Can Large Language Models Still Explain Themselves? Investigating the Impact of Quantization on Self-Explanations

Qianli Wang[1,5] **Nils Feldhus**[2,5,*,†] **Pepa Atanasova**[3,*] **Fedor Splitt**[1]
Simon Ostermann[4,5,6] **Sebastian Möller**[1,5] **Vera Schmitt**[1]
[1]Technische Universität Berlin [2]University of Groningen [3]University of Copenhagen
[4]Saarland Informatics Campus [5]German Research Center for Artificial Intelligence (DFKI)
[6]Centre for European Research in Trusted AI (CERTAIN)
Correspondence: qianli.wang@tu-berlin.de

Abstract

Quantization is widely used to accelerate inference and streamline the deployment of large language models (LLMs), yet its effects on self-explanations (SEs) remain unexplored. SEs, generated by LLMs to justify their own outputs, require reasoning about the model's own decision-making process, a capability that may exhibit particular sensitivity to quantization. As SEs are increasingly relied upon for transparency in high-stakes applications, understanding whether and to what extent quantization degrades SE quality and faithfulness is critical. To address this gap, we examine two types of SEs: *natural language explanations* (NLEs) and *counterfactual examples*, generated by LLMs quantized using three common techniques at distinct bit widths. Our findings indicate that quantization typically leads to moderate declines in both SE quality (up to 4.4%) and faithfulness (up to 3.9%). The user study further demonstrates that quantization considerably diminishes both the coherence and trustworthiness of SEs (by up to 8.5%). Compared to smaller models, larger models show limited resilience to quantization in terms of SE quality but maintain more faithfulness. Moreover, no quantization technique consistently excels across task accuracy, SE quality, and faithfulness. Because quantization's impact varies considerably by context and can be sizable in specific cases, we recommend validating SE quality for the intended use case. Despite these sometimes considerable drops, quantization remains an effective compression technique when its impact on SEs is properly validated.

1 Introduction

Deploying LLMs efficiently at scale has motivated extensive research on quantization (Dettmers et al., 2022; Frantar et al., 2023; Lin et al., 2024; Wang et al., 2026b). Quantization achieves model compression and efficient deployment (e.g., of on-device LLMs) by reducing parameter precision and bit allocation, delivering substantial size reductions while preserving most functionality (Gray and Neuhoff, 1998). Previous work has investigated quantization's influence on various model dimensions, such as multilinguality (Marchisio et al., 2024), bias (Gonçalves and Strubell, 2023), and alignment (Jin et al., 2024). An important capability dimension that may be affected by quantization is the capability of a model to explain itself. Self-explanations (SEs) are statements generated by models to justify their own decisions (Agarwal et al., 2024; Madsen et al., 2024; Wang et al., 2026c), which are deemed to be an effective and convincing way to deliver explanations to users and enhance the transparency of black-box LLMs (Huang et al., 2023; Randl et al., 2025). Nevertheless, SE may obfuscate the true reasoning process of LLMs (Turpin et al., 2023; Tutek et al., 2025; Sun et al., 2026), and we hypothesize that quantization may exacerbate this, since LLMs are directly optimized for task performance but learn to generate faithful SEs more indirectly. Moreover, quantized models have been widely adopted in prior work for many types of SE generation (Wang et al., 2024; Liu et al., 2024; Bhattacharjee et al., 2024; Giorgi et al., 2025). However, the impact of quantization on SEs, specifically on whether SEs remain faithful to a model's inner workings and whether their quality can be largely preserved, remains unexplored and has yet to be comprehensively characterized.

We bridge this gap through a comprehensive study on how quantization affects both the *quality* and *faithfulness* of SEs. Our study encompasses two distinct types of free-text SEs: *natural language explanations* and *counterfactual examples* (Figure 1). **First**, we perform comprehensive automatic evaluations of SE quality across five datasets

[*]Equal contribution and shared second authorship.
[†]Work done at TU Berlin.

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 780–809
October 24-29, 2026 ©2026 Association for Computational Linguistics

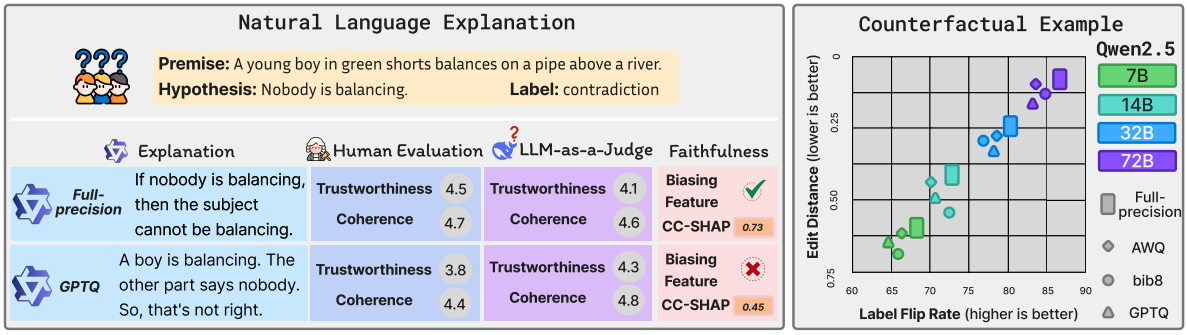


Figure 1: Example effects of quantization on two types of self-explanations, *natural language explanations* (**left**) and *counterfactual examples* (**right**), across different models and quantization methods. Natural language explanations are rated by ① human annotators in terms of perceived trustworthiness and coherence (§5.2); ② LLM-as-a-Judge evaluation (§5.3); ③ faithfulness test by Biasing Feature (Turpin et al., 2023) and CC-SHAP (Parcalabescu and Frank, 2024) (§5.1.1), while counterfactual examples are evaluated on edit distance and label flip rate (§5.1.2).

and six models of varying sizes under full precision and three quantization approaches with different bit widths (§5.1). We show that different types of SEs exhibit varying levels of sensitivity to quantization, though *quantization generally leads to moderate degradation in SE quality* (up to 4.4%). Unexpectedly, larger quantized models do not always outperform smaller full-precision models in generating high-quality SEs, nor do LLMs with lower bit precision consistently lag behind their higher bit-precision counterparts.

Second, we assess SE faithfulness through self-consistency checks for counterfactuals and two metrics for natural language explanations. SE faithfulness exhibits *modest average decline under quantization* (up to 3.9%); however, larger models demonstrate greater robustness in preserving SE faithfulness (§5.1.1). Across our experiments, no quantization method consistently excels across *task performance*, *explanation quality*, and *faithfulness* simultaneously. Furthermore, we observe a distinct trade-off between SE characteristics, namely quality and faithfulness, and overall task performance under quantization. Therefore, SE characteristics should be validated for specific use cases depending on whether task performance or explanation performance is prioritized (§5.3).

Lastly, we conduct a user study with 48 participants, evaluating trustworthiness and coherence of SEs generated by models at different bit-precision levels. Human evaluators perceive *full-precision models as producing more trustworthy and coherent SEs than those produced by their quantized counterparts* (Figure 1). Notably, conducting a similar LLM-as-a-Judge evaluation fails to fully capture the impact of quantization on self-explanation quality, evidenced by the weak or negative, and non-statistically significant correlation observed between human and judge model ratings (§5.2).

In conclusion, the modest reductions in SE quality and faithfulness do not diminish quantization's value as an effective model compression strategy. Nevertheless, for high-stakes scenarios requiring optimal explanation reliability, we recommend application-specific validation before deployment.

2 Preliminaries and Related Work

Quantization. Quantization has been demonstrated to offer a superior balance between compression and accuracy (Li et al., 2024c). The decoding stage during LLM inference is typically memory-bound, where the key-value cache (KV cache) overhead often exceeds the size of the model weights (Li et al., 2024c). Quantization techniques compress LLMs by converting model weights, activations, or the KV cache, originally in 32-bit floating-point format, into lower-precision data types (Zhu et al., 2024), e.g., 8-bit integer (Dettmers et al., 2022). 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 an ad-hoc quantization during inference without necessitating modifications to the model architecture or training process. Among PTQ, weight-only quantization is the most conventional and widely adopted method (Wan et al., 2024; Zhou et al., 2024), which effectively accelerates matrix multiplications during the decoding stage (Li et al., 2024c). Thereby, in this paper, we evaluate the impact of weight-only PTQ

quantization (§3.3) on self-explanations (§4).

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. (2024) 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. In our work, we explicitly explore the impact of quantization on self-explanations.

Self-Explanations. SEs are generated by LLMs to justify their own decisions. Prior work has identified several SE types, including prompting-based feature attribution explanations (Huang et al., 2023), counterfactual explanations (Wang et al., 2025b), redaction explanations (Madsen et al., 2024; Doi et al., 2025), and natural language explanations (Villa-Arenas et al., 2024). Previous research has inspected various aspects of SEs: Huang et al. (2023) compare prompting-based feature attribution with perturbation-based feature importance. Madsen et al. (2024) explore the faithfulness of SEs via self-consistency. Randl et al. (2025) examine how SEs correlate with human judgments and internal model dynamics. We extend this line of research by investigating the impact of quantization on SE quality and faithfulness.

3 Experimental Setup

We evaluate the impact of quantization on SE by examining two representative **free-text** explanation types (§3.1). Specifically, we compare full-precision LLMs (§3.4) with their quantized counterparts employing different quantization techniques and bit-widths (§3.3) across five datasets (§3.2).

3.1 Self-Explanations

The experimental investigation focuses on two well-established types of SEs in the explainability literature (Madsen et al., 2024; Agarwal et al., 2024; Villa-Arenas et al., 2024; Wang et al., 2025a; Monteiro Paes et al., 2025): *natural language explanations* and *counterfactual examples* (Figure 1).

Natural Language Explanations (NLEs) are free-text explanations of predictions made by the model and can be easily understood by humans (Camburu et al., 2018; Wiegreffe et al., 2021). To minimize the confounding effects of quantization that also affect in-context learning capabilities (Liu et al., 2024), we generate NLEs in a zero-shot setting using `ZeroCoT` (Kojima et al., 2022) which elicits step-by-step reasoning from LLMs by simply adding "Let's think step by step" before each answer, without requiring any hand-crafted few-shot examples.

Counterfactual Examples (CFEs) refer to minimally edited inputs that result in different model predictions (Miller, 2019; Madsen et al., 2022; Wang et al., 2026a,d), which can be used to understand the black-box nature of models in a contrastive manner (Wu et al., 2021; Nguyen et al., 2024b). Analogous to NLEs, we generate CFEs in a zero-shot setting using `FIZLE` (Bhattacharjee et al., 2024). `FIZLE` employs a two-stage process: it begins with prompting LLMs to extract salient keywords from the input, which are then employed to guide the generation of counterfactual examples.

3.2 Datasets

Our study employs five widely recognized datasets[1] to evaluate SEs (§3.1). eSNLI (Camburu et al., 2018) categorizes premise-hypothesis relationships as *entailment*, *contradiction*, or *neutrality*, with human-annotated NLEs. HealthFC (Vladika et al., 2024) is a fact-checking dataset with questions, documents, veracity labels (*true*, *false*, *unknown*), and human explanations. ECQA (Aggarwal et al., 2021) extends CommonsenseQA (Talmor et al., 2019) with human-annotated explanations for multiple-choice answers. AG News (Zhang et al., 2015) classifies news articles into four categories: *World*, *Sports*, *Business*, and *Sci/Tech*. IMDb (Maas et al., 2011) is a binary sentiment classification benchmark of movie reviews.

3.3 Quantization Techniques

Building on the prior discussion (§2), we identify three commonly used PTQ techniques applied to the selected LLMs (Appendix B, Table 3):

- `GPTQ` (Frantar et al., 2023) uses a second-order, Hessian-based optimization to quantize weights post-training with minimal accuracy loss;

[1] Dataset examples and label distributions are detailed in Appendix A. eSNLI is used for both SE types, AG News and IMDb for CFEs, and HealthFC and ECQA for NLEs.

- 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[2] (bib4 and bib8) enables fast and memory-efficient inference by using optimized low-bit kernels.

3.4 Models

We employ six open-source LLMs spanning model sizes from 7B to 72B, drawn from two families: Llama3 (8B, 70B) (AI@Meta, 2024) and Qwen2.5 (7B, 14B, 32B, 72B) (Qwen et al., 2024) (Table 1; Appendix C), across all self-explanations. These models are selected because their corresponding quantized versions are (officially) provided. Additionally, for all experiments, we set the temperature=0 during generation to minimize the stochastic nature of LLM outputs, ensuring the reproducibility of our results.

4 Evaluation

We assess the impact of quantization on self-explanations from three perspectives: explanation quality, evaluated through ① automatic evaluation (§4.1), ② human evaluation (§4.3), and ③ explanation faithfulness (§4.2).

4.1 Self-Explanation Quality Evaluation

We assess the self-explanation quality using automatic metrics evaluating NLE plausibility (resemblance to human-annotated explanations; §4.1.1) and CFE performance across validity, fluency, and textual similarity (§4.1.2). All results are averaged over three runs with different seeds (§5.1).

4.1.1 Natural Language Explanation

Following Marasovic et al. (2022); Wang et al. (2025a); Hsu et al. (2025), we employ two automatic metrics:

BARTScore (Yuan et al., 2021) is a reference-based metric that employs BART (Lewis et al., 2020) to evaluate generated explanations based on how well they align with human-annotated references. Furthermore, BARTScore performs bidirectional evaluation, assessing both "generated-to-reference" and "reference-to-generated" directions, thereby offering a more robust assessment. BARTScore measures NLE plausibility based on textual similarity between human NLEs and LLM-generated NLEs.

TIGERScore (Jiang et al., 2024), in contrast, is a reference-free metric that deploys a fine-tuned Llama2 (Touvron et al., 2023) model to identify errors in the generated explanations in terms of, e.g., coherence, informativeness, and accuracy. For each mistake, TIGERScore assigns a penalty score between $[-5, -0.5]$. High-quality explanations that contain no detected errors receive a score of 0.

4.1.2 Counterfactual Example

We evaluate the generated counterfactuals using three automated metrics widely adopted in the literature (Ross et al., 2021; Bhan et al., 2023; Nguyen et al., 2024b; Wang et al., 2025c).

Label Flipping Rate (LFR) For a dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ containing N pairs of original inputs x_i and gold labels y_i, LFR captures the frequency with which the generated counterfactual $\tilde{x}_i$ alters the original model prediction $\hat{y}_i$ on x_i to a different one $\tilde{y}_i$ (Ge et al., 2021; Bhattacharjee et al., 2024; Wang et al., 2025b). The LFR is calculated as:

$$LFR = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(\hat{y}_i \neq \tilde{y}_i)$$

where $\mathbb{I}$ is the indicator function, which returns 1 if the condition is true and 0 otherwise.

