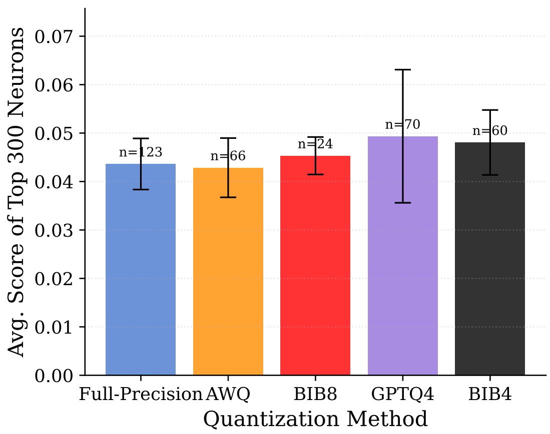
(b) Person father: Layer distribution



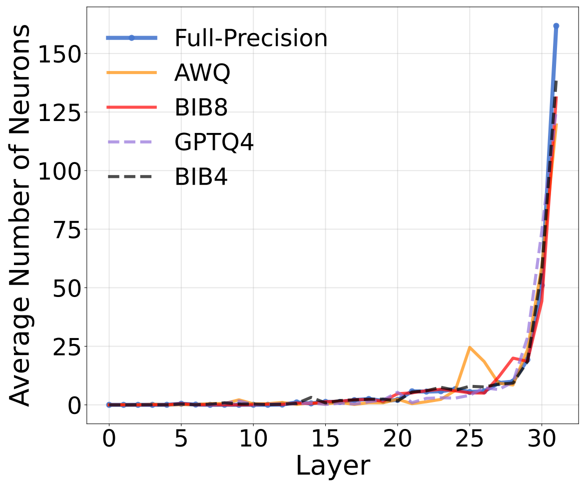
(c) Person mother: Contribution scores



(d) Person mother: Layer distribution

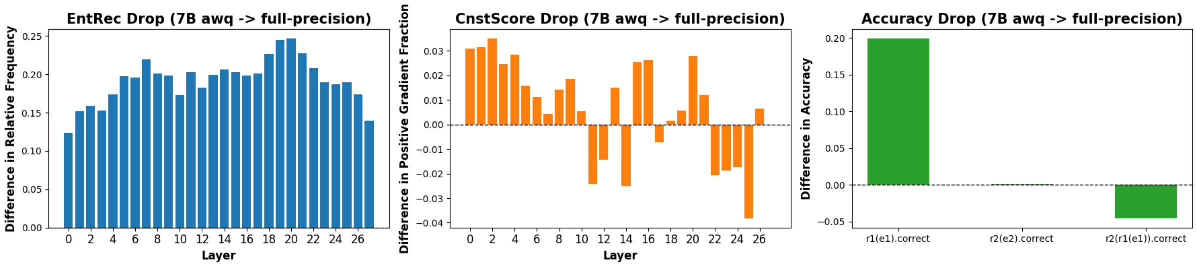


(e) Person sport position: Contribution scores

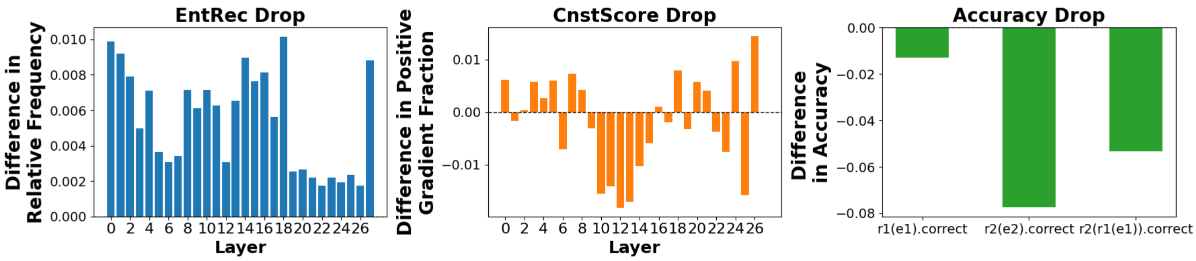


(f) Person sport position: Layer distribution

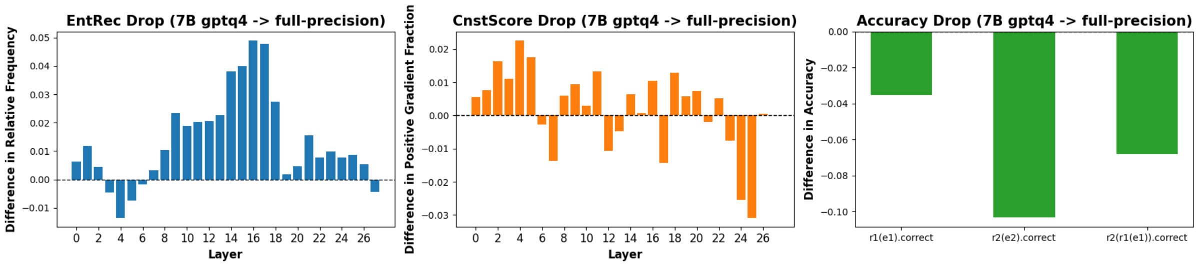
Figure 10: Analysis of other relationship types under different quantization methods applied to `Llama3-8B`. Left column: Average contribution scores of top 300 feed-forward neurons across quantization methods. Right column: Distribution of high-scoring neurons across model layers. Each row represents a different relationship: (a-b) person father, (c-d) person mother, and (e-f) person sport position. These visualizations reveal both the magnitude of contribution score changes and their distribution across the model architecture when applying different quantization techniques to various types of factual knowledge.



