

When Quantization Preserves Accuracy but Not Evidence: Explanation-Aware Post-Training Quantization for Medical LLMs

Yeji Kim, Mi-Young Kim, Randy Goebel
University of Alberta

{yeji7, miyoung2, rgoebel}@ualberta.ca

Abstract

Post-training quantization (PTQ) enables efficient deployment of large language models, and PTQ methods are usually optimized and evaluated with generic reconstruction, perplexity, or answer accuracy. But in explanation-critical domains, preserving only the final answer may be insufficient, since users may also inspect generated rationales to judge whether a prediction is trustworthy. We study this issue in medical multiple-choice question answering, where rationales should provide evidence that supports the selected answer.

We propose an explanation-aware objective for transformation-based PTQ. Our method builds an offline faithfulness cache from full-precision teacher rationales and uses it during optimization to preserve answer-supporting evidence tokens and evidence-conditioned answer behavior. We instantiate it on OSTQuant under W4A4KV4 quantization and evaluate four 7B–8B medical and instruction-tuned LLMs on MedExQA, MedExpQA, and ChallengeClinicalQA. While a same-calibration OSTQuant baseline preserves task accuracy, it can substantially weaken answer-supporting rationales. Our objective is to preserve the full-precision model’s answer-supporting behavior rather than improve gold-label accuracy, and our method better preserves the full-precision model’s answer behavior and rationale-to-answer support. These results suggest that PTQ for explanation-critical settings should evaluate preservation of answer-supporting evidence, not only answer accuracy. Code and evaluation scripts are available at <https://github.com/dut0817/EAQuant>.

1 Introduction

Post-training quantization (PTQ) reduces the memory and inference cost of large language models (LLMs) by lowering the precision of weights, activations, and KV-cache states without full retraining. Standard LLM PTQ methods aim to preserve

Question: A 66-year-old patient with stage IV pancreatic adenocarcinoma is taking delayed-release morphine. He develops paresthesias and occasional lancinating pain in the right lumbar and periumbilical area that does not subside with rescue morphine. Neurological examination is normal. **Gold:** D - administer amitriptyline or gabapentin, dexamethasone, and increase morphine

FP16 Explanation: The patient’s symptoms of **paresthesias and lancinating pain** in the right lumbar and periumbilical area, despite **adequate morphine treatment**, suggest the presence of **neuropathic pain**. **Amitriptyline or gabapentin, dexamethasone, and increasing the dose of morphine** are effective in managing neuropathic pain, which is a common complication of pancreatic cancer.

Quantized Explanation: The patient’s symptoms are consistent with neuropathic pain, which is a common complication of pancreatic cancer. The patient’s history of delayed release **morphine and lactulose may contribute to the development of neuropathic pain**. The selected answer addresses the patient’s symptoms by adding adjuvant therapy with amitriptyline or gabapentin, dexamethasone, and increasing the dose of morphine, which is a combination of treatments commonly used to manage neuropathic pain.

Figure 1: Motivating example of rationale drift after quantization. Yellow marks answer-supporting evidence from the full-precision reference, and red marks an unsupported claim introduced after quantization.

full-precision behavior by optimizing against reconstruction error, perplexity, calibration losses, or downstream answer accuracy (Frantar et al., 2022; Xiao et al., 2023; Lin et al., 2024; Shao et al., 2024). These metrics are appropriate for generic compression fidelity, but they do not directly test whether a quantized model preserves the explanation behavior that users may inspect and rely on.

The difference between preserving what a model answers and preserving how it justifies that answer is especially consequential in explanation-critical domains. For example, in medical QA, users may not only ask whether a model selected the correct option, but also *whether the generated rationale cites the clinical evidence that supports that option*. Prior work distinguishes faithful explanations from plausible explanations and emphasizes that rationales should be meaningfully associated with model decisions, not merely sound convincing (Jacovi and Goldberg, 2020; Wiegreffe et al., 2021; Hase et al., 2020). Recent clinical-generation work

similarly stresses preserving important clinical information and avoiding unsupported content (Xie et al., 2024; Oukelmoun et al., 2025). Therefore a quantized model can be problematic even when answer accuracy appears stable. The quantized model may preserve the answer while changing, omitting, or fabricating the answer-supporting evidence.

Figure 1 illustrates this failure mode. The full-precision and quantized models select the same answer, but the quantized rationale changes the clinical justification. This motivates our central question: *when quantization preserves answers, does it also preserve the evidence that supports those answers?* A recent study of LLM self-explanations under quantization suggests that quantization can affect explanation quality and faithfulness across bit-widths and methods (Wang et al., 2026), but to our knowledge no prior work directly modifies the PTQ objective to preserve explanation behavior. We extend this inquiry to medical multiple-choice QA using MedExQA, MedExpQA, and ChallengeClinicalQA (Kim et al., 2024; Alonso et al., 2024; Chen et al., 2025).

We propose an explanation-aware objective for transformation-based PTQ. Our goal is to preserve the full-precision model’s answer-supporting behavior under quantization, rather than to improve gold-label accuracy. To avoid directly reinforcing known teacher errors, we construct the cache only from calibration examples for which the full-precision teacher’s prediction matches the gold answer. Our method is designed for PTQ pipelines with calibration-time transformation optimization. We instantiate it on OSTQuant (Hu et al., 2025), a strong transformation-based method for low-bit weight, activation, and KV-cache quantization. Our method keeps the calibration pool, deployed W4A4KV4 format, and fused inference path unchanged. During transformation optimization, it builds an offline faithfulness cache from full-precision teacher rationales and adds two extra preservation losses to the optimization objective. The cache selects answer-supporting evidence spans from the teacher rationale, such as phrases or clauses, that support the teacher’s selected answer and become less answer-supporting after quantization. The added losses preserve both evidence-token behavior and evidence-conditioned answer behavior, so that the quantized model better retains the answer-supporting role of the full-precision rationale.

We evaluate four 7B–8B medical and instruction-

tuned LLMs on MedExQA, MedExpQA, and ChallengeClinicalQA under W4A4KV4 quantization. The results show that the OSTQuant baseline can substantially weaken answer-rationale support even when task accuracy is preserved. In contrast, our method better preserves the full-precision model’s answer choices and rationale–answer consistency.

Our contributions are as follows:

- We expose a gap between task accuracy and rationale preservation in medical QA PTQ, where low-bit quantization can preserve task accuracy while degrading answer-supporting rationales.
- We introduce a teacher-rationale cache and explanation-aware transformation losses that preserve answer-supporting evidence tokens and rationale-conditioned answer behavior during calibration.
- We evaluate W4A4KV4 PTQ with answer and evidence metrics, showing improved preservation of full-precision answer behavior and rationale support over a calibration-matched baseline.

2 Related Work

LLM post-training quantization (PTQ) PTQ reduces memory and inference cost by quantizing weights, activations, and KV-cache states after training. generative pre-trained quantization (GPTQ) uses approximate second-order information for weight quantization (Frantar et al., 2022); SmoothQuant addresses the difficulty of quantizing activation outliers by rescaling channels and shifting part of the quantization burden from activations to weights (Xiao et al., 2023); activation-aware weight quantization (AWQ) uses activation statistics to identify and protect salient weight channels (Lin et al., 2024). More recent differentiable or transformation-based methods optimize quantization parameters or equivalent transformations during calibration, including OmniQuant (Shao et al., 2024), SpinQuant (Liu et al., 2025), and OSTQuant (Hu et al., 2025). All these methods mainly target numerical fidelity, perplexity, or task accuracy. Here we study whether the resulting quantized model also preserves answer-supporting rationale behavior.