Perplexity (PPL) represents the exponential of the average negative log-likelihood computed over a sequence. PPL is a commonly used metric in counterfactual evaluation literature to assess the fluency of generated counterfactuals by measuring the model's predictive accuracy for each word given its preceding context (Le et al., 2023; Nguyen et al., 2024a). For a given counterfactual $\tilde{x} = (t_1, t_2, \cdots, t_n)$ and a model parameterized by θ, PPL is computed as follows:

$$PPL(\tilde{x}) = \exp\left\{\frac{1}{n}\sum_{i=1}^{n} \log p_\theta(t_i|t_{<i})\right\}$$

Textual Similarity (TS) As counterfactuals $\tilde{x}$ should closely resemble the original inputs x, we measure this similarity using the Levenshtein distance d on token level, a standard metric in existing literature (Ross et al., 2021; Treviso et al., 2023):

$$TS = \frac{1}{N} \sum_{i=1}^{N} \frac{d(x_i, \tilde{x}_i)}{|x_i|}$$

[2]The implementation of integer quantization provided in BITSANDBYTES is limited to weight-only quantization: https://github.com/bitsandbytes-foundation/bitsandbytes.

4.2 Faithfulness

In addition to assessing the impact of quantization on SE quality, we also evaluate the effect of quantization on the faithfulness of NLEs and CFEs. This evaluation determines whether the SEs generated by quantized LLMs are still able to truly reflect their underlying reasoning process (Doshi-Velez and Kim, 2017; Jacovi and Goldberg, 2020). The faithfulness rate is defined as:

$$r_{\text{faith}} = \frac{N_{\text{faithful}}}{N}$$

where N is the number of evaluated explanations and N_{faith} denotes the number of faithful explanations. This metric enables direct comparison of faithfulness of NLE and CFE across various quantization methods and bit-widths, with higher values indicating greater faithfulness.

Faithfulness of NLEs. For NLE faithfulness evaluation, we employ two widely used faithfulness metrics: *biasing features* and *CC-SHAP*.[3] ① The biasing features metric marks explanations as unfaithful that do not reflect answer biases added in context examples (Turpin et al., 2023). ② CC-SHAP identifies unfaithful explanations when the model's input contribution distributions for prediction and reasoning diverge, with values ranging from -1 to +1 (Parcalabescu and Frank, 2024).

Faithfulness of CFEs. Following Madsen et al. (2024), we employ the self-consistency check to evaluate whether counterfactual predictions satisfy the targeted labels; counterfactuals meeting this criterion are deemed faithful. In this context, r_{faith} corresponds to the label flip rate (§4.1.2) – the ratio of valid and faithful counterfactuals to all instances that successfully alter the model prediction to the target label.

4.3 Human Evaluation

We further assess the effect of quantization on self-explanation quality (§3.1) by conducting a user study in which participants subjectively evaluate NLEs and CFEs along two dimensions (§4.3.1). This analysis identifies explanation qualities that necessitate human judgment, which extend beyond the scope of automatic evaluation metrics (§4.1).

4.3.1 Subjective Ratings

Following the design of Likert scales for explanation evaluation proposed by Feldhus et al. (2023); Chiang and Lee (2023), and user studies on NLEs and CFEs conducted by Domnich et al. (2025); Wang et al. (2025a); Shailya et al. (2025), we ask human annotators to evaluate NLEs and CFEs based on the following dimensions, each rated on a 5-point Likert scale ranging from "strongly disagree" (1) to "strongly agree" (5):

- **Trustworthiness**: Evaluate whether the provided explanation is trustworthy and can be relied upon by humans;
- **Coherence**: Assess whether the provided explanation is sensible and clear, and effectively captures the rationale.

4.3.2 User Study Setup

To assess the quality of explanations produced by various LLMs and quantization techniques, we conduct a user study involving $N = 48$ participants, who are all native English speakers without necessarily possessing expertise in explainability. We randomly sample ($k = 30$) dataset indices. For each model-precision pair, the self-explanations generated by the corresponding model in full precision or quantized using different methods are evaluated by at least two human annotators. Our user study focuses on the overlapping dataset between NLEs and CFEs (eSNLI), and on `Qwen2.5` models of sizes {7B, 32B, 72B} to capture a wide range of model scales. We exclude `Qwen2.5-14B` due to its consistently suboptimal performance, even without quantization. Each annotator is assigned 15 explanations, accompanied by two evaluation dimensions (§4.3.1), and tasked with assigning appropriate scores based on a given Likert scale.[4] We report inter-annotator agreements with Krippendorff's α of 0.71 for NLEs and 0.64 for CFEs.

5 Results

5.1 Automatic Evaluation

5.1.1 Natural Language Explanations

Quality of NLEs. Tables 1 and 5 demonstrate that NLE quality reduction *varies more substantially in smaller models but remains less affected in larger models*. Surprisingly, full-precision LLMs do not consistently outperform their quantized counterparts. Furthermore, LLMs with lower precision, e.g., those using `bib4`, occasionally generate

[3] The description of the faithfulness metrics is in App. J.

[4] Further details about annotator recruitment and annotation guidelines can be found in Appendix F.

Model	Preci -sion	NLE (eSNLI)		NLE (HealthFC)		CFE (eSNLI)			CFE (AG News)		
		BARTScore ↑	TIGERScore ↑	BARTScore ↑	TIGERScore ↑	LFR ↑	PPL ↓	TS ↓	LFR ↑	PPL ↓	TS ↓
Qwen2.5-7B	full	-6.56	**-0.13**	-4.41	**-0.34**	64.80%	94.34	**0.37**	35.40%	74.87	**0.53**
	bib4	-6.69	-0.22	-4.35	-0.64	67.40%	**64.99**	0.57	36.00%	95.26	0.61
	bib8	-6.56	-0.24	-7.84	-0.94	65.40%	88.85	0.37	**42.00%**	**72.97**	0.61
	gptq4	**-6.53**	-0.51	-4.29	-0.65	**68.60%**	92.39	0.57	34.80%	81.52	0.57
	gptq8	-6.60	-0.16	-6.02	-0.74	66.00%	79.48	0.39	31.20%	79.45	0.57
	awq	-6.59	-0.13	**-4.26**	-0.93	67.60%	99.97	0.35	38.40%	80.52	0.57
Qwen2.5-14B	full	**-6.52**	-0.25	-4.25	-0.39	67.20%	**90.67**	0.54	38.40%	102.83	0.66
	bib4	-6.73	-0.24	-4.30	-0.29	**67.60%**	93.03	0.59	**42.20%**	133.47	0.78
	bib8	-6.59	**-0.21**	-4.34	**-0.24**	64.40%	93.54	0.53	34.60%	99.28	0.63
	gptq4	-6.59	-0.25	**-4.23**	-0.41	64.40%	91.45	0.53	39.40%	99.97	0.66
	gptq8	-6.54	-0.30	-4.27	-0.31	63.60%	92.58	0.51	36.80%	95.11	0.66
	awq	-6.57	-0.81	-4.33	-0.48	67.60%	91.44	**0.47**	39.20%	**84.55**	**0.61**
Qwen2.5-32B	full	-7.68	-0.44	-8.17	-2.68	64.20%	87.07	0.43	31.40%	83.13	0.49
	bib4	-10.50	-1.28	**-6.06**	**-1.61**	**64.80%**	**79.35**	0.43	**39.20%**	80.93	0.56
	bib8	-8.77	-0.70	-9.07	-3.00	64.00%	86.11	0.41	34.00%	82.82	0.49
	gptq4	**-6.61**	-0.56	-6.62	-1.88	63.60%	87.07	**0.41**	33.40%	81.48	**0.48**
	gptq8	-7.91	-0.90	-8.54	-2.86	63.60%	90.03	0.43	35.40%	84.81	0.49
	awq	-9.91	**-0.27**	-8.97	-3.27	63.40%	92.82	0.41	37.20%	**80.60**	0.52
Qwen2.5-72B	full	-6.52	-0.47	-4.21	-0.58	61.20%	117.85	0.39	28.80%	96.45	0.43
	bib4	-6.54	**-0.39**	-4.21	-0.58	63.60%	120.36	0.44	25.40%	93.03	0.45
	bib8	**-6.51**	-0.52	-4.22	**-0.57**	61.60%	**112.60**	0.40	**31.20%**	91.51	0.44
	gptq4	-6.52	-0.47	**-4.17**	-0.60	61.40%	119.59	0.39	29.60%	96.05	0.44
	gptq8	-6.55	-0.53	-4.20	-0.65	61.20%	115.94	0.41	26.40%	96.47	0.45
	awq	-6.52	-0.51	-4.22	-0.66	**64.00%**	119.61	**0.37**	24.60%	**88.97**	**0.42**
Llama3-8B	full	-6.62	-0.37	-4.29	-1.01	66.00%	**62.26**	**0.41**	48.80%	**42.88**	1.54
	bib4	**-6.60**	-0.54	-4.65	-1.30	63.60%	81.80	0.42	48.40%	46.01	1.38
	bib8	-6.66	-0.37	**-4.26**	**-0.95**	63.40%	72.17	0.41	51.60%	46.95	1.60
	gptq4	-7.42	-0.29	-4.36	-1.28	**67.60%**	76.55	0.54	**65.20%**	62.61	1.60
	awq	-6.76	**-0.26**	-4.43	-1.31	67.60%	75.21	0.53	54.00%	79.87	**1.21**
Llama3-70B	full	**-6.58**	-0.16	-4.48	-1.04	64.80%	**80.77**	**0.41**	49.00%	**101.97**	0.40
	bib4	-6.62	-0.71	-4.44	-1.56	62.40%	139.80	0.47	49.60%	130.05	0.48
	bib8	-6.76	-0.26	**-4.18**	-1.30	**68.80%**	102.23	0.45	46.20%	140.24	**0.25**
	gptq4	-6.62	**-0.14**	-4.32	-1.12	63.27%	94.48	0.46	**53.20%**	115.03	0.52
	awq	-6.61	-0.15	-4.54	**-1.01**	64.80%	114.89	0.45	47.40%	106.90	0.42

Table 1: Automatic evaluation results of NLEs and CFEs generated by `Llama3` (8B, 70B) and `Qwen2.5` (7B, 14B, 32B, 72B) models with full precision and different quantization methods. For NLEs, we use `ZeroCoT` on eSNLI and HealthFC, evaluated by *BARTScore* and *TIGERScore*. For CFEs, we use `FIZLE` on eSNLI and AG News, evaluated by *label flip rate (LFR)*, *perplexity (PPL)*, and *text similarity (TS)*. **Bold** values indicate the best-performing approach for each model, while underlined values denote the best-performing quantization method. Table 5 further provides results for ECQA and IMDb.

higher-quality NLEs than LLMs with higher precision. In addition, Table 9 reveals that, especially for larger models, quantization-induced task performance degradation generally does not contribute to NLE quality degradation, as indicated by weak or even negative correlations.

Faithfulness of NLEs. Table 2 reveals that quantization generally induces *moderate declines* in NLE faithfulness across all faithfulness metrics (biasing features ↓ 3.4%, CC-SHAP ↓0.05). Analysis of transition patterns confirms that faithfulness is preserved in the majority of cases (Figure 3, Appendix J; Table 10). Notably, compared to `Qwen2.5` models, `Llama3` models are more susceptible to quantization-induced degradation of NLE faithfulness, especially at 4-bit precisions, e.g., `Llama3-8B` with `GPTQ4` on HealthFC assessed by CC-SHAP. Thus, although the drop is on average modest, in some configurations, the drop could be substantial, which is not easily predictable. As a result, *practitioners must validate the impact on SE faithfulness per use case*. Moreover, NLEs generated by larger models tend to be more faithful, aligned with the finding from Siegel et al. (2025), and larger models show greater robustness to quantization in preserving NLE faithfulness.

5.1.2 Counterfactual Examples

Tables 1 and 5 show that quantization negatively affects LLMs' ability to generate CFEs, causing *substantial counterfactual quality degradation* in *validity*, *fluency*, and *textual similarity* (§4.1), with fluency being most affected (on average 6.3%). This degradation is particularly pronounced for smaller LLMs, whereas larger LLMs demonstrate greater robustness. Moreover, as discussed in Section 4.2, the validity of counterfactuals, measured by LFR, simultaneously reflects their faithfulness (Table 1, Figure 3). We observe that counterfactual

Model	Precision	eSNLI		HealthFC		ECQA	
		Bias	CC-SHAP	Bias	CC-SHAP	Bias	CC-SHAP
Qwen2.5-7B	full	90.20	0.843	**93.43**	0.772	0.872	0.039
	bib4	85.60	**0.852**	91.14	0.765	0.902	**0.107**
	bib8	**91.20**	0.849	93.14	0.770	0.884	0.010
	gptq4	86.20	0.840	92.86	**0.774**	**0.904**	0.021
	gptq8	89.60	0.844	93.14	0.768	0.878	0.033
	awq	85.00	0.821	92.29	0.764	0.886	-0.033
Qwen2.5-14B	full	90.00	0.819	94.29	0.804	0.888	0.170
	bib4	91.00	0.832	**95.14**	0.797	0.856	**0.237**
	bib8	89.20	0.790	93.14	0.803	0.872	0.196
	gptq4	41.80	**0.895**	49.43	**0.900**	0.306	0.164
	gptq8	92.00	0.820	94.00	0.808	**0.892**	0.160
	awq	**93.00**	0.845	93.43	0.771	0.874	0.153
Qwen2.5-32B	full	93.00	0.835	95.43	0.808	**0.920**	**0.553**
	bib4	**93.80**	0.839	94.00	0.815	0.902	0.298
	bib8	90.60	**0.842**	95.14	0.805	0.918	0.440
	gptq4	93.60	**0.842**	95.14	**0.820**	0.918	0.379
	gptq8	93.40	0.835	94.57	0.813	0.914	0.401
	awq	92.60	0.840	**96.29**	0.806	0.902	0.330
Qwen2.5-72B	full	95.60	0.876	**97.71**	**0.826**	**0.922**	**0.797**
	bib4	95.40	**0.884**	96.86	0.821	0.880	0.675
	bib8	95.80	0.874	97.43	0.824	0.916	0.714
	gptq4	**96.40**	0.875	96.57	0.817	0.900	0.692
	gptq8	**96.40**	0.880	**97.71**	0.821	0.920	0.663
	awq	95.60	0.875	96.86	0.813	0.908	0.548
Llama3-8B	full	52.60	0.742	**84.30**	0.745	0.616	0.946
	bib4	**56.60**	**0.776**	64.00	**0.749**	**0.696**	**0.994**
	bib8	47.20	0.712	82.30	0.736	0.600	0.954
	gptq4	50.80	0.651	68.00	0.584	0.486	0.711
	awq	42.40	0.682	78.00	0.567	0.576	0.993
Llama3-70B	full	87.00	**0.741**	90.86	**0.409**	**0.870**	**0.895**
	bib4	83.40	0.278	89.43	0.313	0.854	0.834
	bib8	77.00	0.628	80.29	0.319	0.782	0.886
	gptq4	83.40	0.512	91.71	0.407	0.858	0.523
	awq	**88.60**	0.469	**93.14**	0.368	**0.870**	0.658

Table 2: Faithfulness rate (in %) of natural language explanation evaluated using biasing features (Bias), and CC-SHAP values on eSNLI, HealthFC, and ECQA across various quantization configurations. **Bold** values indicate the best-performing approach for each model, while underlined values denote the best-performing quantization method.

faithfulness decreases by an average of 1.5% under quantization and smaller LLMs exhibit more noticeable faithfulness drops. Counterintuitively, we find that full-precision models may underperform quantized models in generating effective counterfactuals, and larger quantized models sometimes generate lower-quality counterfactuals compared to smaller full-precision models.