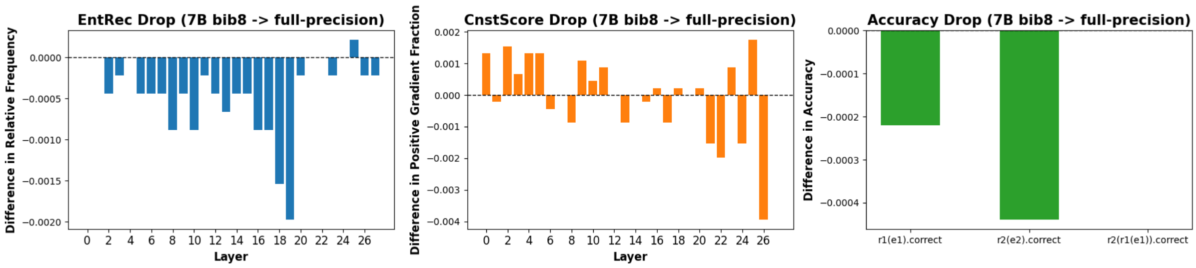
(a) AWQ



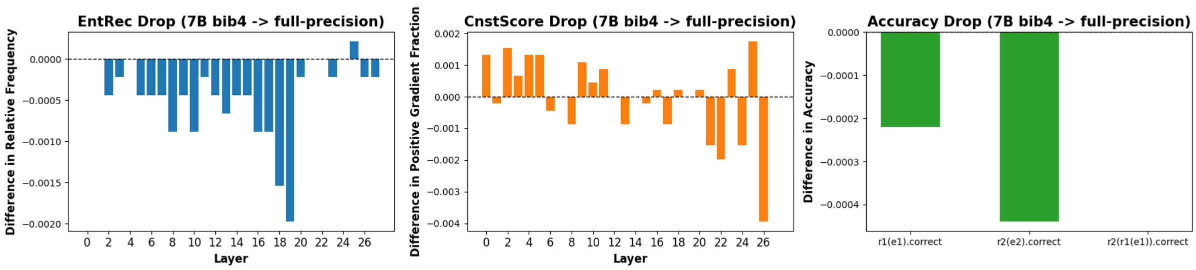
(b) GPTQ8



(c) GPTQ4

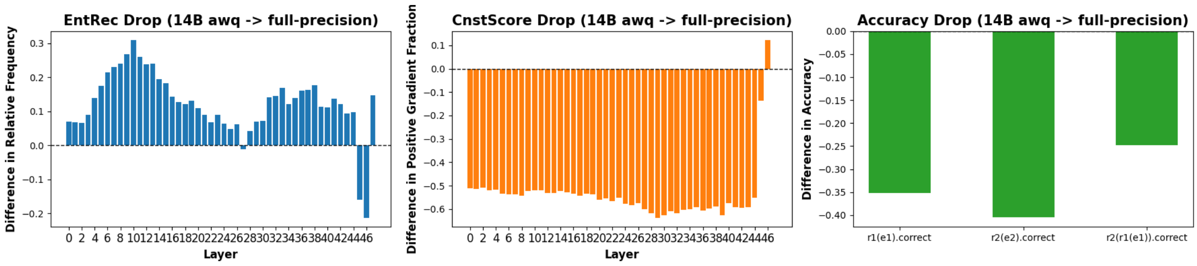


(d) bib8

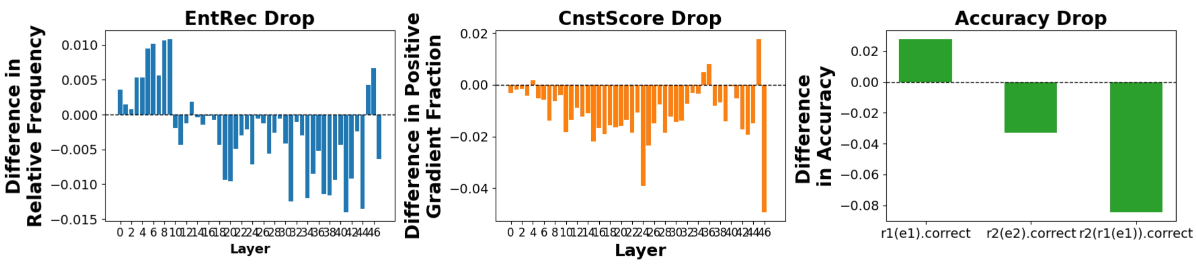


(e) bib4

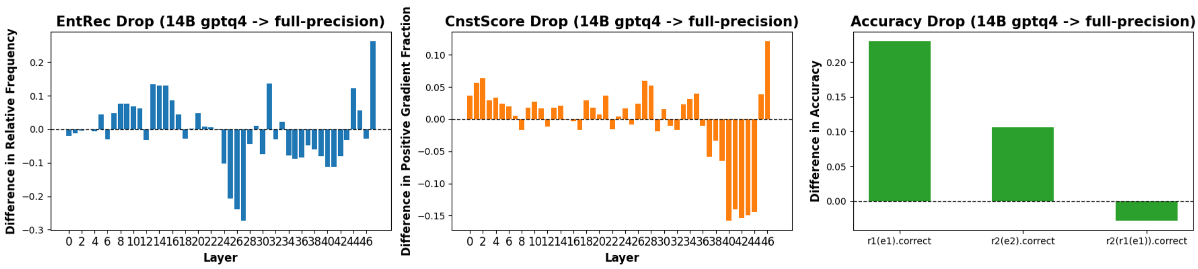
Figure 11: Difference in the *entity recall score* (ENTREC), *consistency score* (CNSTSCORE), and *accuracy* between the AWQ, GPTQ8, GPTQ4, bib8, bib4 quantized and full-precision models of Qwen2.5-7B, evaluated across all layers.