Medical explanations and behavior-aware compression Medical QA benchmarks increasingly include rationales or explanations beyond answer labels (Kim et al., 2024; Alonso et al., 2024; Chen et al., 2025). Prior work distinguishes plausible explanations from faithful ones and argues that ratio-

Token group	Token %	Signal %
Numeric cues	0.67	0.38
Answer/option	1.80	0.99
Clinical cues	22.19	15.56
Generic/discourse	75.34	83.07

Table 1: Distribution of OSTQuant token-level fidelity signal in the diagnostic setting. Token % is the fraction of full-precision rationale tokens in each group. Signal % is the fraction of total token-level KL divergence assigned to each group.

nales should be meaningfully coupled with model decisions (Jacovi and Goldberg, 2020; Wiegrefe et al., 2021; Hase et al., 2020). Clinical generation work also emphasizes preserving important clinical information and avoiding unsupported content (Xie et al., 2024; Oukelmoun et al., 2025). Recent studies show that quantization can affect LLM self-explanations (Wang et al., 2026), and that PTQ objectives can be adapted to preserve behavior-critical properties such as safety alignment (Wee et al., 2025). We extend this direction to medical rationale preservation under low-bit PTQ.

3 Diagnostic Analysis

Before introducing our method, we examine which rationale tokens receive signal from the original OSTQuant token-level fidelity loss. For this diagnostic, we compute token-level Kullback–Leibler (KL) (FP vs. OSTQuant next-token distribution) on full-precision rationales generated by Llama-3-8B-Instruct on MedExpQA, and cluster tokens into numeric cues, answer/option tokens, clinical cues, and generic/discourse tokens. The full clustering procedure is described in Appendix H.

Table 1 shows that generic/discourse tokens account for 75.3% of rationale tokens but receive 83.1% of the KL signal, whereas clinical cues receive a smaller signal share (15.6%) than their token share (22.2%). Thus, the base fidelity objective is dominated by common discourse tokens rather than the sparse clinical cues that make a rationale answer-supporting. This mirrors the failure mode in Figure 1: a quantized model may preserve a fluent-looking rationale and even the same answer while changing the clinical justification. We therefore add explanation-aware losses that concentrate preservation pressure on answer-supporting clinical evidence.

4 Method

This section presents our explanation-aware objective for transformation-based PTQ. Our method targets PTQ pipelines that optimize transformation parameters during calibration and fuse them into the final low-bit model. We add calibration-time losses that preserve answer-supporting evidence tokens and evidence-conditioned answer behavior from full-precision teacher rationales. We instantiate the method on OSTQuant (Hu et al., 2025), a strong recent transformation-based PTQ baseline.

PTQ optimization setup Figure 2 shows the intervention point. The OSTQuant baseline performs a single calibration-time transformation-optimization pass with the base PTQ objective, then quantizes and fuses the optimized transformations into the deployed model. Our method uses two calibration-time passes. In Pass 1, we run the same standard OSTQuant procedure on the same calibration pool to obtain a fixed quantized probe M_p . We use this probe only to construct the offline cache. In Pass 2, with the cache fixed, we optimize the final transformation parameters using the base PTQ loss together with the rationale-token and rationale-conditioned recovery losses. The deployed architecture, W4A4KV4 format, and inference path are identical to the OSTQuant baseline.

Within a transformation-optimization pass, let M_ϕ^{fp} denote the transformed model before quantization, and let M_ϕ^{q} denote the same transformed model under fake quantization, which simulates the target low-bit quantization during calibration. Here, ϕ denotes the calibration-optimized transformation parameters. OSTQuant optimizes ϕ by matching the next-token prediction distributions of these two paths on calibration text. Let u denote a calibration text sequence sampled from a calibration pool $\mathcal{D}_{\text{cal}}$ (defined below). We denote this base objective as

$$\mathcal{L}_{\text{OST}} = \mathbb{E}_{u \sim \mathcal{D}_{\text{cal}}} \left[\ell_{\text{KL-Top}} \left(M_\phi^{\text{fp}}, M_\phi^{\text{q}}; u \right) \right], \quad (1)$$

Here, $\ell_{\text{KL-Top}}$ is the original OSTQuant KL loss computed over the top-K next-token probabilities of the quantization-disabled path M_ϕ^{fp} . It matches the fake-quantized path to the quantization-disabled transformed path. Our method keeps $\mathcal{L}_{\text{OST}}$ and adds explanation-preserving losses during the second optimization pass.

Both OSTQuant and our method use the same non-test medical calibration pool $\mathcal{D}_{\text{cal}}$. This controls for calibration-data effects, which can affect

post-training compression performance (Williams and Aletras, 2024). Let $x_i = (q_i, \mathcal{O}_i, y_i)$ be a medical multiple-choice example, where q_i is the question, $\mathcal{O}_i = \{o_i^1, \dots, o_i^{K_i}\}$ is the answer-option set with K_i options, and $y_i \in \{1, \dots, K_i\}$ is the gold option index. We distinguish two full-precision objects. M_0 is the original full-precision teacher used only to build the fixed offline cache. M_ϕ^{fp} is the quantization-disabled path of the transformed model during OSTQuant optimization.

Answer-option scoring To decide whether a rationale or evidence span supports an answer, we use deterministic forced-choice option scoring (Holtzman et al., 2021; Robinson et al., 2023) rather than free-form answer parsing, which can be sensitive to decoding and answer-format choices (Kumar, 2022; Biderman et al., 2024; Tsvilodub et al., 2024). For a model M , option text $o_i^k = (o_{i,1}^k, \dots, o_{i,L_{i,k}}^k)$, where $L_{i,k}$ is the number of tokens in option o_i^k , and prompt u , we define

$$a_{M,i}(k | u) = \frac{1}{L_{i,k}} \sum_{t=1}^{L_{i,k}} \log p_M(o_{i,t}^k | u, o_{i,<t}^k). \quad (2)$$

The predicted option is $\hat{y}_{M,i}(u) = \arg \max_k a_{M,i}(k | u)$. When a distribution over options is needed, we apply a softmax over option scores and denote it by $\pi_{M,i}(\cdot | u)$. For target option y_i , the answer margin is

$$m_{M,i}(u; y_i) = a_{M,i}(y_i | u) - \max_{k \neq y_i} a_{M,i}(k | u). \quad (3)$$

4.1 Faithfulness Cache Construction

After obtaining the fixed same-calibration quantized probe M_p in Pass 1, we build an offline faithfulness cache $\mathcal{A}$ from the shared non-test calibration pool. Its purpose is to identify which parts of a full-precision teacher rationale should receive additional preservation pressure under quantization. We use faithfulness cache as an internal term. Operationally, it stores evidence that supports the full-precision teacher’s selected answer rather than establishing clinical factuality or causal explanation faithfulness.

For each calibration example, the full-precision teacher M_0 selects an answer using Eq. 2 and then generates an answer-conditioned rationale r_i (Hasan et al., 2026). We keep only examples where the teacher prediction matches the gold option index y_i . Note that this filter is a reliability

check, not a supervised answer loss. The gold label is used only to discard cache entries where the teacher itself is wrong. This prevents the cache from preserving rationales for teacher mistakes.