5.2 Human Evaluation

Self-explanations generated by full-precision models are more trustworthy and coherent. Table 6 presents results from the human evaluation, showing that, overall, NLEs and CFEs generated by full-precision models are perceived as generally more trustworthy and coherent than those generated by quantized models (Figure 1 and 2). This can be attributed to quantization's effect on confidence calibration: By introducing truncation or rounding, it is harder for the model to capture contextual semantics due to distribution shifts (Proskurina et al., 2024). As a result, the coherence of

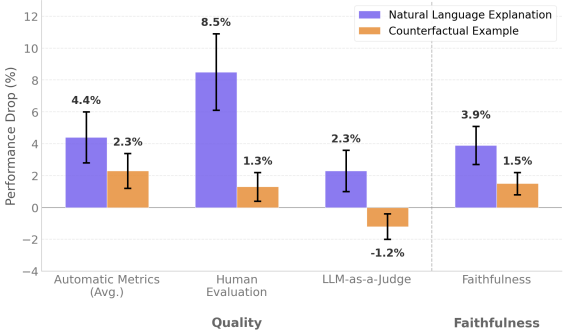


Figure 2: Averaged drop in self-explanation *quality* and *faithfulness* with confidence intervals, as assessed by various automatic evaluation metrics (§4.1), human and LLM-as-a-Judge evaluation dimensions (§4.3.1), and faithfulness metrics (§4.2).

generated text is impaired (Resendiz and Klinger, 2025) and the likelihood of hallucinations increases (Li et al., 2024a), ultimately diminishing annotators' trust in the self-explanations.

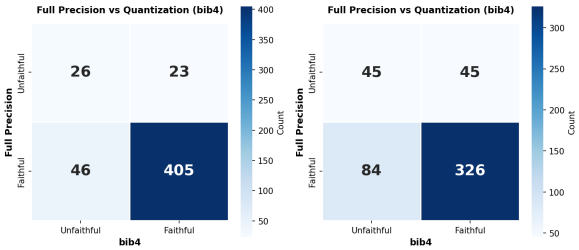
LLM-as-a-Judge evaluation fails to fully capture the impact of quantization. To facilitate comparative analysis, an LLM-as-a-Judge (LaaJ) evaluation is conducted on the identical subset of data examples selected for human assessment of trustworthiness and coherence (§4.3.2). Setup details and full correlation results are in Appendix G. We find a striking discrepancy: while judge models generally agree with each other (Figure 6 and 8, especially on coherence, with κ overall above 0.69), their agreement with human raters is consistently weak or even occasionally negative and is not statistically significant (all p-values > 0.05; Figure 7 and 9). This misalignment is particularly pronounced for CFEs, where judges systematically diverge from human judgments far more than for NLEs. These findings indicate that, in our setting, LaaJ cannot substitute for human evaluation of quantization effects. One plausible explanation is that the SE quality differences induced by quantization in our setting are subtle (Figure 2; particularly for CFEs where the full-precision vs. quantized gap is only 1.3%), and within this narrow margin judges may tend to converge on surface features (e.g., fluency, length) that diverge from the subtle quality signals humans pick up on. Therefore, human evaluation should preferably be considered when subjective judgment is involved in assessing the effect of quantization of SE quality.

5.3 Discussion

Various methods for self-explanations have different sensitivities to quantization. Figure 2 illustrates that quantization impacts self-explanations differently, though it moderately degrades self-explanation quality by up to 8.5%. NLEs exhibit greater sensitivity, while CFEs demonstrate relative robustness to quantization, with NLE quality degradation being substantially more pronounced (Table 1). Thus, for applications where CFEs are suitable, quantization presents lower risk to explanation quality and faithfulness.

Quantization generally leads to declines in self-explanation quality. Table 1 displays that no single quantization method invariably outperforms others across all experimental configurations, making it challenging to predict accurately the quantitative impact of quantization on self-explanation quality, since the magnitude of quality degradation depends on the specific *quantization techniques* and *deployed models*. Nevertheless, quantization generally leads to self-explanation quality degradation (Figure 2; most $p < 0.05$), though surprisingly, it can sometimes even improve self-explanation quality. The increase in SE quality from quantization may arise from the reduced entropy of the output distribution and from more consistent, simple language use, as quantization narrows the diversity of LLM outputs (Guo et al., 2025). Additionally, we observe that *larger quantized models do not consistently generate higher-quality self-explanations than smaller full-precision models*, contradicting the finding of Badshah and Sajjad (2024). Furthermore, LLMs with lower-bit precision do not invariably perform worse than those with higher bit precision. These findings may stem from regularization effects (Park et al., 2022) or noise (Li et al., 2024b) introduced by quantization, which limits weight precision and may inadvertently enhance self-explanation quality.

Self-explanation faithfulness is adversely impacted by quantization. Figure 2 reveals that, overall, quantization does not notably affect the self-explanation faithfulness, with average degradation of only 1.5% for CFEs and 3.9% for NLEs. This minimal impact is evidenced by the fact that the faithfulness of SEs generated by the full-precision and quantized models remains largely unchanged. However, there are more cases with full-precision explanations remaining faithful while



(a) Natural language expl. (b) Counterfactual example

Figure 3: Self-explanation faithfulness variation due to quantization for `Qwen2.5-7B` with `bib4` on eSNLI measured by biasing feature and self-consistency.

quantized versions become unfaithful (Figure 3), although faithfulness can occasionally be surprisingly enhanced through quantization. A possible assumption is that the model retains core reasoning pathways while discards spurious correlations that lead to unfaithful or lower-quality explanations (Mulchandani and Kim, 2025). Moreover, we observe that *self-explanations from larger quantized models are more frequently faithful than those from smaller full-precision models*, as smaller models experience more pronounced faithfulness degradation from quantization (Appendix J).[5] Consequently, *when SE faithfulness is critical, practitioners should consider employing larger quantized models rather than smaller full-precision models*.

Ranking of quantization methods based on the extent of degradation. We assign rankings to quantization methods for each model based on quality changes compared to full-precision models. Subsequently, we calculate the mean ranking across all experimental configurations. We find that no single quantization method consistently outperforms others across *task performance*, self-explanation *quality*, and *faithfulness*. Figure 11 shows that `GPTQ8`, `AWQ`, and `bib8` excel at preserving these metrics, respectively. Moreover, we observe a trade-off among quantization methods between self-explanation (both quality and faithfulness) and task performance preservation. Notably, lower-bit methods can generate explanations with comparable quality to their higher-bit counterparts.

Summary. Quantization leads to degradation in self-explanation quality and faithfulness, with this

[5]This highlights that quantization affects SE quality differently from faithfulness (App. K). We hypothesize that under quantization, weights may be more sensitive compared to the decision boundary due to overparametrization, indicating that faithfulness can be better preserved compared to SE quality.

trend becoming **more pronounced in smaller models** (Table 1 and 2).[6] Surprisingly, quantized models occasionally generate self-explanations of higher quality than their full-precision counterparts. Moreover, lower-bit quantization does not necessarily produce inferior self-explanations compared to higher-bit quantization. Although **quantization effects exhibit variability across models and techniques**, our findings indicate only modest degradation in self-explanation quality and faithfulness (Figure 2 and 3), **rendering quantization a viable compression strategy**. Nevertheless, practitioners should proceed cautiously when deploying quantized LLMs in transparency-critical applications.

6 Conclusion

In this work, we examine the impact of quantization on two free-text self-explanation types concerning explanation quality and faithfulness, employing three quantization techniques across six LLMs of varying sizes. Quantization generally causes degradation in both self-explanation quality and faithfulness. While larger models demonstrate limited robustness to quantization regarding explanation quality, they are more robust in preserving faithfulness. Across our experiments, the impact of quantization on self-explanations is highly context-dependent, and no single quantization method consistently outperforms others across *task performance*, explanation *quality*, and *faithfulness*. Our user study further reveals that quantization reduces the coherence and trustworthiness of self-explanations. This heterogeneity suggests practitioners should empirically test multiple quantization strategies for their specific use case rather than assuming a one-size-fits-all solution. Nevertheless, the modest explanation quality and faithfulness degradation indicates that quantized models retain their competence for self-explanation and does not undermine quantization's viability as a model compression strategy.

Limitations

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, following Singh and Sajjad (2025), the scope of our experiments is limited to post-training quantization (PTQ) techniques. The rationale for focusing on PTQ is twofold: PTQ facilitates an ad-hoc quantization during inference and it offers computational efficiency without necessitating modifications to the model architecture or training process.

Although quantization can simultaneously affect other model capabilities, we argue that disentangling the impact of quantization from other confounding factors is infeasible, due to the black-box nature of LLMs. Consistent with prior work across multiple domains, e.g., model calibration (Singh and Sajjad, 2025), multilinguality (Marchisio et al., 2024), and alignment (Jin et al., 2024), which similarly does not disentangle confounding factors, we adopt established experimental protocols while focusing on patterns that demonstrate notable divergence from full-precision models.

Ethics Statement

The conducted user study was ethically approved by the Ethics Committee of Faculty IV of Technische Universität Berlin. The participants in our user studies were compensated at or above the minimum wage in accordance with the standards of our host institutions' regions. The annotation took each annotator 45 minutes on average.

Author Contributions

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

- QW: Writing, idea conceptualization, experiments and evaluations, formal analysis, visualization.
- NF: Writing – review & editing, supervision, idea conceptualization.
- PA: Writing – review & editing and idea conceptualization of the comparison between LLM-as-a-Judge evaluation and human evaluation.
- FS: NLE faithfulness evaluation.
- SO: Writing – review & editing.
- SM: Supervision, review & editing, and funding acquisition.
- VS: Funding acquisition.

Acknowledgment

We sincerely thank **Martin Tutek** and CopeNLU members, including **Jingyi Sun**, **Nadav Borenstein**, and **Sekh Mainul Islam**, for their thorough review of an early draft and the anonymous reviewers of EMNLP 2025 and ACL 2026 for their helpful

[6] To demonstrate the generalizability of our findings to non-free-text SEs, we conduct an additional evaluation of feature attribution (Madsen et al., 2024) in Appendix D.

and rigorous feedback. This work has been supported by the Federal Ministry of Research, Technology and Space (BMFTR) as part of the projects BIFOLD 24B and VERANDA (16KIS2047).

References

Chirag Agarwal, Sree Harsha Tanneru, and Himabindu Lakkaraju. 2024. Faithfulness vs. plausibility: On the (un)reliability of explanations from large language models. *Preprint*, arXiv:2402.04614.

Shourya Aggarwal, Divyanshu Mandowara, Vishwajeet Agrawal, Dinesh Khandelwal, Parag Singla, and Dinesh Garg. 2021. Explanations for CommonsenseQA: New Dataset and Models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3050–3065, Online. Association for Computational Linguistics.

AI@Meta. 2024. Llama 3 model card.

Sher Badshah and Hassan Sajjad. 2024. Quantifying the capabilities of llms across scale and precision. *Preprint*, arXiv:2405.03146.

Milan Bhan, Jean-noel Vittaut, Nicolas Chesneau, and Marie-jeanne Lesot. 2023. Enhancing textual counterfactual explanation intelligibility through counterfactual feature importance. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, pages 221–231, Toronto, Canada. Association for Computational Linguistics.

Amrita Bhattacharjee, Raha Moraffah, Joshua Garland, and Huan Liu. 2024. Zero-shot LLM-guided Counterfactual Generation: A Case Study on NLP Model Evaluation . In *2024 IEEE International Conference on Big Data (BigData)*, pages 1243–1248, Los Alamitos, CA, USA. IEEE Computer Society.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: natural language inference with natural language explanations. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 9560–9572, Red Hook, NY, USA. Curran Associates Inc.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada. Association for Computational Linguistics.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

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.

Tomoki Doi, Masaru Isonuma, and Hitomi Yanaka. 2025. Investigating training and generalization in faithful self-explanations of large language models. *Preprint*, arXiv:2512.07288.

Marharyta Domnich, Julius Välja, Rasmus Moorits Veski, Giacomo Magnifico, Kadi Tulver, Eduard Barbu, and Raul Vicente. 2025. Towards unifying evaluation of counterfactual explanations: Leveraging large language models for human-centric assessments. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(15):16308–16316.

Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *Preprint*, arXiv:1702.08608.

Nils Feldhus, Qianli Wang, Tatiana Anikina, Sahil Chopra, Cennet Oguz, and Sebastian Möller. 2023. InterroLang: Exploring NLP models and datasets through dialogue-based explanations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5399–5421, Singapore. Association for Computational Linguistics.

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*.

Yingqiang Ge, Shuchang Liu, Zelong Li, Shuyuan Xu, Shijie Geng, Yunqi Li, Juntao Tan, Fei Sun, and Yongfeng Zhang. 2021. Counterfactual evaluation for explainable ai. *Preprint*, arXiv:2109.01962.

Flavio Giorgi, Cesare Campagnano, Fabrizio Silvestri, and Gabriele Tolomei. 2025. Natural language counterfactual explanations for graphs using large language models. In *The 28th International Conference on Artificial Intelligence and Statistics*.

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.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2025. A survey on llm-as-a-judge. *Preprint*, arXiv:2411.15594.

Yanzhu Guo, Guokan Shang, and Chloé Clavel. 2025. Benchmarking linguistic diversity of large language models. *Transactions of the Association for Computational Linguistics*, 13:1507–1526.

Yi-Sheng Hsu, Nils Feldhus, and Sherzod Hakimov. 2025. Free-text rationale generation under readability level control. In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM²)*, pages 129–150, Vienna, Austria and virtual meeting. Association for Computational Linguistics.

Fan Huang, Haewoon Kwak, Kunwoo Park, and Jisun An. 2024. ChatGPT rates natural language explanation quality like humans: But on which scales? In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3111–3132, Torino, Italia. ELRA and ICCL.

Shiyuan Huang, Siddarth Mamidanna, Shreedhar Jangam, Yilun Zhou, and Leilani H. Gilpin. 2023. Can large language models explain themselves? a study of llm-generated self-explanations. *Preprint*, arXiv:2310.11207.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205, Online. Association for Computational Linguistics.

Dongfu Jiang, Yishan Li, Ge Zhang, Wenhao Huang, Bill Yuchen Lin, and Wenhu Chen. 2024. TIGERScore: Towards building explainable metric for all text generation tasks. *Transactions on Machine Learning Research*.

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. 2024. The impact of inference acceleration strategies on bias of large language models. In *Neurips Safe Generative AI Workshop 2024*.

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213. Curran Associates, Inc.

Tiep Le, Vasudev Lal, and Phillip Howard. 2023. COCO-counterfactuals: Automatically constructed counterfactual examples for image-text pairs. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Junyi Li, Jie Chen, Ruiyang Ren, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2024a. The dawn after the dark: An empirical study on factuality hallucination in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10879–10899, Bangkok, Thailand. Association for Computational Linguistics.

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

Shiyao Li, Xuefei Ning, Luning Wang, Tengxuan Liu, Xiangsheng Shi, Shengen Yan, Guohao Dai, Huazhong Yang, and Yu Wang. 2024c. Evaluating quantized large language models. In *Forty-first International Conference on Machine Learning*.