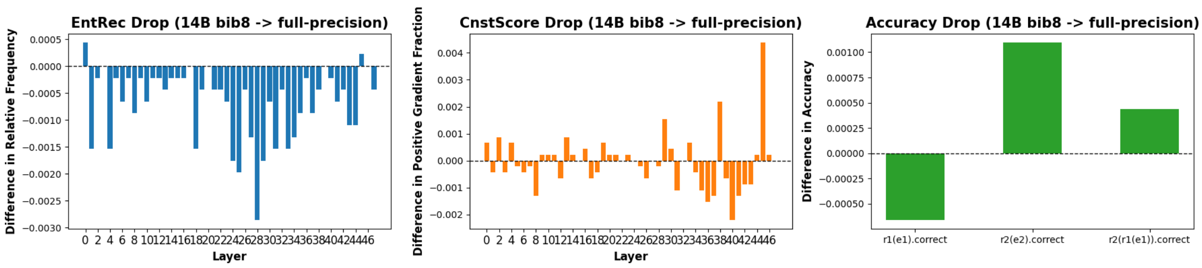
(a) AWQ



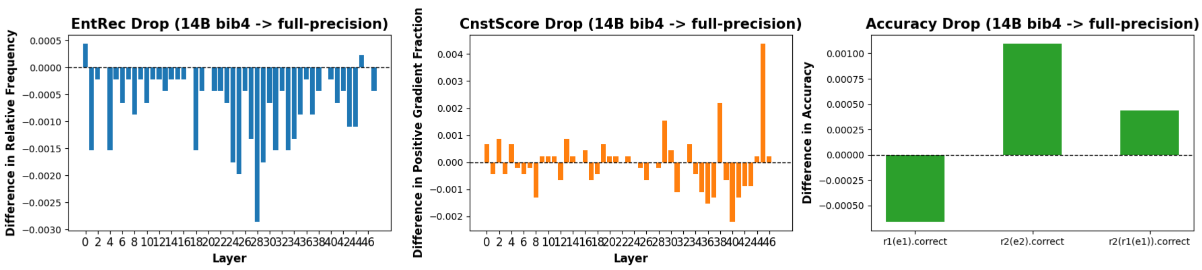
(b) GPTQ8



(c) GPTQ4

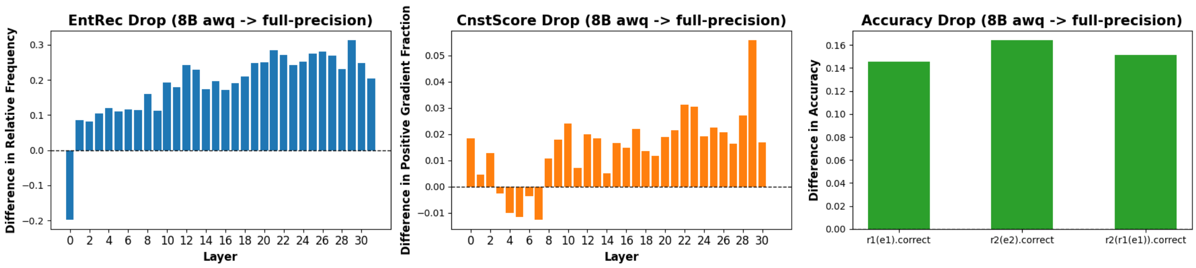


(d) bib8

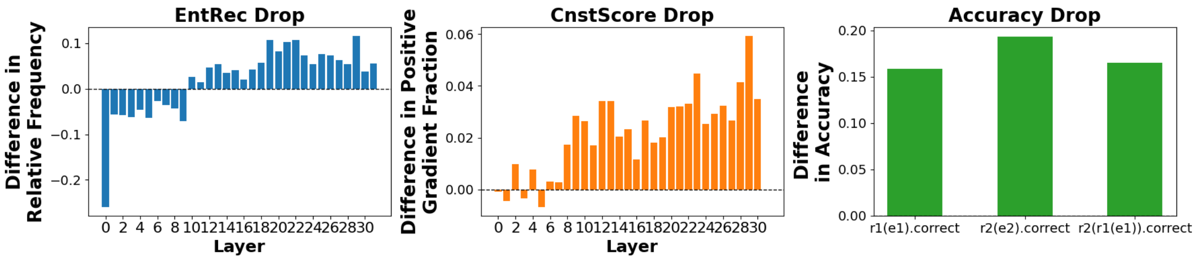


(e) bib4

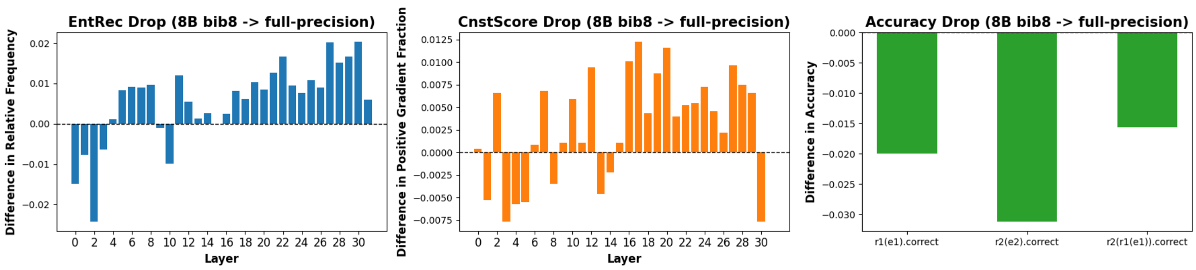
Figure 12: Difference in the *entity recall score* (EntRec), *consistency score* (CnstScore), and *accuracy* between the AWQ, GPTQ8, GPTQ4, bib8, bib4 