We use a prompted Qwen2.5-72B-Instruct model (Qwen Team, 2024) to generate high-recall candidate spans from the teacher rationale. For each candidate span e , we measure how strongly it supports the teacher answer using forced-choice option margins. We compare the teacher-answer margin both when e is provided as evidence and when e is removed from the rationale. A span is retained as teacher-answer-supporting evidence if it provides strong support as defined in Appendix I for the teacher answer under these behavioral tests (DeYoung et al., 2020). We then repeat the same scoring with a same-calibration quantized probe M_p and define $\mathcal{Q}_i \subseteq \mathcal{F}_i$ as the spans whose answer-supporting effect is weakened after quantization. To avoid trivially derivable answers, direct answer-label cues are removed before evidence extraction. The prompts used for cache construction are provided in Appendix J.

Each retained cache entry is

$$\xi_i = (x_i, r_i, \mathcal{F}_i, \mathcal{Q}_i, \omega_i) \in \mathcal{A}.$$

Here, $\mathcal{F}_i$ is the set of teacher-answer-supporting evidence spans, $\mathcal{Q}_i \subseteq \mathcal{F}_i$ is the subset whose answer-supporting effect is weakened by the quantized probe, and ω_i is a normalized token-level weight vector supported on tokens covered by $\mathcal{Q}_i$. We use $\mathcal{Q}_i$ for the rationale-token loss (focusing on tokens most affected by quantization) and $\mathcal{F}_i$ for the rationale-conditioned recovery loss (preserving the answer-supporting behavior of all teacher-answer-supporting evidence). The full cache construction details are provided in Appendix I.

4.2 Explanation-preserving Transformation Optimization

During the second transformation optimization pass, the base loss $\mathcal{L}_{\text{OST}}$ is computed on the full calibration pool. As previously noted, for cache entries $\xi_i = (x_i, r_i, \mathcal{F}_i, \mathcal{Q}_i, \omega_i) \sim \mathcal{A}$, we add two explanation-preserving losses. Both losses compare the fake-quantized path M_ϕ^q against the quantization-disabled path M_ϕ^{fp} . They are not supervised answer losses. They preserve full-precision behavior on selected rationale evidence.

Rationale-token loss Let c_i be the rationale-generation prompt, and let $d_{i,t}$ be the next-token

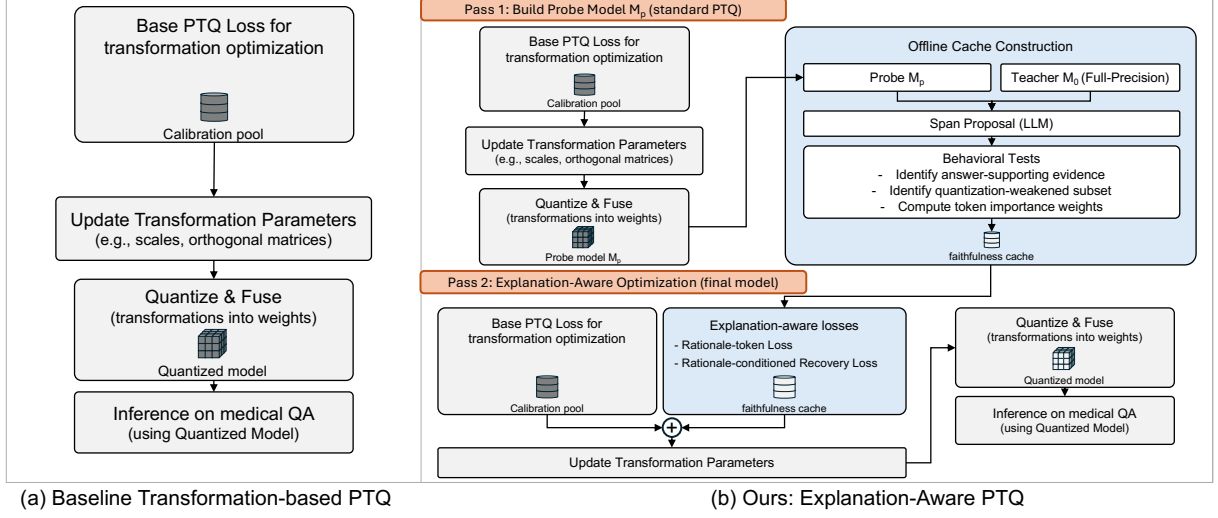


Figure 2: Overview of baseline transformation-based PTQ and our explanation-aware PTQ. Ours builds a quantized probe M_p and an offline cache before cache-conditioned final optimization, while keeping the deployed quantized model format and inference cost unchanged.

KL divergence between M_ϕ^{fp} and M_ϕ^{q} at prefix $(c_i, r_{i,<t})$:

$$d_{i,t} = \text{KL} \left(p_{M_\phi^{\text{fp}}}(\cdot \mid c_i, r_{i,<t}) \parallel p_{M_\phi^{\text{q}}}(\cdot \mid c_i, r_{i,<t}) \right).$$

The rationale-token loss is

$$\mathcal{L}_{\text{rat}} = \mathbb{E}_{\xi_i \sim \mathcal{A}} \left[\sum_{t=1}^{T_i} \omega_{i,t} d_{i,t} \right]. \quad (4)$$

This loss focuses token-level preservation on evidence that is both teacher-answer-supporting and quantization-affected, rather than imitating the entire rationale surface form.

Rationale-conditioned recovery loss Token preservation alone does not ensure that the same evidence induces the same answer preference after quantization. For each teacher-answer-supporting evidence span $e \in \mathcal{F}_i$, we build a recovery prompt $u_i(e)$. Let $h_{i,e}$ be the KL divergence between the option distributions of M_ϕ^{fp} and M_ϕ^{q} under this prompt:

$$h_{i,e} = \text{KL} \left(\pi_{M_\phi^{\text{fp}},i}(\cdot \mid u_i(e)) \parallel \pi_{M_\phi^{\text{q}},i}(\cdot \mid u_i(e)) \right).$$

The recovery loss is

$$\mathcal{L}_{\text{rec}} = \mathbb{E}_{\xi_i \sim \mathcal{A}} \left[\frac{1}{|\mathcal{F}_i|} \sum_{e \in \mathcal{F}_i} h_{i,e} \right]. \quad (5)$$

This loss preserves what the evidence does for the answer decision, rather than only what the rationale confirms lexically.

4.3 Overall Objective and Inference

The final objective is

$$\mathcal{L}_{\text{ours}} = \mathcal{L}_{\text{OST}} + \lambda_{\text{rat}} \mathcal{L}_{\text{rat}} + \lambda_{\text{rec}} \mathcal{L}_{\text{rec}}, \quad (6)$$

where λ_{rat} and λ_{rec} are hyperparameters controlling the contribution of the rationale-token and rationale-conditioned recovery losses. The base loss preserves generic next-token quantization fidelity on the full calibration pool. The two additional losses preserve answer-supporting rationale behavior on the faithfulness cache. After optimization, the optimized transformations are fused into the model weights and W4A4KV4 quantization is applied as in OSTQuant. The faithfulness cache, teacher model, gold labels, quantized probe, and evidence extractor are used only during calibration-time optimization. At inference time, the deployed model has the same quantization format, architecture, and runtime cost as the OSTQuant baseline.

5 Experiments

We evaluate whether explanation-aware transformation optimization preserves both medical QA performance and answer-supporting rationales under W4A4KV4 quantization.