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.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Andreas Madsen, Sarath Chandar, and Siva Reddy. 2024. Are self-explanations from large language models faithful? In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 295–337,

Bangkok, Thailand. Association for Computational Linguistics.

Andreas Madsen, Siva Reddy, and Sarath Chandar. 2022. Post-hoc interpretability for neural nlp: A survey. *ACM Computing Survey*, 55(8).

Ana Marasovic, Iz Beltagy, Doug Downey, and Matthew Peters. 2022. Few-shot self-rationalization with natural language prompts. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 410–424, Seattle, United States. Association for Computational Linguistics.

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.

Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38.

Lucas Monteiro Paes, Dennis Wei, Hyo Jin Do, Hendrik Strobelt, Ronny Luss, Amit Dhurandhar, Manish Nagireddy, Karthikeyan Natesan Ramamurthy, Prasanna Sattigeri, Werner Geyer, and Soumya Ghosh. 2025. Multi-level explanations for generative language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32291–32317, Vienna, Austria. Association for Computational Linguistics.

Varun Mulchandani and Jung-Eun Kim. 2025. Severing spurious correlations with data pruning. In *The Thirteenth International Conference on Learning Representations*.

Van Bach Nguyen, Christin Seifert, and Jörg Schlötterer. 2024a. CEval: A benchmark for evaluating counterfactual text generation. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 55–69, Tokyo, Japan. Association for Computational Linguistics.

Van Bach Nguyen, Paul Youssef, Christin Seifert, and Jörg Schlötterer. 2024b. LLMs for generating and evaluating counterfactuals: A comprehensive study. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14809–14824, Miami, Florida, USA. Association for Computational Linguistics.

OpenAI, :, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, and 108 others. 2025. gpt-oss-120b & gpt-oss-20b model card. *Preprint*, arXiv:2508.10925.

Letitia Parcalabescu and Anette Frank. 2024. On measuring faithfulness or self-consistency of natural language explanations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6048–6089, Bangkok, Thailand. 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.

Irina Proskurina, Luc Brun, Guillaume Metzler, and Julien Velcin. 2024. When quantization affects confidence of large language models? In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1918–1928, Mexico City, Mexico. Association for Computational Linguistics.

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, and 23 others. 2024. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Korbinian Randl, John Pavlopoulos, Aron Henriksson, and Tony Lindgren. 2025. Evaluating the reliability of self-explanations in large language models. In *Discovery Science: 27th International Conference, DS 2024, Pisa, Italy, October 14–16, 2024, Proceedings, Part I*, page 36–51, Berlin, Heidelberg. Springer-Verlag.

Yarik Menchaca Resendiz and Roman Klinger. 2025. Llm-based affective text generation quality based on different quantization values. *Preprint*, arXiv:2501.19317.

Alexis Ross, Ana Marasović, and Matthew Peters. 2021. Explaining NLP models via minimal contrastive editing (MiCE). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3840–3852, Online. Association for Computational Linguistics.

Krithi Shailya, Shreya Rajpal, Gokul S Krishnan, and Balaraman Ravindran. 2025. Lext: Towards evaluating trustworthiness of natural language explanations. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '25, page 1565–1587, New York, NY, USA. Association for Computing Machinery.

Noah Y. Siegel, Nicolas Heess, Maria Perez-Ortiz, and Oana-Maria Camburu. 2025. Verbosity tradeoffs and the impact of scale on the faithfulness of llm self-explanations. *Preprint*, arXiv:2503.13445.

Manpreet Singh and Hassan Sajjad. 2025. Interpreting the effects of quantization on llms. *Preprint*, arXiv:2508.16785.

Jingyi Sun, Qianli Wang, Pepa Atanasova, Nils Feldhus, and Isabelle Augenstein. 2026. Investigating the interplay between contextual and parametric chain-of-thought faithfulness under optimization. *Preprint*, arXiv:2605.24960.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Marcos Treviso, Alexis Ross, Nuno M. Guerreiro, and André Martins. 2023. CREST: A joint framework for rationalization and counterfactual text generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15109–15126, Toronto, Canada. Association for Computational Linguistics.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Martin Tutek, Fateme Hashemi Chaleshtori, Ana Marasovic, and Yonatan Belinkov. 2025. Measuring chain of thought faithfulness by unlearning reasoning steps. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 9946–9971, Suzhou, China. Association for Computational Linguistics.

Luis Felipe Villa-Arenas, Ata Nizamoglu, Qianli Wang, Sebastian Möller, and Vera Schmitt. 2024. Anchored alignment for self-explanations enhancement. *Preprint*, arXiv:2410.13216.

Juraj Vladika, Phillip Schneider, and Florian Matthes. 2024. HealthFC: Verifying health claims with evidence-based medical fact-checking. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8095–8107, Torino, Italia. ELRA and ICCL.

Zhongwei Wan, Xin Wang, Che Liu, Samiul Alam, Yu Zheng, Jiachen Liu, Zhongnan Qu, Shen Yan, Yi Zhu, Quanlu Zhang, Mosharaf Chowdhury, and Mi Zhang. 2024. Efficient large language models: A survey. *Transactions on Machine Learning Research*. Survey Certification.

Qianli Wang, Tatiana Anikina, Nils Feldhus, Josef Genabith, Leonhard Hennig, and Sebastian Möller. 2024. LLMCheckup: Conversational examination of large language models via interpretability tools and self-explanations. In *Proceedings of the Third Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, pages 89–104, Mexico City, Mexico. Association for Computational Linguistics.

Qianli Wang, Tatiana Anikina, Nils Feldhus, Simon Ostermann, Sebastian Möller, and Vera Schmitt. 2025a. Cross-refine: Improving natural language explanation generation by learning in tandem. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1150–1167, Abu Dhabi, UAE. Association for Computational Linguistics.

Qianli Wang, Nils Feldhus, Simon Ostermann, Luis Felipe Villa-Arenas, Sebastian Möller, and Vera Schmitt. 2025b. FitCF: A framework for automatic feature importance-guided counterfactual example generation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 1176–1191, Vienna, Austria. Association for Computational Linguistics.

Qianli Wang, Van Bach Nguyen, Nils Feldhus, Luis Felipe Villa-Arenas, Christin Seifert, Sebastian Möller, and Vera Schmitt. 2025c. Truth or twist? optimal model selection for reliable label flipping evaluation in LLM-based counterfactuals. In *Proceedings of the 18th International Natural Language Generation Conference*, pages 80–97, Hanoi, Vietnam. Association for Computational Linguistics.

Qianli Wang, Van Bach Nguyen, Yihong Liu, Fedor Splitt, Nils Feldhus, Christin Seifert, Hinrich Schuetze, Sebastian Möller, and Vera Schmitt. 2026a. Parallel universes, parallel languages: A comprehensive study on LLM-based multilingual counterfactual example generation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 323–346, San Diego, California, United States. Association for Computational Linguistics.

Qianli Wang, Mingyang Wang, Nils Feldhus, Simon Ostermann, Yuan Cao, Hinrich Schuetze, Sebastian Möller, and Vera Schmitt. 2026b. Through a compressed lens: Investigating the impact of quantization on factual knowledge recall. In *Proceedings of the 6th Workshop on Trustworthy NLP (TrustNLP 2026)*, pages 21–39, San Diego, California. Association for Computational Linguistics.

Yilong Wang, Qianli Wang, Bohao Chu, Yihong Liu, Jing Yang, and Simon Ostermann. 2026c. Macro: Enhancing multilingual counterfactual explanations through alignment-as-preference optimization. *Preprint*, arXiv:2605.11632.

Yilong Wang, Qianli Wang, and Nils Feldhus. 2026d. iflip: Iterative feedback-driven counterfactual example refinement. *Preprint*, arXiv:2601.01446.

Sarah Wiegreffe, Ana Marasović, and Noah A. Smith. 2021. Measuring association between labels and free-text rationales. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10266–10284, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel Weld. 2021. Polyjuice: Generating counterfactuals for explaining, evaluating, and improving models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6707–6723, Online. Association for Computational Linguistics.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. In *Advances in Neural Information Processing Systems*, volume 34, pages 27263–27277. Curran Associates, Inc.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Zixuan Zhou, Xuefei Ning, Ke Hong, Tianyu Fu, Jiaming Xu, Shiyao Li, Yuming Lou, Luning Wang, Zhihang Yuan, Xiuhong Li, Shengen Yan, Guohao Dai, Xiao-Ping Zhang, Yuhan Dong, and Yu Wang. 2024. A survey on efficient inference for large language models. *Preprint*, arXiv:2404.14294.

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

A.1 Dataset Examples

Figure 4 presents examples from the eSNLI, AG News, and HealthFC datasets.

A.2 Label Distribution

Label distributions of eSNLI, AG News, and HealthFC are shown in Figure 5.

B Quantization Method

We further provide a detailed overview of three selected quantization methods employed in our experiments (§3.3).

GPTQ. GPTQ is a post-training quantization technique that compresses a large language model by reducing its weights to a low precision (typically 4-bit) without needing to retrain the model (Frantar et al., 2023). It works layer-by-layer and group-by-group, solving a local least-squares optimization problem for each set of weights. Crucially, it uses second-order information (Hessian estimates) to intelligently decide which weights can be approximated (quantized) with the least impact on the model's overall output accuracy.

AWQ. AWQ is an Activation-aware Weight Quantization technique that compresses Large Language Models to low precision by prioritizing accuracy (Lin et al., 2024). It uses a calibration dataset to find salient channels (groups of weights) that are highly sensitive to the model's activations and scales up these critical weights before quantization to protect them from accuracy loss when their precision is reduced.

BitSandBytes BITSANDBYTES (Dettmers et al., 2022) identifies and isolates outliers, which are model weights or data points with values significantly deviating from the norm. To maintain high precision, these outliers are preserved in 16-bit floating-point format. The remaining non-outliers (standard-range values) are efficiently quantized to 4- or 8-bit integers.

C Models & Inference Time

Table 3 presents details of 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. Explanation generation across the entire dataset, including both natural language explanations (NLEs) and counterfactual examples (CFEs), can be completed within 10 hours.

eSNLI (Natural Language Inference)

Premise: This church choir sings to the masses as they sing joyous songs from the book at a church.

Hypothesis: The church has cracks in the ceiling.

Label: Neutral

Explanation: Not all churches have cracks in the ceiling.

AG News (News Topic Classification)

News: E-mail scam targets police chief Wiltshire Police warns about "phishing" after its fraud squad chief was targeted.

Label: sci/tech

HealthFC (Fact-checking)

Question: Does chicken soup help with colds?

Document: However, such experiments – even though they may sound so promising – do not provide any evidence that the soup also works the same in the human body. Chicken soup in case of cold: theories without evidence Extensive studies on the effect of chicken soup do not exist, but all the more attempts to explain them. As long as there are no studies with human subjects, we simply cannot assess whether and which chicken soup ingredients could help cold-stricken people.

Label: unknown

Explanation: "So far, this has only been investigated in laboratory experiments. Studies on efficacy in humans are missing so far. Therefore, we can not judge whether chicken soup is helpful for colds."

Figure 4: Examples from eSNLI, AG News and HealthFC.

D Prompting-based Feature Attribution

Motivation. In this paper, we specifically focus on two representative *free-text* self-explanations: natural language explanations (NLEs) and counterfactual examples (CFEs) (§3.1). The selection of self-explanations is motivated by two considerations: (1) their widespread adoption in the literature (Madsen et al., 2024; Agarwal et al., 2024; Wang et al., 2025b); and (2) their free-text nature, which facilitates human comprehension and enables more straightforward quantification of quantization-induced degradation through user studies. Furthermore, NLEs and CFEs necessitate multi-dimensional evaluation, whereas feature attribution and redaction-based explanations generally require assessment only along task performance and faithfulness dimensions rather than text quality dimensions. Additionally, the other explanation methods largely depend on the type of technique instead of the underlying model, indicating that quantization may not have large effects on them.

Setup for Feature Attribution Evaluation. In order to further show the generalizability of our findings, we conduct a small-scale experiments on prompt-based feature attribution (Madsen et al., 2024), which is a not necessarily faithful approximation of model internal-based feature attribution, on eSNLI with `Qwen2.5-{7,14,32,72B}` and `Llama3-{8,70B}` with full-precision, AWQ, GPTQ4/8, and bib4/8, while the degradation of

Name	Citation	Size	Precision	Link
Llama3-8B	AI@Meta (2024)	8B	Full	https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct
Llama3-8B	AI@Meta (2024)	2B	GPTQ4	https://huggingface.co/TechxGenus/Meta-Llama-3-8B-Instruct-GPTQ
Llama3-8B	AI@Meta (2024)	2B	AWQ	https://huggingface.co/TechxGenus/Meta-Llama-3-8B-Instruct-AWQ
Llama3-70B	AI@Meta (2024)	70B	full	https://huggingface.co/meta-llama/Meta-Llama-3-70B-Instruct
Llama3-70B	AI@Meta (2024)	11B	GPTQ4	https://huggingface.co/TechxGenus/Meta-Llama-3-70B-Instruct-GPTQ
Llama3-70B	AI@Meta (2024)	11B	AWQ	https://huggingface.co/TechxGenus/Meta-Llama-3-70B-Instruct-AWQ
Qwen2.5-7B	Qwen et al. (2024)	7B	Full	https://huggingface.co/Qwen/Qwen2.5-7B-Instruct
Qwen2.5-7B	Qwen et al. (2024)	2B	AWQ	https://huggingface.co/Qwen/Qwen2.5-7B-Instruct-AWQ
Qwen2.5-7B	Qwen et al. (2024)	2B	GPTQ4	https://huggingface.co/Qwen/Qwen2.5-7B-Instruct-GPTQ-Int4
Qwen2.5-7B	Qwen et al. (2024)	3B	GPTQ8	https://huggingface.co/Qwen/Qwen2.5-7B-Instruct-GPTQ-Int8
Qwen2.5-14B	Qwen et al. (2024)	14B	Full	https://huggingface.co/Qwen/Qwen2.5-14B-Instruct
Qwen2.5-14B	Qwen et al. (2024)	3B	AWQ	https://huggingface.co/Qwen/Qwen2.5-14B-Instruct-AWQ
Qwen2.5-14B	Qwen et al. (2024)	3B	GPTQ4	https://huggingface.co/Qwen/Qwen2.5-14B-Instruct-GPTQ-Int4
Qwen2.5-14B	Qwen et al. (2024)	5B	GPTQ8	https://huggingface.co/Qwen/Qwen2.5-14B-Instruct-GPTQ-Int8
Qwen2.5-32B	Qwen et al. (2024)	32B	Full	https://huggingface.co/Qwen/Qwen2.5-32B-Instruct
Qwen2.5-32B	Qwen et al. (2024)	6B	AWQ	https://huggingface.co/Qwen/Qwen2.5-32B-Instruct-AWQ
Qwen2.5-32B	Qwen et al. (2024)	6B	GPTQ4	https://huggingface.co/Qwen/Qwen2.5-32B-Instruct-GPTQ-Int4
Qwen2.5-32B	Qwen et al. (2024)	10B	GPTQ8	https://huggingface.co/Qwen/Qwen2.5-32B-Instruct-GPTQ-Int8
Qwen2.5-72B	Qwen et al. (2024)	72B	Full	https://huggingface.co/Qwen/Qwen2.5-72B-Instruct
Qwen2.5-72B	Qwen et al. (2024)	12B	AWQ	https://huggingface.co/Qwen/Qwen2.5-72B-Instruct-AWQ
Qwen2.5-72B	Qwen et al. (2024)	12B	GPTQ4	https://huggingface.co/Qwen/Qwen2.5-72B-Instruct-GPTQ-Int4
Qwen2.5-72B	Qwen et al. (2024)	21B	GPTQ8	https://huggingface.co/Qwen/Qwen2.5-72B-Instruct-GPTQ-Int8

Table 3: Detailed information about used LLMs in our experiments.