quantized and full-precision models of Qwen2.5-14B, evaluated across all layers.



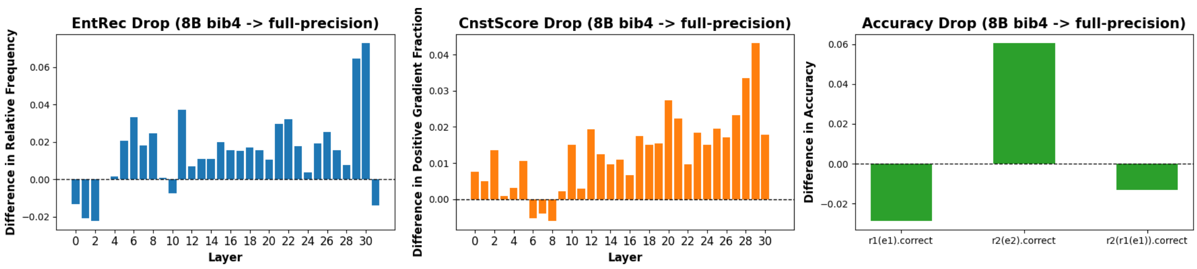
(a) AWQ



(b) GPTQ



(c) bib8



(d) bib4

Figure 13: Difference in the *entity recall score* (ENTREC), *consistency score* (CNSTSCORE), and *accuracy* between the AWQ, GPTQ, bib8, bib4 quantized and full-precision models of Llama3-8B, evaluated across all layers.