5.1 Experimental Setup

Datasets We evaluate on MedExQA (Kim et al., 2024), MedExpQA (Alonso et al., 2024), and ChallengeClinicalQA (Chen et al., 2025). For quantization calibration, we construct a mixed pool of

128 non-test medical QA examples from the three datasets. The same calibration pool is used for OSTQuant and our method. Our method additionally builds the faithfulness cache from this pool, and the cache is used only during calibration-time transformation optimization. Dataset statistics and split details are in Appendix B.

Models and quantization We evaluate OpenBioLLM-8B (Pal and Sankarasubbu, 2024), Llama-3-8B-Instruct (Grattafiori et al., 2024), BioMistral-7B (Labrak et al., 2024), and Mistral-7B-Instruct (Jiang et al., 2023). For each model, the full-precision version is used as the reference teacher. The primary quantized systems use OSTQuant (Hu et al., 2025) under W4A4KV4 quantization. We set $\lambda_{\text{rat}} = \lambda_{\text{rec}} = \lambda$ to keep the hyperparameter search space tractable and select a single global $\lambda = 0.1$ on the MedExpQA validation split. This value is fixed for all datasets, models, and seeds. Appendix E reports sensitivity.

As an additional PTQ-method check, we also adapt the explanation-aware objective to OmniQuant (Shao et al., 2024) on Llama-3-8B-Instruct and Mistral-7B-Instruct across the three evaluation datasets. Full quantization, optimization, and hyperparameter settings are reported in Appendix C.

Compared systems We compare three systems. **Full precision** is the original unquantized model. **OSTQuant** is the same-calibration PTQ baseline using the original OSTQuant objective. **Ours** uses the same OSTQuant pipeline, calibration pool, W4A4KV4, and fused inference path, but adds the rationale-token and rationale-conditioned recovery losses during transformation optimization.

5.2 Evaluation Protocol

Answer accuracy and Full-precision agreement

Accuracy measures whether the selected option matches the gold answer. However, as previously noted, accuracy alone can be misleading under PTQ. Prior work on compressed LLM evaluation shows that aggregate accuracy can hide instance-level answer flips, where compressed models change predictions relative to the baseline even when benchmark accuracy remains similar or improves (Dutta et al., 2024). We report both answer accuracy and full-precision agreement. Full-precision agreement is the fraction of examples where a quantized system selects the same option as the full-precision reference.

Rationale-based answer reconstruction Answer accuracy and reference agreement do not show whether a generated rationale supports the model’s own answer. We therefore evaluate predicted-answer reconstruction, a simulatability-style test (Hase et al., 2020; Wiegrefe et al., 2021). For each system, we generate a concise rationale. We then remove direct answer-label cues and ask the same system to recover an answer from the question, answer options, and generated rationale. Predicted-answer reconstruction, or PredRec, is the fraction of examples where the recovered answer matches the system’s original selected answer. We also compute gold-answer reconstruction, or GoldRec, where the target is the gold answer. Because PredRec and $\mathcal{L}_{\text{rec}}$ use related forced-choice scoring, we also perform additional control experiments. Appendix F reports the full control results.

Full-precision evidence retention PredRec does not directly measure whether quantized rationales retain the clinical evidence used by the full-precision model. We therefore add Full-precision Evidence Retention (FER), inspired by the atomic-claim support style of FActScore (Min et al., 2023). Unlike FActScore’s external factuality evaluation, FER asks whether the system rationale preserves atomic clinical claims from the full-precision rationale for the same question. To focus on evidence retention rather than answer differences, we compute FER only on examples where full-precision, OSTQuant, and our method select the same answer.

For each example, we decompose the full-precision rationale into atomic clinical claims and judge whether each claim is supported by the system rationale in the context of the question and answer options. Let $\mathcal{C}_i^{\text{fp}}$ be the extracted full-precision claims, and let $J(c, r, q, \mathcal{O}) = 1$ if claim c is supported by rationale r under question q and options $\mathcal{O}$. We define

$$\text{FER}(s) = \frac{1}{N} \sum_{i=1}^N \frac{1}{|\mathcal{C}_i^{\text{fp}}|} \sum_{c \in \mathcal{C}_i^{\text{fp}}} J(c, r_i^s, q_i, \mathcal{O}_i). \quad (7)$$

FER measures retention of full-precision clinical evidence, not absolute medical factuality, and does not penalize additional unsupported claims. To assess GPT-judge reliability, we conduct a small human validation on 20 Llama-3/MedExpQA cases (126 claims, 252 candidate–claim judgments per annotator). Two annotators agree on 79.37% of judgments (Cohen’s $\kappa = 0.58$), and the automatic

judge agrees with the pooled human labels on 80.95% ($\kappa = 0.63$). The automatic judge is stricter in absolute support rate but shows the same directional gap between the two systems (+9.52 for humans versus +11.90 for the automatic judge). Full details are in Appendix K.

Unsupported-Claim Rate FER measures recall of evidence from the full-precision rationale, but it does not penalize additional claims introduced by a system rationale. We therefore report the Unsupported-Claim Rate (UCR) as a complementary metric. For each system rationale r_i^s , we extract atomic clinical claims and judge each claim relative to the full-precision rationale. The judge labels each claim as source-supported, valid background, unsupported added, contradicted, or not a claim. Let $\mathcal{C}_i^s$ denote the extracted claims receiving a valid claim judgment, excluding items labeled not a claim, and let $U(c, r_i^{\text{fp}}, q_i, \mathcal{O}_i) = 1$ if claim c is labeled unsupported added or contradicted relative to the full-precision rationale, and 0 otherwise. We define

$$\text{UCR}(s) = \frac{1}{N} \sum_{i=1}^N \frac{1}{|\mathcal{C}_i^s|} \sum_{c \in \mathcal{C}_i^s} U(c, r_i^{\text{fp}}, q_i, \mathcal{O}_i). \quad (8)$$

UCR measures the rate of claims that are unsupported by or contradictory to the full-precision rationale, with lower values indicating fewer unsupported additions. Like FER, UCR is reference-based and does not measure absolute medical factuality.

Pairwise rationale judging As a supplementary analysis, we conduct an anonymized LLM-as-a-judge pairwise comparison on MedExpQA. We focus on MedExpQA because it shows the largest answer–rationale degradation in Table 2, making rationale-level differences most measurable. We treat this analysis as supporting evidence rather than a primary metric because prior work suggests that LLM judges may miss subtle quantization-induced changes in self-explanations (Wang et al., 2026). To focus on rationale quality rather than answer differences, we restrict the comparison to cases where OSTQuant and our method select the same answer. The judge receives the question, answer options, selected answer, and two anonymized rationales, and chooses which rationale is more clinically supported. Full counts and qualitative examples are reported in Appendix L and Appendix M.

All evaluation prompts are provided in Appendix J. All LLM-based evaluations use GPT-5.4 (as of 2026-05).

6 Results and Analysis

All main results use $\lambda = 0.1$ for OSTQuant and 0.5 for OmniQuant. Quantized results are reported as mean $\pm$ standard deviation over three random seeds (0,1,2). Answer scoring is deterministic and rationales are generated with fixed decoding settings. Full-precision results use one deterministic run.

6.1 Answer Preservation and Rationale Reconstruction

Table 2 reports three complementary metrics. Accuracy measures correctness with respect to the gold answer. Full-precision agreement measures whether the quantized model preserves the answer chosen by the full-precision reference. PredRec measures whether the generated rationale supports the system’s own selected answer.