Quantization	Qwen2.5-7B		Qwen2.5-14B		Qwen2.5-32B		Qwen2.5-72B		Llama3-8B		Llama3-70B	
	FR	Diff	FR	Diff	FR	Diff	FR	Diff	FR	Diff	FR	Diff
Full-Precision	0.362	0	0.568	0	0.624	0	0.562	0	0.442	0	0.514	0
AWQ	0.362	0	0.6092	0.0412	0.612	-0.012	0.546	-0.016	0.428	-0.014	0.544	0.03
GPTQ4	0.277	-0.085	0.531	-0.037	0.593	-0.031	0.511	-0.051	0.379	-0.063	0.472	-0.042
GPTQ8	0.283	-0.079	0.552	-0.016	0.599	-0.025	0.52	-0.042	-	-	-	-
bit4	0.278	-0.084	0.5691	0.0011	0.604	-0.02	0.538	-0.024	0.3618	-0.0802	0.4609	-0.0531
bit8	0.344	-0.018	0.552	-0.016	0.62	-0.004	0.554	-0.008	0.436	-0.006	0.481	-0.033
Average	-	**-0.0532**	-	**-0.00534**	-	**-0.0184**	-	**-0.0282**	-	**-0.0408**	-	**-0.0245**

Table 4: Faithfulness rate (FR) and the value differences between full-precision and quantized models on eSNLI.

task performance is reported in Table 8.

Results. The findings derived from Table 4 align with those reported in the main body of the paper. Quantization negatively impacts feature attribution faithfulness, as evidenced by the average faithfulness decline across all models (§5.1.1). Notably, no single quantization method outperforms the others or better preserves feature attribution faithfulness (§5.3). Lastly, consistent with the findings for NLEs and CFEs, feature attributions from smaller models exhibit more pronounced faithfulness degradation under quantization.

E Results with Additional Datasets

Table 5 displays the automatic evaluation results of natural language explanations on ECQA and counterfactual examples on IMDb.

F Annotation

Figure 10 displays the annotation guideline that we provide to human annotators. NLEs and CFEs are presented to annotators in the form of questionnaires. We use the Crowdee[7] crowdsourcing platform to recruit annotators, distribute the questionnaires, and store their responses. A total of 48 annotators were recruited, all of whom are native English speakers without requiring specific expertise in explainable AI (XAI). Each annotator was given 15 explanations, along with two evaluation dimensions (§4.3.1). Each explanation was evaluated by at least two annotators.

Table 6 summarizes the observed self-explanation degradation in terms of trustworthiness and coherence.

[7] https://www.crowdee.com/

Model	Preci-sion	CFE (IMDb) LFR ↑	PPL ↓	TS ↓	NLE (ECQA) BARTScore ↑	TIGERScore ↑	Model	Preci-sion	CFE (IMDb) LFR ↑	PPL ↓	TS ↓	NLE (ECQA) BARTScore ↑	TIGERScore ↑
Qwen2.5-7B	full	87.75	34.77	0.47	-4.75	**-0.60**	Llama3-8B	full	84.60	32.02	0.67	-4.44	**-0.73**
	bib4	87.20	**34.11**	0.58	-4.74	-1.11		bib4	87.00	35.58	**0.54**	**-4.42**	-0.75
	bib8	**88.20**	36.80	0.47	**-4.71**	-1.04		bib8	82.40	**31.60**	0.68	-4.43	-0.75
	gptq4	87.80	37.60	1.00	-4.82	-0.97		gptq4	76.40	33.57	0.77	-4.55	-1.07
	gptq8	86.27	39.46	1.00	-4.82	-0.94		gptq8	–	–	–	–	–
	awq	86.40	38.22	**0.43**	-4.76	-1.28		awq	88.20	38.39	**0.54**	-4.45	-0.87
Qwen2.5-14B	full	88.60	35.37	**0.47**	-4.57	**-0.21**	Llama3-70B	full	**89.29**	**38.74**	**0.21**	**-4.37**	**-0.15**
	bib4	88.80	35.19	0.62	**-4.51**	-0.33		bib4	86.00	46.67	0.30	-4.46	-1.36
	bib8	88.40	36.20	0.52	-4.55	-0.29		bib8	87.01	48.23	0.33	-4.39	-0.27
	gptq4	**90.00**	31.06	1.00	-4.61	-0.30		gptq4	84.89	42.78	0.24	-4.38	-0.76
	gptq8	**90.00**	**30.58**	1.00	-4.59	-0.28		gptq8	–	–	–	–	–
	awq	88.00	35.44	0.51	-4.57	-0.44		awq	86.60	40.43	0.31	-4.41	-0.55
Qwen2.5-32B	full	89.92	**37.08**	**0.25**	-4.14	**-0.59**	Qwen2.5-72B	full	**93.79**	**34.59**	**0.21**	**-3.77**	-0.41
	bib4	85.00	37.23	0.38	-4.29	-0.98		bib4	83.80	39.95	0.23	-3.94	-0.58
	bib8	82.20	39.95	0.28	-4.21	-0.76		bib8	82.20	40.15	0.22	-3.86	-0.51
	gptq4	**90.00**	40.08	1.00	-4.33	-0.61		gptq4	90.00	37.35	1.00	**-3.77**	-0.55
	gptq8	**90.00**	42.15	1.00	-4.22	-1.32		gptq8	90.00	39.82	1.00	-3.82	**-0.39**
	awq	81.60	40.85	0.28	**-4.13**	-0.74		awq	81.20	40.08	0.21	-4.13	-0.43

Table 5: Automatic evaluation results of CFEs and NLEs generated by `Llama3` (8B, 70B) and `Qwen2.5` (7B, 14B, 32B, 72B) models with full precision and different quantization methods. We use `FIZLE` on IMDb, evaluated by *label flip rate (LFR)*, *perplexity (PPL)*, and *text similarity (TS)*; and `ZeroCoT` on ECQA, evaluated by *BARTScore* and *TIGERScore*. **Bold** values indicate the best-performing approach for each model, while underlined values denote the best-performing quantization method.

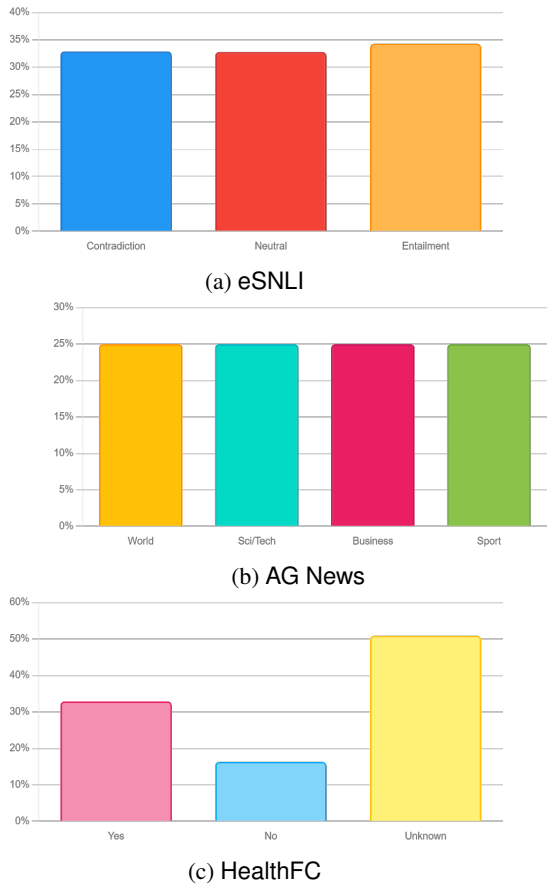


Figure 5: Label distributions of eSNLI, AG News and HealthFC.

Model	Metric	NLE Trust.	NLE Cohere.	CFE Trust.	CFE Cohere.
Qwen2.5-7B	full	**3.47**	**3.22**	**3.90**	**3.40**
	gptq4	3.06	3.14	3.20	2.84
	gptq8	2.96	2.63	3.26	2.86
	awq	2.98	2.96	2.69	2.49
Qwen2.5-32B	full	3.41	3.41	**3.50**	3.21
	gptq4	3.32	**3.60**	3.25	2.96
	gptq8	**3.51**	3.00	3.45	**3.26**
	awq	2.55	2.47	3.26	3.26
Qwen2.5-72B	full	**3.38**	3.50	**4.30**	**4.30**
	gptq4	3.21	3.23	3.04	2.61
	gptq8	2.97	2.86	2.92	3.02
	awq	2.92	**3.71**	3.40	3.37

Table 6: User study results for generated NLEs and CFEs on eSNLI, evaluated based on Trustworthiness (Trust.) and Coherence (Cohere.).

G LLM-as-a-Judge Evaluation

G.1 Setup

The adoption of LLMs as evaluators for complex tasks, referred to as "LLM-as-a-Judge", has gained popularity to perform evaluations by assigning quality scores in accordance with human intuition (Zheng et al., 2023; Huang et al., 2024). In addition to automatic and human evaluation, we investigate how well LLMs can quantitatively assess the explanation quality degradation caused by quantization. For this purpose, we select three open-source LLMs of varying sizes that are commonly used in the literature (Gu et al., 2025): `DeepSeek-R1` (DeepSeek-AI et al., 2025), `Gemma3-27B` (Team et al., 2025) and `GPT-OSS-120B` (OpenAI et al.,

Model	Precision	NLE						CFE					
Metric		Trustworthiness			Coherence			Trustworthiness			Coherence		
Judge Model		DS	OSS	Gemma	DS	OSS	Gemma	DS	OSS	Gemma	DS	OSS	Gemma
Qwen2.5-7B	full	4.93	4.67	3.67	5.00	4.86	4.87	2.73	1.73	2.40	3.47	2.67	3.13
	gptq4	4.73	4.00	3.53	4.86	4.13	4.67	2.87	1.73	2.60	3.93	2.67	3.33
	gptq8	4.93	4.53	3.67	5.00	4.67	4.80	3.40	1.93	2.87	3.93	2.80	3.73
	awq	5.00	4.80	3.60	5.00	5.00	4.87	2.79	1.67	2.87	3.93	2.87	3.60
Qwen2.5-32B	full	4.86	4.67	3.53	4.53	4.67	4.67	2.46	2.87	2.46	3.60	3.73	3.20
	gptq4	5.00	4.87	3.73	5.00	4.93	4.86	2.67	1.60	2.46	3.67	2.67	3.33
	gptq8	4.73	4.07	3.80	4.47	4.20	4.47	1.87	1.53	2.33	3.00	2.73	3.27
	awq	4.21	3.87	3.60	4.07	3.73	4.67	2.40	1.60	2.40	3.53	2.53	3.33
Qwen2.5-72B	full	5.00	5.00	3.67	4.93	4.80	4.87	2.20	1.67	2.66	3.20	2.93	3.47
	gptq4	5.00	4.87	3.47	5.00	4.80	4.80	2.46	1.73	2.66	4.13	2.93	3.53
	gptq8	5.00	4.80	3.60	5.00	4.80	4.80	2.73	1.73	2.53	3.46	2.67	3.47
	awq	4.93	4.73	3.60	5.00	4.67	4.73	3.27	1.67	2.73	4.00	2.67	3.53

Table 7: LLM-as-a-Judge evaluation on data examples selected for the user study using `DeepSeek-R1` (DS), `GPT-OSS-120B` (OSS), and `Gemma3-27B` (Gemma).

2025). Judge models assess the trustworthiness and coherence of self-explanations generated by LLMs with varying levels of precision (§4.3.1).

G.2 Results

The LLM-as-a-Judge evaluation results for data examples selected for human evaluation (§4.3.2) are displayed in Table 7.

G.2.1 Natural Language Explanations

Inter Judge Model Agreement. The within-category correlation analysis reveals notable differences in judge agreement across the two evaluation metrics (Figure 6). For **trustworthiness**, `DeepSeek-R1` and `GPT-OSS-120B` judges demonstrate strong agreement with a Pearson correlation of 0.862 and Spearman correlation of 0.938, both highly significant ($p < 0.001$). However, `Gemma3-27B` shows essentially no correlation with either `DeepSeek-R1` ($r = 0.012$) or `GPT-OSS-120B` (r = -0.088) when evaluating trustworthiness. The pattern shifts considerably for **coherence**, where all three judges show moderate to strong agreement with each other. `DeepSeek-R1` and `GPT-OSS-120B` maintain solid agreement ($r = 0.824$, $\rho = 0.743$), while `Gemma3-27B` now correlates moderately well with `DeepSeek-R1` ($r = 0.690$) and strongly with `GPT-OSS-120B` ($\rho = 0.877$). This indicates that while judges can reach reasonable consensus on what constitutes coherent output, they diverge substantially on trustworthiness assessments, with `Gemma3-27B` being the outlier.

Correlation with the User Study. Figure 7 shows the correlation between judge models and the user study. None of the correlations are statistically significant (all p-values > 0.05), indicating weak to moderate alignment between automated judge models and human user study evaluations. The highest correlation is `GPT-OSS-120B` for *Coherence* ($r = 0.498$, $p = 0.100$), which approaches but doesn't reach significance. `DeepSeek-R1` shows the highest correlation for *Trustworthiness* ($r = 0.507$, $p = 0.093$), also approaching significance. However, the Spearman correlations are even weaker, particularly for trustworthiness.

G.2.2 Counterfactual Explanations

Inter Judge Model Agreement. Figure 8 reveals that the CFE judgments show dramatically different correlation patterns compared to the NLE judgments. `DeepSeek-R1` and `GPT-OSS-120B` judges show essentially no correlation with each other for either metric. The only statistically significant correlation is `DeepSeek-R1` vs `Gemma3-27B` for trustworthiness ($r = 0.646$, $p = 0.023$). However, `GPT-OSS-120B` shows no meaningful correlation with either `DeepSeek-R1` or `Gemma3-27B` across both metrics. These findings highlight that judge alignment is task-dependent and cannot be assumed to generalize across different explanation types.

Correlation with the User Study. Figure 9 shows that all three judge models demonstrate weak or negative correlations with human evaluation, raising concerns about using these judge models for evaluating quantization's impact on counterfactual quality without careful calibration.