Qwen2.5-7B	full	bib4	bib8	gptq4	gptq8	awq
city in country	93.00	85.00	93.00	93.00	93.00	93.00
company CEO	49.00	39.00	50.00	45.00	50.00	46.00
company hq	59.00	54.00	59.00	56.00	58.00	59.00
country capital city	100.00	100.00	100.00	100.00	100.00	100.00
country currency	93.00	90.00	93.00	87.00	93.00	97.00
country language	92.00	88.00	92.00	88.00	92.00	92.00
country largest city	100.00	100.00	100.00	100.00	100.00	100.00
food from country	83.00	83.00	83.00	80.00	83.00	80.00
landmark in country	81.00	80.00	81.00	79.00	81.00	80.00
landmark on continent	87.00	84.00	85.00	83.00	87.00	85.00
person father	45.00	39.00	45.00	40.00	45.00	41.00
person lead singer of band	100.00	100.00	100.00	100.00	100.00	100.00
person mother	39.00	32.00	39.00	34.00	38.00	33.00
person native language	90.00	91.00	90.00	88.00	90.00	86.00
person occupation	33.00	34.00	33.00	31.00	33.00	30.00
person plays instrument	44.00	46.00	43.00	42.00	44.00	43.00
person sport position	66.00	63.00	66.00	59.00	66.00	62.00
person university	47.00	46.00	46.00	48.00	46.00	47.00
plays pro sport	79.00	77.00	81.00	75.00	79.00	76.00
pokemon evolution	100.00	89.00	100.00	91.00	100.00	95.00
president birth year	0.00	0.00	0.00	0.00	0.00	0.00
president election year	0.00	0.00	0.00	0.00	0.00	0.00
product by company	80.00	77.00	79.00	77.00	80.00	79.00
star constellation name	85.00	85.00	85.00	85.00	85.00	84.00
superhero archnemesis	34.00	33.00	33.00	33.00	34.00	31.00
superhero person	49.00	50.00	50.00	49.00	50.00	48.00
AVG	63.25	60.72	63.01	60.10	63.22	60.00

Table 4: Per-relation knowledge recall accuracy results (%) on the LRE dataset for `Qwen2.5-7B` across different quantization methods.

Qwen2.5-14B	full	bib4	bib8	gptq4	gptq8	awq
city in country	96.00	96.00	96.00	81.00	96.00	96.00
company CEO	58.00	54.00	59.00	2.00	58.00	57.00
company hq	68.00	65.00	66.00	10.00	69.00	65.00
country capital city	100.00	100.00	100.00	79.00	100.00	100.00
country currency	90.00	83.00	87.00	77.00	87.00	90.00
country language	96.00	96.00	96.00	92.00	96.00	96.00
country largest city	96.00	100.00	96.00	83.00	96.00	96.00
food from country	80.00	77.00	80.00	50.00	80.00	80.00
landmark in country	87.00	86.00	87.00	31.00	87.00	87.00
landmark on continent	85.00	88.00	84.00	48.00	85.00	84.00
person father	63.00	56.00	62.00	7.00	63.00	57.00
person lead singer of band	100.00	100.00	100.00	29.00	100.00	100.00
person mother	59.00	49.00	59.00	1.00	59.00	53.00
person native language	94.00	94.00	94.00	39.00	93.00	93.00
person occupation	47.00	45.00	49.00	12.00	49.00	45.00
person plays instrument	67.00	61.00	68.00	52.00	66.00	63.00
person sport position	76.00	74.00	76.00	5.00	74.00	71.00
person university	48.00	45.00	52.00	43.00	47.00	46.00
plays pro sport	83.00	83.00	81.00	48.00	82.00	85.00
pokemon evolution	100.00	100.00	100.00	30.00	100.00	100.00
president birth year	0.00	0.00	0.00	0.00	0.00	0.00
president election year	0.00	0.00	0.00	0.00	0.00	0.00
product by company	85.00	82.00	85.00	38.00	85.00	84.00
star constellation name	87.00	87.00	87.00	75.00	87.00	87.00
superhero archnemesis	48.00	43.00	46.00	4.00	48.00	44.00
superhero person	76.00	73.00	75.00	5.00	77.00	72.00
AVG	73.08	70.33	73.06	25.20	73.03	70.61

Table 5: Per-relation knowledge recall accuracy results (%) on the LRE dataset for Qwen2.5-14B across different quantization methods.