Accuracy and Full-precision answer agreement

On MedExpQA, OSTQuant raises average accuracy from 31.80 to 36.80 but reduces full-precision agreement to 48.66, suggesting that many gains reflect quantization-induced answer shifts rather than preservation of reference behavior. Our method keeps accuracy close to the full-precision reference and improves agreement to 79.47. Overall agreement also rises from 68.44 to 79.86.

Predicted-answer reconstruction

PredRec reveals the strongest rationale-level effect. On MedExpQA, OSTQuant’s average PredRec drops from the full-precision value of 90.40 to 56.54, even though its accuracy increases. *This means that a quantized model can achieve favorable answer accuracy while producing rationales that no longer support its selected answers.* Our method substantially reduces this failure, close to the reference. On MedExQA and ChallengeClinicalQA, OSTQuant causes less severe rationale degradation, but our method remains comparable or slightly better on average. These results support the central claim that preserving or improving final-answer accuracy does not guarantee preservation of answer-supporting rationales. Because our objective is preservation of full-precision behavior rather than maximization of gold-label accuracy, we treat accuracy as a secondary outcome. The higher OSTQuant accuracy on MedExpQA should therefore not be interpreted as stronger preservation.

Dataset	Model	Full-precision		OSTQuant			Ours		
		Acc.	PredRec	Acc.	FP Agreement	PredRec	Acc.	FP Agreement	PredRec
MedExQA	OpenBioLLM	38.62	86.81	34.82±1.23	77.77±2.35	83.48±0.16	36.45±1.05	81.03±0.71	86.60±1.81
	Llama-3	41.60	94.57	37.45±0.28	84.47±1.39	92.20±0.64	38.97±1.14	87.23±1.15	93.62±0.32
	BioMistral	58.72	78.94	55.57±0.93	78.48±0.96	77.52±3.82	56.17±1.15	78.19±0.21	78.23±2.83
	Mistral	64.47	94.68	60.57±1.66	79.50±0.59	87.52±3.96	60.14±1.38	79.57±1.08	89.86±1.72
	Avg.	50.85	88.75	47.10	80.06	85.18	47.93	81.50	87.08
MedExpQA	OpenBioLLM	24.80	91.20	23.20±2.12	38.40±3.49	39.73±5.21	24.00±1.39	82.13±6.47	86.40±4.23
	Llama-3	24.80	100.00	34.93±5.45	33.33±2.44	49.07±4.41	26.67±1.85	83.73±2.44	99.20±0.80
	BioMistral	34.40	76.80	41.87±1.67	63.20±1.39	57.87±1.67	32.27±1.67	76.27±2.44	74.67±3.23
	Mistral	43.20	93.60	47.20±2.77	59.73±3.33	79.47±2.81	43.20±0.80	75.73±4.62	93.60±4.23
	Avg.	31.80	90.40	36.80	48.66	56.54	31.54	79.47	88.47
ChallengeClin.	OpenBioLLM	25.00	90.91	24.24±2.76	76.62±0.86	86.47±3.62	24.13±1.14	79.76±0.50	89.07±3.47
	Llama-3	24.03	98.38	23.16±0.19	88.31±0.97	95.24±0.50	23.81±0.75	90.80±1.46	96.32±0.50
	BioMistral	29.55	66.23	29.11±2.30	71.10±1.17	69.70±1.23	29.00±0.50	71.86±2.52	69.48±5.96
	Mistral	33.77	97.40	31.39±0.37	70.35±1.67	93.29±2.48	30.09±2.62	71.97±2.16	95.13±0.56
	Avg.	28.09	88.23	26.98	76.60	86.18	26.76	78.60	87.50
Overall avg.		36.91	89.13	36.96	68.44	75.96	35.41	79.86	87.68

Table 2: Model-wise answer and rationale-preservation results under W4A4KV4 quantization. Full-precision agreement is 100 by construction and omitted for compactness. Bold indicates the better quantized system for reference agreement or PredRec.

6.2 Evidence Retention and Unsupported Claims

FER and UCR capture complementary rationale drift on MedExpQA same-prediction examples: FER measures retention of full-precision evidence, while UCR measures unsupported or contradictory additions.

Table 3 reports paired differences for FER and UCR on MedExpQA same-prediction examples. For FER, our method improves three of four model-level point estimates and raises the model average by +5.34 points, with a 95% CI of [1.50, 9.18]. For UCR, lower is better; our method lowers all four point estimates and reduces the model average by −4.65 points, with a 95% CI of [−9.10, −0.28]. BioMistral is the only FER exception. This likely reflects its already-low full-precision PredRec (76.8, the lowest among the four models) and weaker rationale structure, leaving less answer-supporting behavior to preserve. Its UCR still decreases, suggesting fewer unsupported additions despite no evidence-retention gain. These average intervals exclude zero, but most model-specific intervals do not, so we interpret the results as complementary evidence of reduced rationale drift on average rather than uniform per-model improvement. Both metrics remain reference-based and do not establish absolute medical factuality.

6.3 Additional Check with OmniQuant

We additionally adapt the objective to OmniQuant for Llama-3-8B-Instruct and Mistral-7B-Instruct.

Model	FER $\Delta \uparrow$	UCR $\Delta \downarrow$
OpenBioLLM	+5.92 [−3.09, 14.94]	−7.15 [−16.65, +2.37]
Llama-3	+12.35 [8.77, 15.94]	−2.35 [−13.62, +8.70]
BioMistral	−1.07 [−2.14, 0.01]	−6.56 [−13.96, +1.01]
Mistral	+4.15 [−6.32, 14.63]	−2.53 [−8.51, +3.68]
Avg.	+5.34 [1.50, 9.18]	−4.65 [−9.10, −0.28]

Table 3: FER and UCR differences on MedExpQA same-prediction examples. Values are Ours minus OSTQuant with 95% CIs. Positive favors Ours for FER, negative favors Ours for UCR.

Table 4 shows answer-behavior and rationale-support drift under OmniQuant. Adding our objective improves full-precision agreement and PredRec on all three datasets and partially recovers accuracy. Although the gains are smaller than in the OSTQuant setting, the results suggest that the proposed objective can improve rationale preservation beyond OSTQuant, with effectiveness depending on the PTQ pipeline and optimization interface.

7 Ablation Study

We conduct ablations on Llama-3-8B-Instruct with MedExpQA. Accuracy and PredRec are mean±standard deviation over three seeds, while FER is computed for one fixed seed on the corresponding same-prediction subset and reported with a 95% confidence interval. For FER, the same-prediction subset is constructed separately for each ablation table, so FER values are directly comparable within a table but may differ for the same

System	Acc.	FP Ag.	PredRec
Llama-3-8B-Instruct			
<i>MedExpQA</i>			
OmniQuant	23.37±0.34	28.79±2.29	47.91±3.47
Ours	29.50±0.62	55.67±1.55	57.45±0.98
<i>MedExpQA</i>			
OmniQuant	22.13±3.03	23.73±2.01	51.47±2.01
Ours	24.27±3.33	53.87±5.45	58.13±7.43
<i>ChallengeClin.</i>			
OmniQuant	20.78±2.34	24.03±2.58	48.59±1.60
Ours	23.81±1.63	50.97±2.03	57.58±2.72
Mistral-7B-Instruct			
<i>MedExpQA</i>			
OmniQuant	30.28±2.72	52.76±3.54	53.37±2.43
Ours	37.91±1.29	62.52±0.82	65.82±0.69
<i>MedExpQA</i>			
OmniQuant	25.60±2.88	47.91±4.55	49.87±4.03
Ours	31.20±2.40	66.36±3.48	66.93±0.92
<i>ChallengeClin.</i>			
OmniQuant	21.97±3.69	44.96±4.17	45.24±1.60
Ours	27.06±2.93	57.31±4.41	58.23±4.62

Table 4: Two-model OmniQuant check. Bold marks the better quantized system for FP agreement and PredRec.