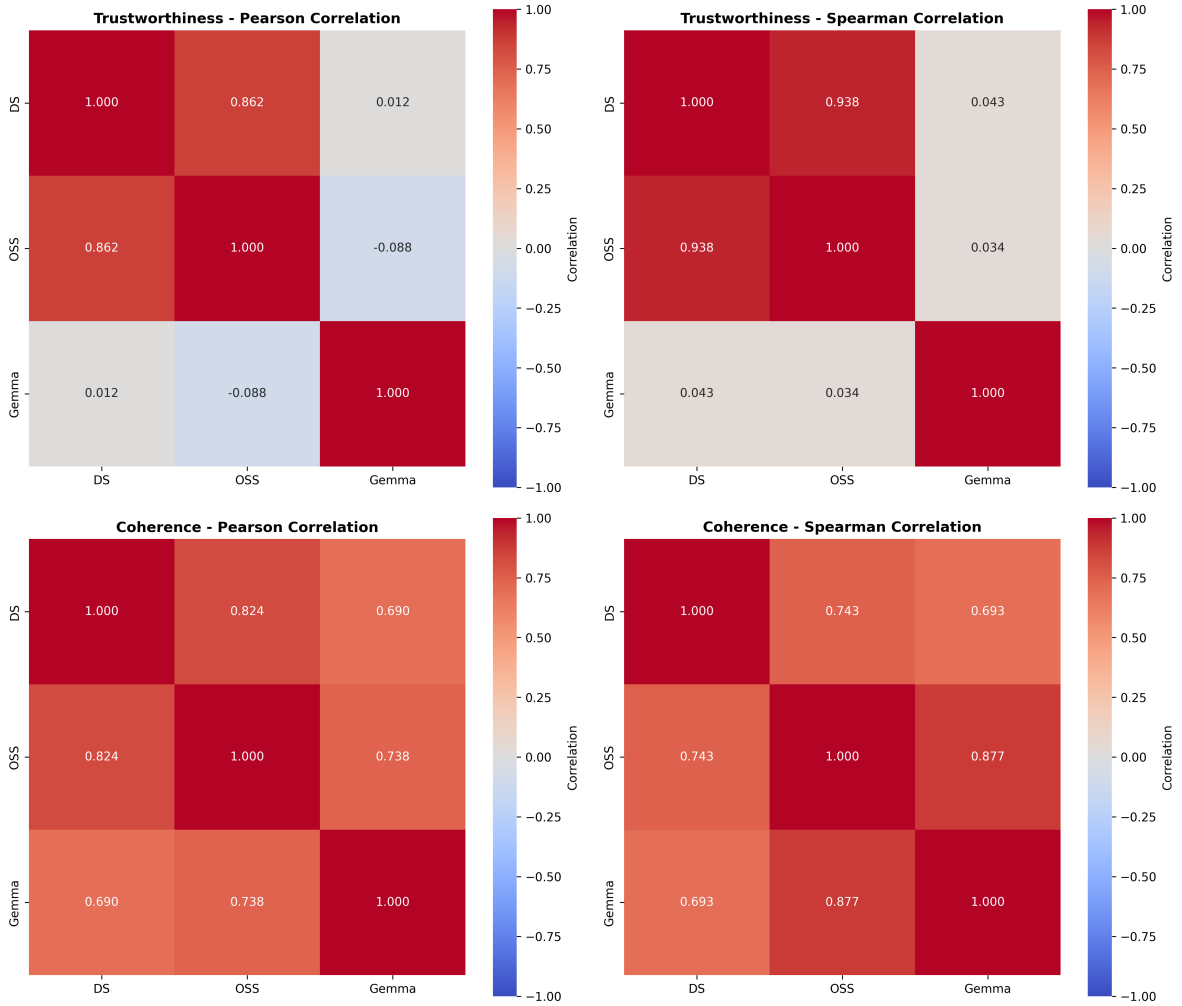


Figure 6: Pearson and Spearman correlation heatmaps for `DeepSeek-R1` (DS), `GPT-OSS-120B` (OSS), and `Gemma3-27B` (Gemma) for **natural language explanations** evaluated on *trustworthiness* and *coherence*.

H Task Performance

Table 8 illustrates the task performance of various quantization methods applied to deployed models on the eSNLI and HealthFC datasets.

Table 9 displays the Spearman correlation coefficient between task performance and natural language explanation quality, with values averaged across different quantization methods for each individual model. We find that, overall, the correlation between task performance degradation and self-explanation quality degradation is rather weakly positive and occasionally weakly negative. This indicates that task performance degradation contributes to self-explanation quality degradation to some extent.

I Quantization Method Ranking

We assign rankings to quantization methods for each model based on their results (Table 1, Table 8) and calculate the mean ranking across all experimental configurations. Figure 11 shows the quantization method ranking based on self-explanation quality and task performance, respectively. Quantization methods demonstrating superior preservation of full-precision LLM capabilities are assigned lower ranking values. We observe that while `AWQ` is optimal for preserving self-explanation quality and `GPTQ8` is suboptimal, `GPTQ8` is optimal for preserving task performance while `AWQ` is suboptimal. Furthermore, no quantization method can simultaneously be optimal in preserving self-explanation quality and task performance.

J Faithfulness

J.1 Faithfulness Metrics

We detail the specific faithfulness metrics utilized for evaluating natural language explanations in the subsequent discussion.

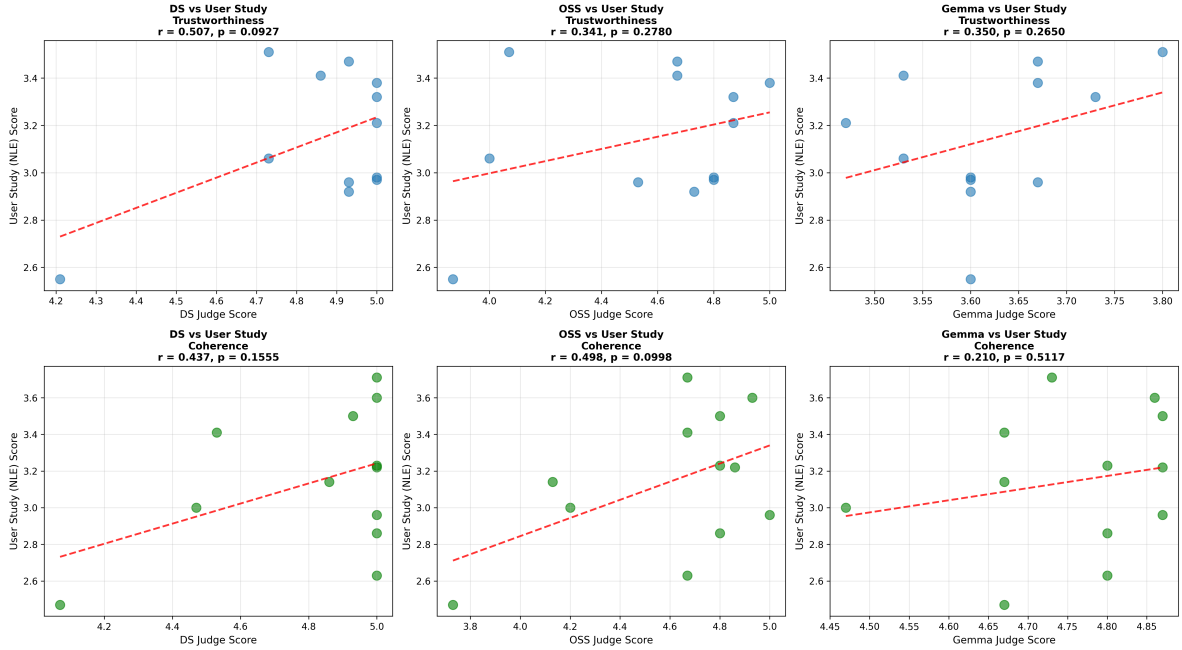


Figure 7: Correlations and significance between judge models (`DeepSeek-R1` (DS), `GPT-OSS-120B` (OSS), and `Gemma3-27B` (Gemma)) and user study for **natural language explanations** evaluated on *trustworthiness* and *coherence*.

Model	Precision	eSNLI	HealthFC	Model	Precision	eSNLI	HealthFC
Qwen2.5-7B	full	**87.40%**	**51.43%**	Qwen2.5-14B	full	69.50%	**45.14%**
	bib4	84.80%	36.57%		bib4	63.90%	40.29%
	bib8	86.70%	36.57%		bib8	67.70%	40.29%
	gptq4	86.80%	38.86%		gptq4	**78.50%**	36.57%
	gptq8	87.30%	39.71%		gptq8	67.90%	39.42%
	awq	86.10%	36.86%		awq	70.00%	46.29%
Qwen2.5-32B	full	88.40%	42.57%	Qwen2.5-72B	full	78.80%	51.43%
	bib4	89.50%	42.86%		bib4	**82.00%**	53.71%
	bib8	88.10%	**47.43%**		bib8	78.90%	**54.00%**
	gptq4	87.90%	41.43%		gptq4	78.00%	53.42%
	gptq8	88.60%	42.29%		gptq8	78.10%	52.09%
	awq	**90.60%**	39.43%		awq	77.70%	52.28%
Llama3-8B	full	34.53%	59.43%	Llama3-70B	full	60.46%	**67.14%**
	bib4	**37.44%**	39.43%		bib4	62.36%	63.71%
	bib8	33.73%	59.43%		bib8	38.44%	58.29%
	gptq4	27.33%	**63.14%**		gptq4	**64.26%**	61.71%
	awq	32.03%	62.00%		awq	64.26%	66.00%

Table 8: Task performance of all deployed models with different data type precisions on eSNLI and HealthFC.

Biasing Features. Turpin et al. (2023) examine the faithfulness of Chain-of-Thought (CoT) explanations that appear before the answer. Their methodology relies on introducing biasing features, such as "Suggested Answer" or "Answer is always A" in few-shot learning, or stereotype-inducing input edits, to the context. Unfaithfulness is established when the model's answer changes due to the bias, but the explanation does not explicitly state the bias as the reason for the decision (e.g., not generating a phrase like "Because you suggested A.").

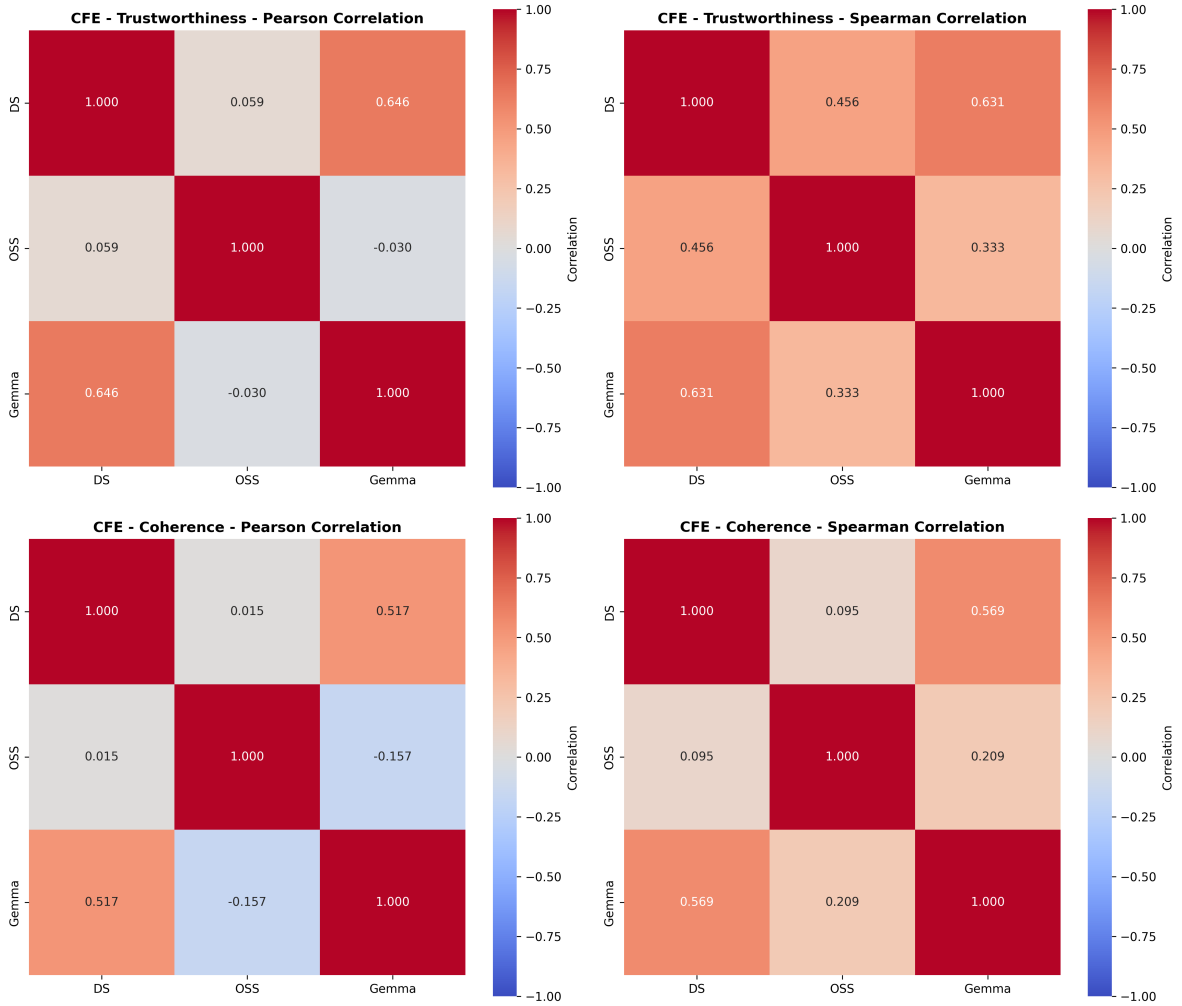


Figure 8: Pearson and Spearman correlation heatmaps for `DeepSeek-R1` (DS), `GPT-OSS-120B` (OSS), and `Gemma3-27B` (Gemma) for **counterfactual explanations** evaluated on *trustworthiness* and *coherence*.

Model	eSNLI	HealthFC
`Qwen2.5-7B`	0.13	-0.01
`Qwen2.5-14B`	-0.15	0.04
`Qwen2.5-32B`	-0.16	0.08
`Qwen2.5-72B`	0.13	0.03
`Llama3-8B`	0.02	0.05
`Llama3-70B`	0.11	-0.11

Table 9: Spearman correlation between the task performance and natural language explanation quality.

CC-SHAP. CC-SHAP assess the self-consistency of LLM explanations (Parcalabescu and Frank, 2024). It works by using SHAP values to compare how a model's input contributes to generating the predicted answer versus generating the explanation. The core idea is that a highly consistent explanation should rely on the same important input tokens as the prediction, allowing the method to measure the alignment between the input's importance for the answer and its importance for the explanation, all without needing to edit or perturb the model's input.

J.2 Results

Table 10 shows the natural language explanation faithfulness measured by biasing features. We observe that faithfulness is largely preserved, as evidenced by the predominant portion of instances maintaining their original state (faithful → faithful and unfaithful → unfaithful). Nevertheless, a greater number of cases exist in which natural language explanations become unfaithful due to quantization. More fine-grained faithfulness transitions are shown in Figure 12, 13, 14, 15, 16 and 17.

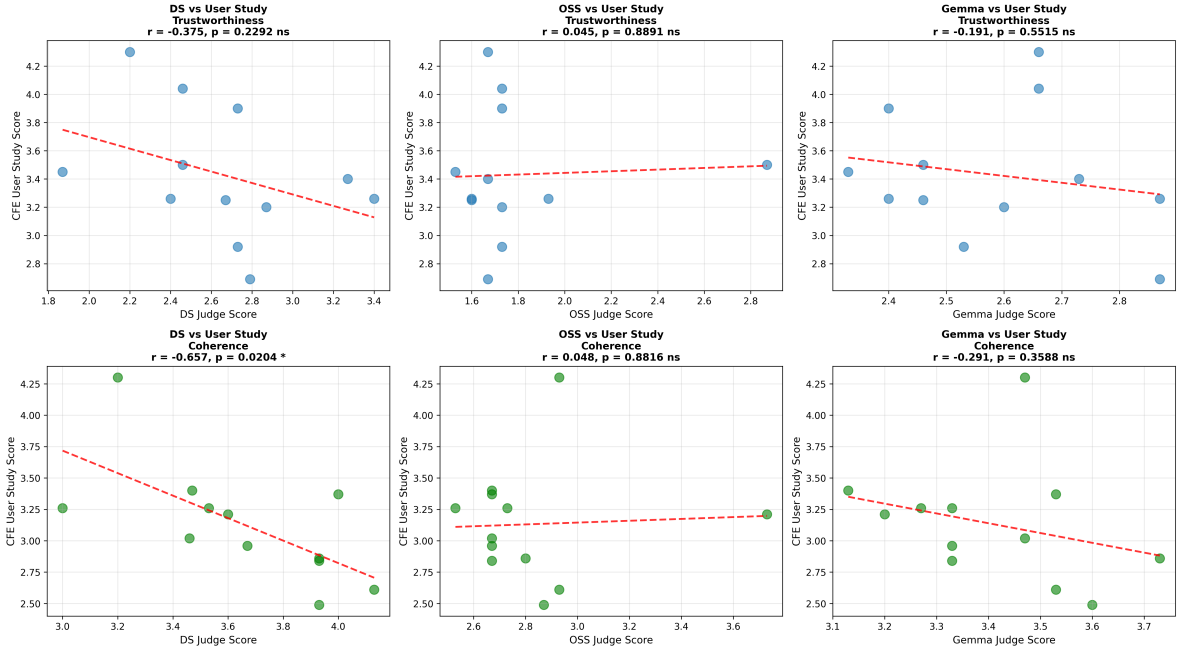


Figure 9: Correlations between judge models (`DeepSeek-R1` (DS), `GPT-OSS-120B` (OSS), and `Gemma3-27B` (Gemma)) and user study for **counterfactual explanations** evaluated on *trustworthiness* and *coherence*.

Quantization	F→N **(Degrade)**	F→F **(Maintain)**	N→N **(Maintain)**	N→F **(Improve)**
AWQ	7.11%	80.78%	7.20%	4.92%
GPTQ4	16.88%	70.32%	7.76%	5.05%
GPTQ8	4.57%	83.54%	8.52%	3.37%
bib4	8.20%	79.91%	5.52%	6.37%
bib8	5.57%	82.54%	8.24%	3.65%
Average	**8.47%**	**79.42%**	**7.45%**	**4.67%**

Table 10: Faithfulness transition rates (in %) measured by biasing features.

K Quantization Affects Explanation Quality Differently From Faithfulness?

Self-explanation quality and faithfulness are separate phenomena. We provide the following hypothesis regarding the different behavior of quality and faithfulness under quantization, i.e., under quantization, weights may be more sensitive compared to the decision boundary due to overparametrization, indicating that faithfulness can be better preserved compared to text quality. However, rigorously validating this hypothesis would constitute a substantial independent investigation beyond the scope of the present work.

Self-explanation quality, e.g., coherence or trustworthiness, depends heavily on the language modeling head and the token-by-token generation process, i.e., the precise logit distributions over vocabulary. These distributions are highly sensitive to small perturbations in weight values, because the softmax over a large vocabulary amplifies even minor shifts in pre-softmax scores. Quantization introduces exactly this kind of uniform noise across weights, degrading the fine-grained distributional properties that make text read coherent and well-structured (Resendiz and Klinger, 2025).

Faithfulness, by contrast, reflects whether an explanation causally tracks the model's actual decision-making process, contingent upon the decision boundary in the representation space. Quantization perturbs weights but tends to preserve the topology of these boundaries, particularly in over-parameterized models, where numerous weight configurations yield functionally equivalent deci-

Annotation Guideline for Human Evaluation

User Study Description:
Dear participants,
Thanks for attending our user study. This study focuses on evaluating model-generated explanations. We present two types of explanations:

- **Counterfactual Example (CFE):** A minimally edited version of the input text that results in a change in the model's prediction.
- **Natural Language Explanation (NLE)**: A textual justification generated by the model to explain its decision-making process for a given input.

Each explanation should be evaluated along the following two dimensions. Please assign a score to each dimension on a scale from 1 **(strongly disagree)** to 5 **(strongly agree)**.

- **Trustworthiness**: Evaluate whether the provided explanation is trustworthy and can be relied upon by humans;
- **Coherence**: Assess whether the provided explanation is sensible, clear, and coherent, and effectively captures the rationale;

Dataset Structure:
e-SNLI (Stanford Natural Language Inference): Each example consists of a premise and a hypothesis. The task is to determine the relationship between the two, categorizing it as either Entailment, Contradiction, or Neutral based on the information in the premise.

e-SNLI Example: {example}

Explanation: {example explanation}

Rating:

Trustworthiness: {score}

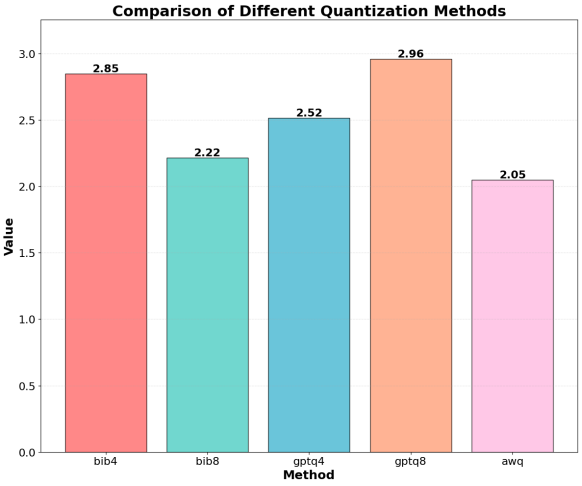
Coherence: {score}

Entailment means the hypothesis must be true if the premise is true. *Contradiction* means the hypothesis must be false if the premise is true. *Neutral* means the hypothesis might be true, or might not — we can't tell just from the premise.

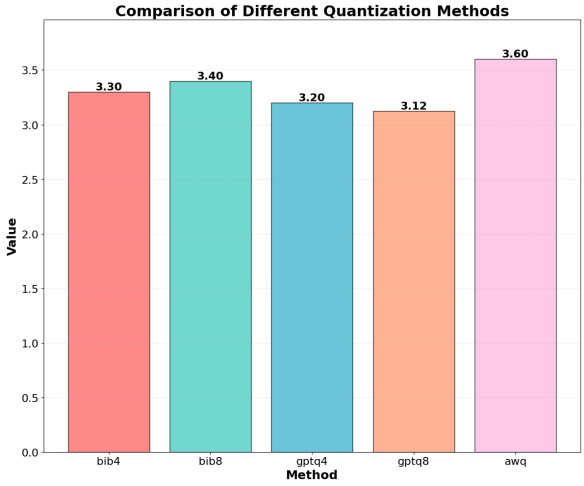
User Study Instruction:
You will be provided with 15 instances to evaluate. For the **counterfactual example** evaluation, each instance includes a premise–hypothesis pair. Your task is to evaluate only the quality of the **premise** according to the two dimensions described above. For the **natural language explanation** evaluation, each instance also includes a premise–hypothesis pair, along with a **model-generated justification** in natural language. In this case, your task is to evaluate the provided justification based on the same two dimensions.

Figure 10: Annotation Guideline.

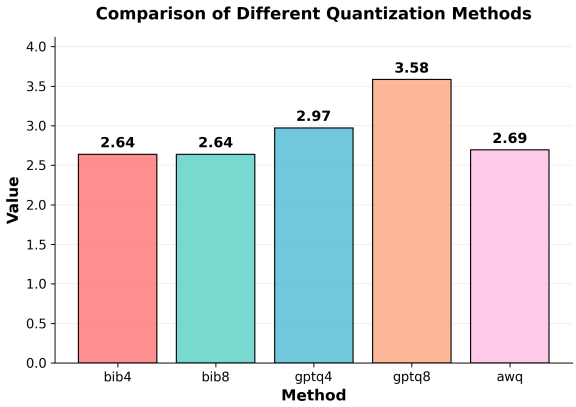
sion surfaces. Consequently, faithfulness exhibits lower sensitivity to quantization than explanation quality.



(a) Self-explanation quality

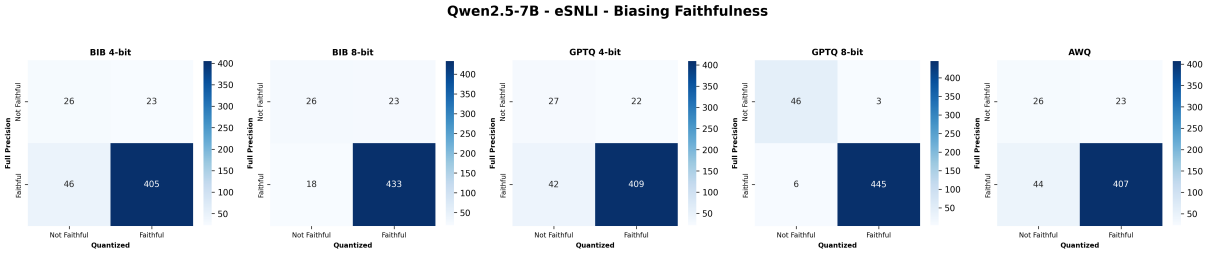


(b) Task performance

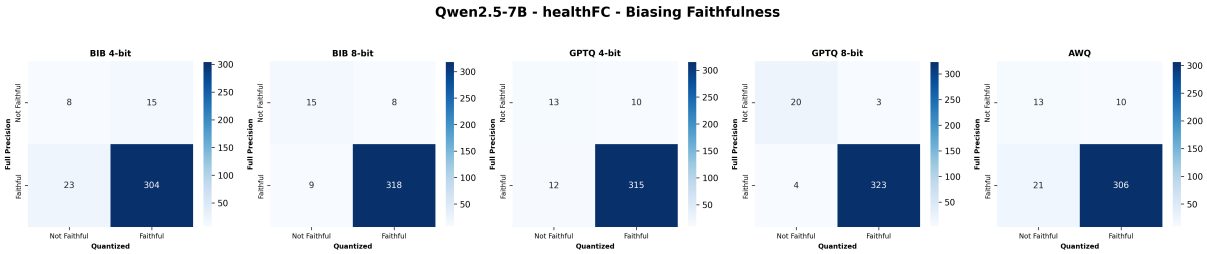


(c) Faithfulness

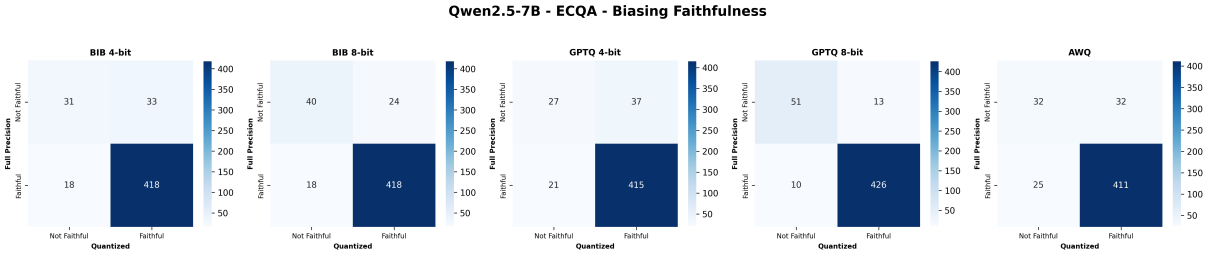
Figure 11: Quantization method ranking.



(a) eSNLI

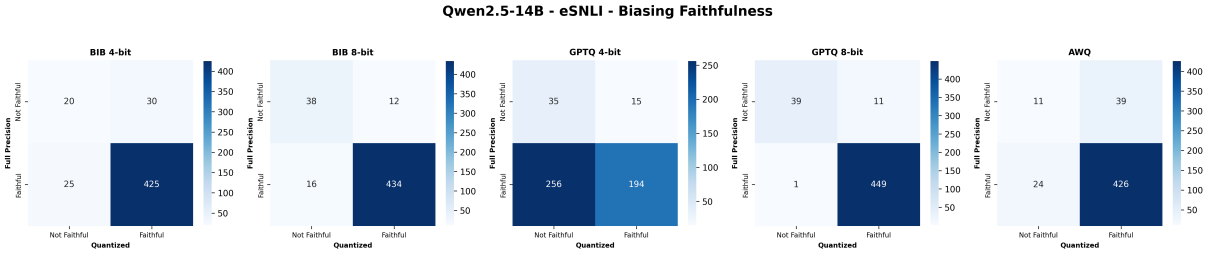


(b) HealthFC

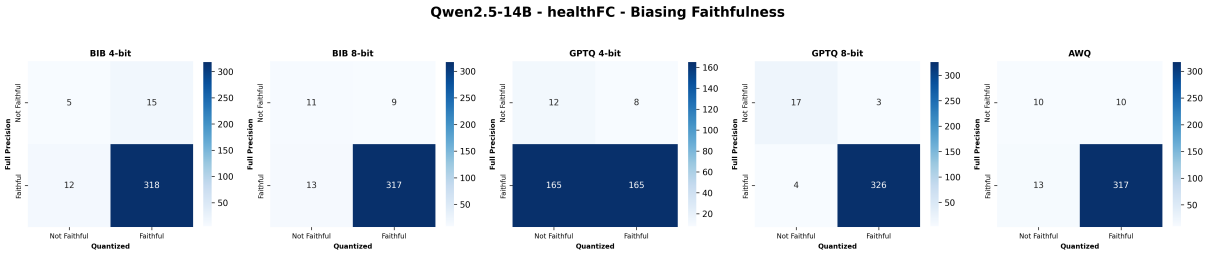


(c) ECQA

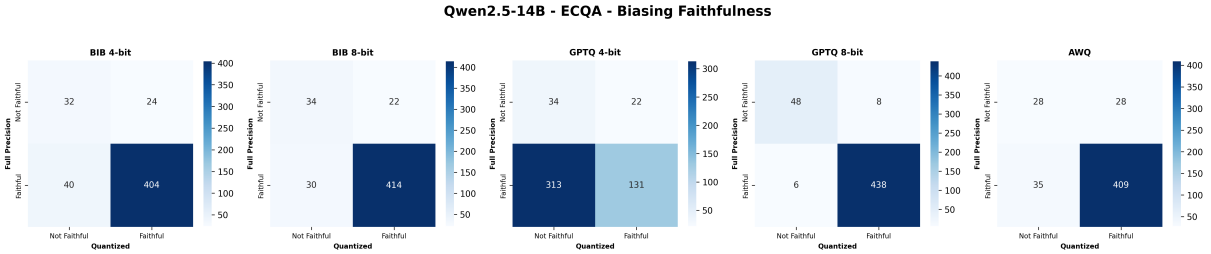
Figure 12: Faithfulness Variation `Qwen2.5-7B` (§5.3)