System	Acc.	PredRec	FER
OSTQuant	34.93±5.45	49.07±4.41	53.25 [45.90, 60.58]
Full KD	27.20±2.12	95.20±1.60	56.30 [50.95, 61.51]
Random	24.80±0.80	97.87±1.22	55.99 [51.68, 60.12]
Ours	26.67±1.85	99.20±0.80	59.15 [54.68, 63.51]

Table 5: Simple rationale baselines on Llama-3/MedExpQA.

setting across tables.

Simple Rationale Baselines Table 5 compares against full-rationale KD and random span selection. All rationale-aware variants improve PredRec over OSTQuant, but they also reduce task accuracy, showing the accuracy–preservation trade-off in this setting. Among the rationale-aware variants, our method has comparable accuracy, the highest PredRec, and the highest FER. This suggests that evidence selection provides benefit beyond generic rationale distillation or arbitrary span selection.

Loss Component Ablation Table 6 ablates the two auxiliary losses. Either loss improves PredRec over the base objective, while accuracy remains in a similar range across the auxiliary-loss variants. FER separates their effects: $\mathcal{L}_{\text{rec}}$ alone retains more full-precision evidence than $\mathcal{L}_{\text{rat}}$ alone, and the combined objective obtains the highest observed

Objective	Acc.	PredRec	FER
Base	34.93±5.45	49.07±4.41	51.64 [40.41, 62.00]
+ $\mathcal{L}_{\text{rat}}$	24.80±0.80	95.73±2.44	46.01 [36.36, 55.79]
+ $\mathcal{L}_{\text{rec}}$	24.27±1.22	96.27±2.01	53.99 [45.93, 62.37]
Both	26.67±1.85	99.20±0.80	58.22 [49.33, 66.82]

Table 6: Loss-component ablation on Llama-3/MedExpQA.

Cache	Acc.	PredRec	FER
All spans	25.33±2.01	97.60±1.39	51.31 [46.58, 56.16]
+ support	24.80±1.60	97.07±1.67	51.15 [45.53, 56.75]
+ weakened	25.33±2.31	95.47±3.03	54.99 [49.69, 60.09]
+ margin	26.67±1.85	99.20±0.80	55.30 [50.41, 60.31]

Table 7: Evidence-selection ablation on Llama-3/MedExpQA.

FER and PredRec. This suggests that token-level preservation and rationale-conditioned answer recovery are complementary.

Evidence Selection Ablation Table 7 compares cumulative cache variants. PredRec is already high with all proposed spans, so additional filtering is not well distinguished by PredRec alone. FER shows that the largest gain appears after adding the quantization-weakened filter, and the full margin-drop criterion gives the highest observed FER and PredRec. Thus, the probe-based filtering stages help focus the cache on evidence most affected by quantization while maintaining strong rationale-to-answer consistency.

8 Conclusion

We studied whether low-bit PTQ preserves the evidence behind medical QA answers. Our results show that a calibration-matched OSTQuant baseline can keep aggregate accuracy similar while weakening full-precision agreement and rationale-to-answer support, suggesting that accuracy alone may mask explanation-level degradation. Our explanation-aware objective uses full-precision teacher rationales to preserve evidence-token and evidence-conditioned answer behavior during calibration. Across four 7B–8B models and three benchmarks, it better preserves answer behavior and answer-supporting rationales. The OmniQuant check further suggests broader use as a preservation regularizer, although recovery remains method-dependent. Overall, these findings motivate evaluating PTQ with evidence-preservation metrics, not only final-answer accuracy.

Limitations

This study focuses on medical multiple-choice QA and does not evaluate open-ended clinical generation, real-world clinical decision support, or patient-facing deployment. Although our method improves answer-rationale consistency, it does not guarantee clinical correctness, factuality, or complete explanation faithfulness. The faithfulness cache is built from full-precision teacher rationales, so it can inherit teacher errors or preserve teacher-specific reasoning patterns. The prompted extractor is used only as a candidate span proposer and is followed by behavior-based filtering, but extraction errors can still affect which evidence receives preservation pressure.

Our full evaluation primarily instantiates the method on OSTQuant under W4A4KV4 quantization with four 7B–8B models. We include an OmniQuant generalization check, but broader conclusions may differ for other PTQ backbones, bit-widths, model sizes, calibration pools, or non-medical domains. The pairwise LLM-as-a-judge analysis may also reflect judge-model preferences despite anonymization, a fixed rubric, and an explicit tie option. Finally, this work should be viewed as a compression and evaluation study for explanation-critical settings, not as evidence that any quantized model is safe for clinical use.

Ethical Considerations

This work studies compression and explanation preservation for medical QA models, but it is not intended for clinical decision-making or patient-facing deployment. Quantized models may preserve final answers while altering the rationale evidence supporting those answers, which could affect user trust and interpretation in high-stakes settings. Our results therefore highlight the importance of evaluating evidence preservation and rationale consistency; however, such evaluation does not guarantee clinical safety, factual correctness, or deployment readiness. All experiments are conducted offline on benchmark datasets intended for research use. We used AI assistants for limited writing and editorial support, such as improving clarity and readability. All technical content, experiments, analyses, and reported results were produced and verified by the authors.

Acknowledgements

This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) Collaborative Research and Training Experience (CREATE) program “From Data to Decision” (FD2D). This research was supported by the Alberta Machine Intelligence Institute (Amii), NSERC (including grants DGEER-2022-00369, RGPIN-2022-03469, and RGPIN-2025-05572), and Alberta Innovates (Enabling Better Health through Artificial Intelligence (AI-Better Health) Program). We thank Brett Siemens and Miron Nekhoroshkov for assisting with the human validation study.

References

- Iñigo Alonso, Maite Oronoz, and Rodrigo Agerri. 2024. MedExpQA: Multilingual benchmarking of large language models for medical question answering. *Artificial intelligence in medicine*, 155:102938.
- Stella Biderman, Hailey Schoelkopf, Lintang Sutawika, Leo Gao, Jonathan Tow, Baber Abbasi, Alham Fikri Aji, Pawan Sasanka Ammanamanchi, Sidney Black, Jordan Clive, and 1 others. 2024. Lessons from the trenches on reproducible evaluation of language models. *arXiv preprint arXiv:2405.14782*.
- Hanjie Chen, Zhouxiang Fang, Yash Singla, and Mark Dredze. 2025. [Benchmarking large language models on answering and explaining challenging medical questions](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3563–3599, Albuquerque, New Mexico. Association for Computational Linguistics.
- Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C Wallace. 2020. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 4443–4458.
- Abhinav Dutta, Sanjeev Krishnan, Nipun Kwatra, and Ramachandran Ramjee. 2024. Accuracy is not all you need. *Advances in Neural Information Processing Systems*, 37:124347–124390.
- Elias Frantar, Saleh Ashkboos, Torsten Hoefer, and Dan Alistarh. 2022. GPTQ: Accurate post-training quantization for generative pre-trained transformers. *arXiv preprint arXiv:2210.17323*.
- Gemma Team. 2026. [Gemma 4 technical report](#). *Preprint*, arXiv:2607.02770.

- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- H M Quamran Hasan, Housam Khalifa Bashier, Jiayi Dai, Mi-Young Kim, and Randy Goebel. 2026. [Reason2Decide: Rationale-driven multi-task learning](#). In *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 2309–2322, Palma, Mallorca, Spain. European Language Resources Association (ELRA).
- Peter Hase, Shiyue Zhang, Harry Xie, and Mohit Bansal. 2020. Leakage-adjusted simulatability: Can models generate non-trivial explanations of their behavior in natural language? In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4351–4367.
- Ari Holtzman, Peter West, Vered Shwartz, Yejin Choi, and Luke Zettlemoyer. 2021. Surface form competition: Why the highest probability answer isn’t always right. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7038–7051.
- Xing Hu, Yuan Cheng, Dawei Yang, Zukang Xu, Zhihang Yuan, Jiangyong Yu, Chen Xu, Zhe Jiang, and Sifan Zhou. 2025. OSTQuant: Refining large language model quantization with orthogonal and scaling transformations for better distribution fitting. *arXiv preprint arXiv:2501.13987*.
- 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.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L  lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, and William El Sayed. 2023. *Mistral 7B*. *Preprint*, arXiv:2310.06825.
- Yunsoo Kim, Jinge Wu, Yusuf Abdulle, and Honghan Wu. 2024. [MedExQA: Medical question answering benchmark with multiple explanations](#). In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 167–181, Bangkok, Thailand. Association for Computational Linguistics.
- Sawan Kumar. 2022. [Answer-level calibration for free-form multiple choice question answering](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 665–679, Dublin, Ireland. Association for Computational Linguistics.
- Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. BioMistral: A collection of open-source pretrained large language models for medical domains. In *Findings of the association for computational linguistics: acl 2024*, pages 5848–5864.
- 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 LLM compression and acceleration. In *MLSys*.
- Zechun Liu, Changsheng Zhao, Igor Fedorov, Bilge Soran, Dhruv Choudhary, Raghuraman Krishnamoorthi, Vikas Chandra, Yuandong Tian, and Tijmen Blankevoort. 2025. SpinQuant: LLM quantization with learned rotations. In *International Conference on Learning Representations*, volume 2025, pages 92009–92032.
- Sewon Min, Kalpesh Krishna, Xinxu Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100.
- Achir Oukelmoun, Nasredine Semmar, Ga  l de Chalendar, Cl  ment Cormi, Mariame Oukelmoun, Eric Vibert, and Marc-Antoine Allard. 2025. Detecting omissions in LLM-generated medical summaries. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 325–337.
- Ankit Pal and Malaikannan Sankarasubbu. 2024. OpenBioLLMs: Advancing open-source large language models for healthcare and life sciences. <https://huggingface.co/aaditya/OpenBioLLM-Llama3-70B>.
- Qwen Team. 2024. [Qwen2.5: A party of foundation models](#).
- Joshua Robinson, Christopher Michael Rytting, and David Wingate. 2023. [Leveraging large language models for multiple choice question answering](#). *Preprint*, arXiv:2210.12353.
- Wenqi Shao, Mengzhao Chen, Zhaoyang Zhang, Peng Xu, Lirui Zhao, Zhiqian Li, Kaipeng Zhang, Gao Peng, Yu Qiao, and Ping Luo. 2024. OmniQuant: Omnidirectionally calibrated quantization for large language models. In *International Conference on Learning Representations*, volume 2024, pages 45472–45496.
- Polina Tsvilodub, Hening Wang, Sharon Grosch, and Michael Franke. 2024. Predictions from language models for multiple-choice tasks are not robust under variation of scoring methods. *arXiv preprint arXiv:2403.00998*.

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

Sunghyun Wee, Suyoung Kim, Hyeonjin Kim, Kyomin Hwang, and Nojun Kwak. 2025. Alignment-aware quantization for LLM safety. *arXiv preprint arXiv:2511.07842*.

Sarah Wiegrefe, 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.

Miles Williams and Nikolaos Aletras. 2024. On the impact of calibration data in post-training quantization and pruning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10100–10118.

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

Yiqing Xie, Sheng Zhang, Hao Cheng, Pengfei Liu, Zelalem Gero, Cliff Wong, Tristan Naumann, Hoifung Poon, and Carolyn Rose. 2024. [DocLens: Multi-aspect fine-grained evaluation for medical text generation](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 649–679, Bangkok, Thailand. Association for Computational Linguistics.

A Artifact licenses and terms

We use all artifacts only for non-redistributive offline research experiments. MedExQA is released under CC BY-NC-SA 4.0, and MedExpQA is released under CC BY 4.0. For ChallengeClinicalQA, we use the released Medbullets files from the public repository. The evaluated model checkpoints are used under their corresponding terms: OpenBioLLM-8B and Llama-3-8B-Instruct under the Meta Llama 3 license, BioMistral-7B and Mistral-7B-Instruct-v0.1, and Gemma-4-12B-IT under Apache-2.0, and Qwen2.5-72B-Instruct under the Qwen license. OSTQuant and OmniQuant are used under Apache-2.0 and MIT license, respectively, and GPT-based evaluation is performed through the OpenAI API under the applicable service terms.

Dataset	Train/Pool	Dev	Test	Calib. used
MedExQA	–	25	940	25
MedExpQA	434	63	125	51
ChallengeClin.	308 (4-opt.)	–	308 (5-opt.)	52

Table 8: Dataset statistics and calibration-pool composition. A total of 128 non-test calibration examples were used.

B Dataset and Calibration Details

Table 8 summarizes datasets, evaluation splits, and calibration-pool composition. All calibration examples are sampled from non-test portions of the corresponding datasets. No test example is used for transformation optimization, cache construction, hyperparameter selection, or prompt development. For ChallengeClinicalQA, we use only the Medbullets subset, and the calibration pool is drawn from the four-option Medbullets subset and evaluation is performed on the disjoint five-option Medbullets subset.

Calibration-pool construction We sample 128 non-test examples from the three datasets. The same 128 examples are used by OSTQuant and our method. OSTQuant uses this pool for the base transformation-optimization loss. Our method uses the same pool for the base loss and additionally builds a faithfulness cache.

C Quantization and Optimization Details

This section reports the quantization and calibration-time optimization settings for the primary OSTQuant experiments and the additional OmniQuant check. OSTQuant and our method use the same calibration pool, deployed W4A4KV4 format, and fused inference path. The OmniQuant check uses the same mixed medical calibration pool, but evaluates a W4A4 fake-quantized inference path, where the weights and activations are quantized-dequantized during forward passes using learned OmniQuant parameters.

Shared explanation-aware objective settings

Table 9 lists the settings for the faithfulness cache and the explanation-aware objective. These components are used only during cache construction and calibration-time optimization, and are discarded before inference. The only objective hyperparameters that differ between the OSTQuant-backed and OmniQuant-backed runs are the two auxiliary loss weights.

Item	Value
Rationale-token loss weight λ_{rat}	0.1 for OSTQuant, 0.5 for OmniQuant
Recovery loss weight λ_{rec}	0.1 for OSTQuant, 0.5 for OmniQuant
Faithfulness-cache max sequence length	640
Teacher-answer-support threshold η_F	0.02
Probe-weakening threshold η_Q	0.02
Evidence-support sufficiency weight α	1.0
Margin-drop weight β	0.01
Span proposer	Qwen2.5-72B-Instruct
Maximum evidence units	6
Maximum evidence words	20
Minimum evidence words	3

Table 9: Shared settings for the explanation-aware objective and faithfulness cache. Only the auxiliary loss weights differ between the OSTQuant-backed and OmniQuant-backed runs.