(a) eSNLI

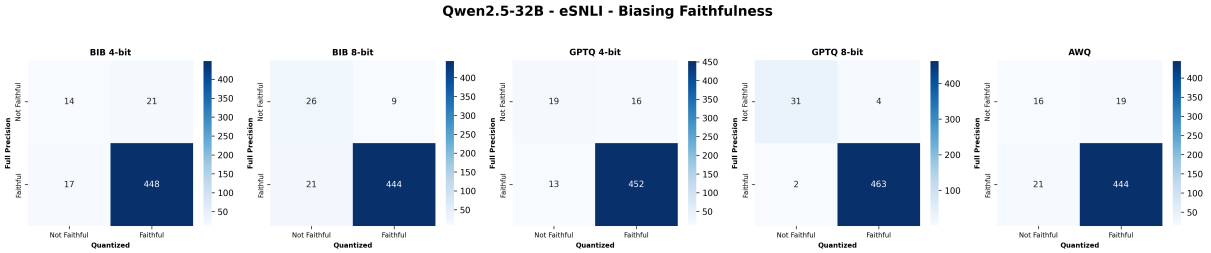


(b) HealthFC

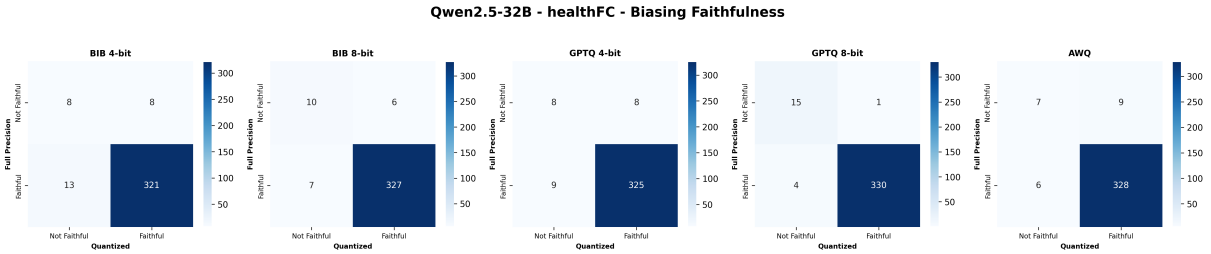


(c) ECQA

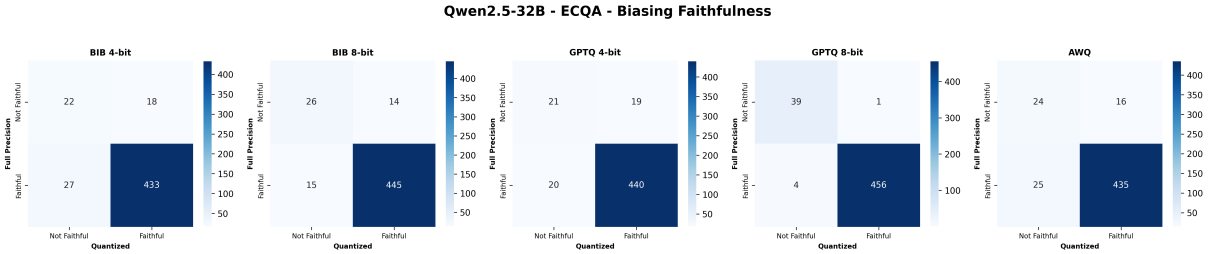
Figure 13: Faithfulness Variation `Qwen2.5-14B` (§5.3)



(a) eSNLI

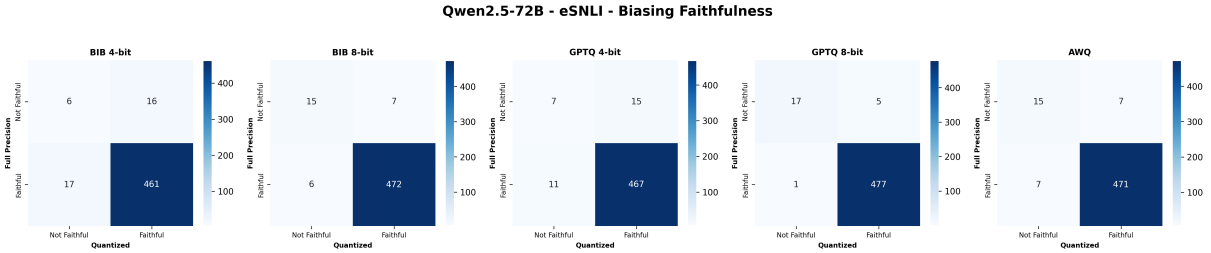


(b) HealthFC

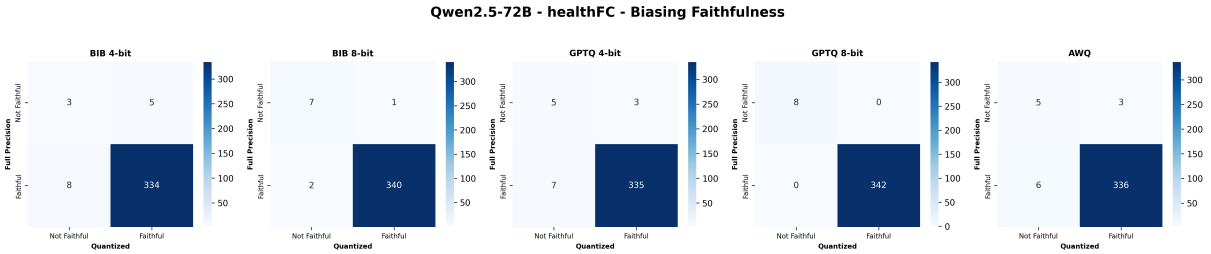


(c) ECQA

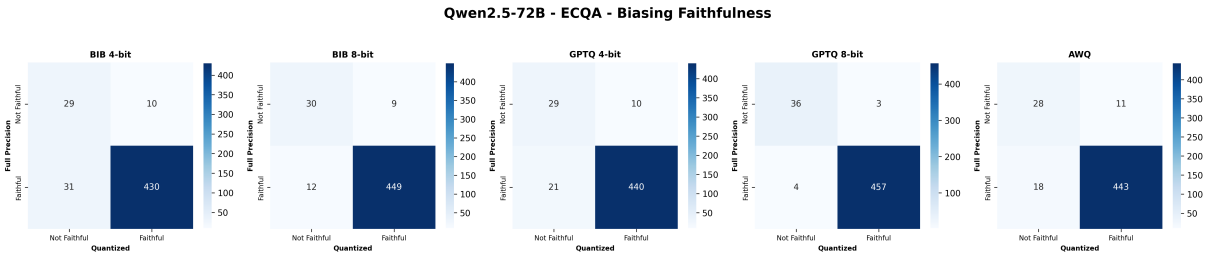
Figure 14: Faithfulness Variation `Qwen2.5-32B` (§5.3)



(a) eSNLI

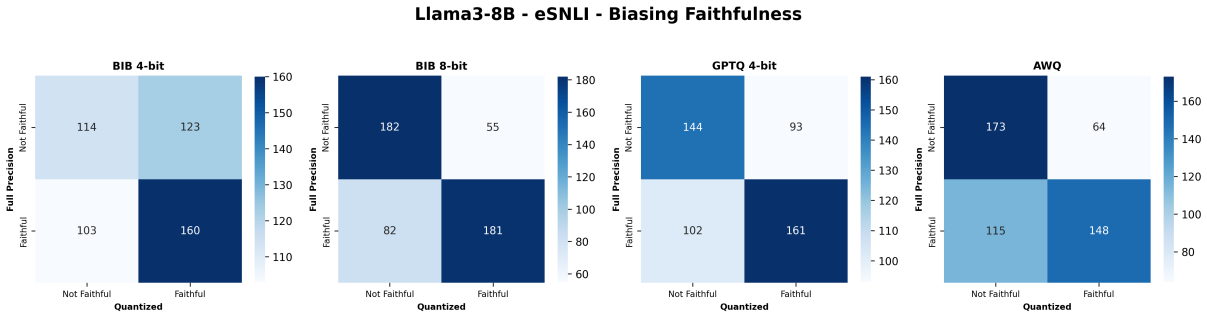


(b) HealthFC

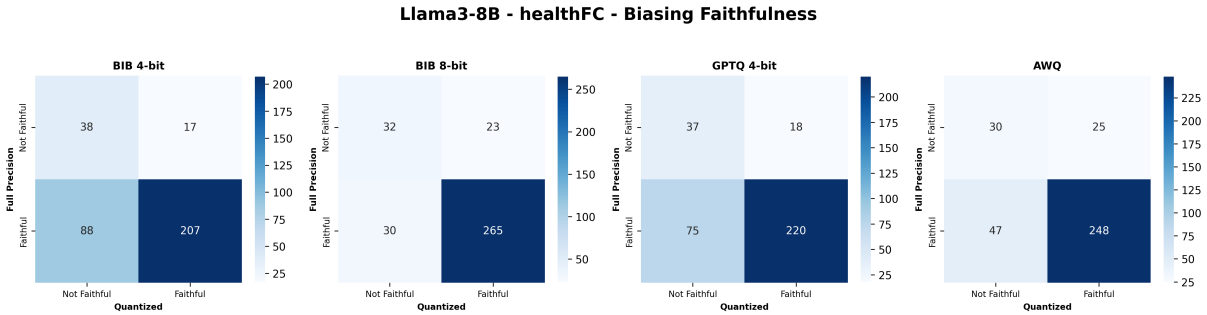


(c) ECQA

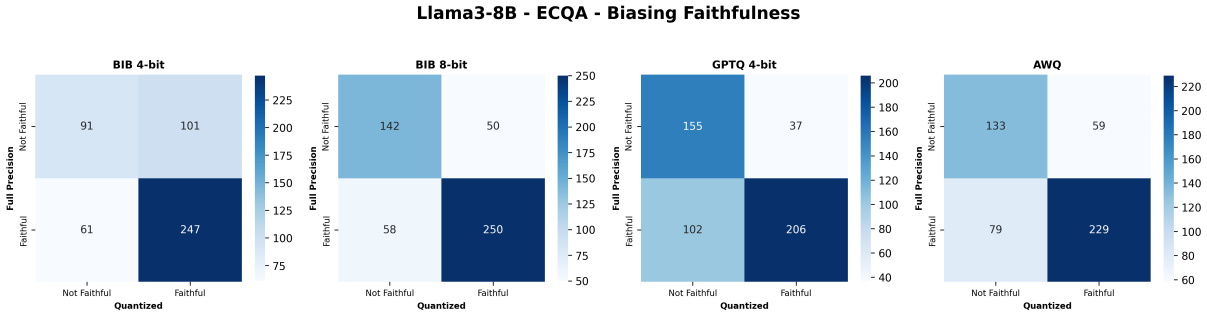
Figure 15: Faithfulness Variation `Qwen2.5-72B` (§5.3)



(a) eSNLI

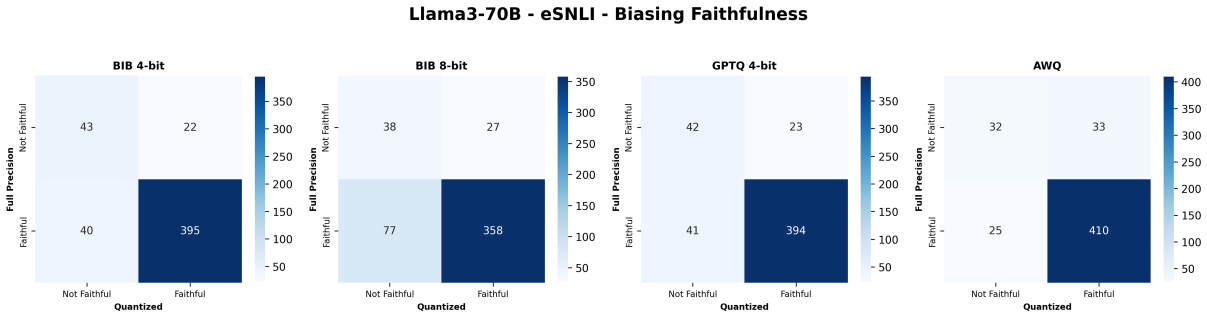


(b) HealthFC

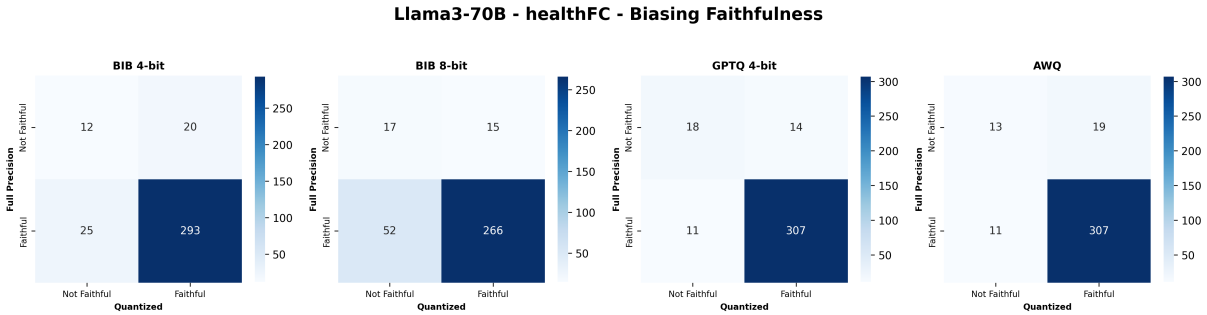


(c) ECQA

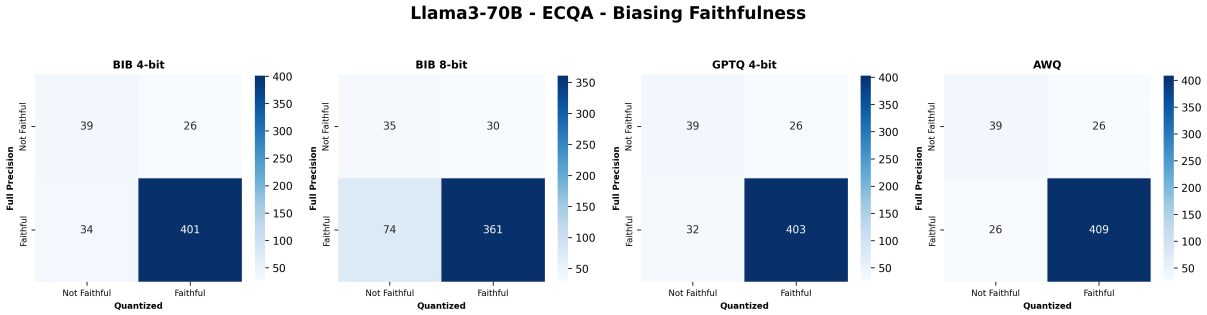
Figure 16: Faithfulness Variation `Llama3-8B` (§5.3)



(a) eSNLI



(b) HealthFC



(c) ECQA

Figure 17: Faithfulness Variation `Llama3-70B` (§5.3)