OSTQuant-backed primary experiments Table 10 lists the quantization and calibration-time optimization settings shared by OSTQuant and our OSTQuant-backed method. The sequence length and calibration-sample count refer to calibration-time transformation optimization, not to a restriction on downstream inference length.

OmniQuant-backed additional check Table 11 lists the settings for the OmniQuant baseline and our OmniQuant-backed variant. This check is intended as a PTQ-method transfer check for the preservation objective, not as a runtime-equivalent comparison to the packed W4A4KV4 OSTQuant setting.

D Computational Budget and Infrastructure

All local experiments were run on NVIDIA L40S GPUs with 48GB memory. Quantization and transformation-optimization runs generally used 4 GPUs per run, while rationale generation, answer scoring, and label-recovery jobs used 1 GPU per run. Qwen2.5-72B-Instruct extraction stages were run locally with HuggingFace inference, typically on 4 GPUs. LLM-as-a-judge evaluation stages, including FER, UCR, and pairwise rationale judging, used the GPT-5.4 API.

Representative calibration cost To make the additional calibration cost of our method explicit, Table 12 reports one representative Llama-3-8B-Instruct run. Standard OSTQuant requires 3.11 GPU-hours. Our two-pass procedure first uses the

Item	Value
Evaluated format	W4A4KV4
Calibration data	128 mixed non-test medical QA examples, sequence length 512
Base objective	Top- K KL matching between the quantization-disabled transformed path and the fake-quantized path
Implementation loss	kl_top, post_attn=True
Weight quantization	GPTQ, 4-bit symmetric, group size -1 , MSE-based clipping
GPTQ damping	0.01
Activation quantization	4-bit dynamic per-token asymmetric
K/V cache quantization	4-bit, group size 128
Unquantized paths	Residual, attention, and output paths kept in 16-bit, following OSTQuant
Computation dtype	BF16
Optimized transformations	Global rotation, online Q/K Hadamard, O/V rotation, and smoothing
Smoothing modules	Q/K, O/V, up/down, and nonlinear smoothing
Optimizer	Riemannian Adam
Optimization steps	Maximum 100
Learning rates	0.0169 for rotation, 0.00179 for smoothing
Smoothing momentum	0.9
Gradient checkpointing	Enabled

Table 10: Quantization and calibration-time optimization settings for the OSTQuant primary experiments. Both OSTQuant and our method use the same W4A4KV4 format, calibration pool, and fused inference path.

same-calibration OSTQuant run to obtain the fixed quantized probe M_p , then spends 1.66 GPU-hours on span proposal and cache construction, followed by 7.83 GPU-hours for the cache-conditioned transformation optimization. The complete procedure therefore requires 12.60 GPU-hours, corresponding to an additional 9.49 GPU-hours relative to standard OSTQuant.

Reuse and inference cost The quantized probe and faithfulness cache are constructed once per model and reused across random seeds. Subsequent seeds therefore require only the final cache-conditioned optimization stage. LLM-as-a-judge evaluations are performed through the GPT-5.4 API and are not included in the local GPU-hour totals. All probe, teacher, extractor, and cache components are discarded after calibration. Consequently, the deployed model retains the same W4A4KV4 format, architecture, and inference cost as the OSTQuant baseline.

Item	Value
Quantization format	W4A4
Calibration data	128 mixed non-test medical QA examples, sequence length 512
Base objective	Layer-wise block-output reconstruction
Implementation loss	MSE between full-precision and fake-quantized layer outputs
Weight quantization	4-bit asymmetric per-channel affine, no group size
Learned weight clipping	Enabled (LWC)
Activation quantization	4-bit dynamic per-token asymmetric
Attention operand quantization	Q/K/V matmul operands dynamically fake-quantized to 4-bit
K/V cache quantization	No separate packed K/V-cache quantization; cache stored in the FP16 path
Compute and attention	FP16 / AMP, eager attention, attention probabilities kept 16-bit
Optimized transformations	LET smoothing scales and LWC clipping
Smoothing modules	QKV, V/O, Q/K, and MLP up/gate smoothing; shifts disabled for LLaMA-style models
Optimizer	AdamW
Optimization schedule	20 epochs, batch size 1, 128 steps/epoch, 2560 updates per layer
Learning rates	0.005 for LET, 0.01 for LWC
Weight decay	0
LET initialization α	0.5
Activation statistics	Precomputed per model/seed and reused
Auxiliary-loss integration	Logit-KL auxiliary loss every 32 OmniQuant optimization steps, auxiliary batch size 4

Table 11: Quantization and calibration-time optimization settings for the OmniQuant check. The baseline uses only the layer-wise reconstruction objective, while our variant adds the explanation-aware auxiliary losses using the shared settings in Table 9.

Study-wide budget Based on Slurm accounting, the completed experiments reported in this paper used approximately 3.0k L40S GPU-hours, computed as allocated GPUs multiplied by wall-clock runtime. This total includes the main OSTQuant experiments, ablations, additional OmniQuant checks, rationale generation, cache construction, and supporting analyses run locally.

E Lambda Sensitivity

We tune the shared explanation-aware loss weight λ , where $\lambda_{\text{rat}} = \lambda_{\text{rec}} = \lambda$, on the MedExpQA validation split. The validation split is small, so we use this analysis only to choose a single global hyperparameter, not as an independent result. As shown in Table 13, we select $\lambda = 0.1$ since it gives the best average validation PredRec while

Stage	GPU-hours
Standard OSTQuant / probe construction	3.11
Span proposal and cache construction	1.66
Cache-conditioned optimization	7.83
Ours total	12.60
Additional cost over OSTQuant	+9.49

Table 12: Representative calibration cost for one Llama-3-8B-Instruct run. GPU-hours are computed as the number of allocated GPUs multiplied by wall-clock runtime. All additional computation is incurred only during calibration.

λ	Avg. Acc.	Avg. PredRec
0.01	34.39	88.63
0.05	35.32	88.21
0.10	34.52	90.21
0.30	35.10	89.29
0.50	35.85	88.36

Table 13: Sensitivity to the shared explanation-aware loss weight on the MedExpQA validation split. Values are averaged over the four evaluated models. Main experiments use $\lambda = 0.1$.

keeping accuracy in the same range as the other settings. The selected value is fixed for all main experiments.

F PredRec Content Controls

Because PredRec and $\mathcal{L}_{\text{rec}}$ use a related forced-choice option-scoring form, we evaluate whether PredRec depends on rationale content or mainly reflects the recovery protocol. We construct several controls on MedExpQA. Standard recovery uses the question, answer options, and system rationale. Question-only recovery removes the rationale. Rationale-only recovery removes the question and uses the answer options and system rationale. Shuffled recovery replaces the rationale with one from another example. External-reconstructor recovery uses the same standard input setting, but replaces the original recovery model with a fixed Gemma-4-12B-IT (Gemma Team, 2026) reconstructor that is not one of the evaluated quantized systems and is not optimized by our method.

The controls show that PredRec is affected by the available rationale signal rather than reflecting only the forced-choice protocol. Removing the rationale or replacing it with a shuffled rationale lowers reconstruction scores for both systems, while the non-trivial question-only and rationale-only scores show that both the question/options

System	Std.	Q-only	R-only	Shuf.	Ext.
OSTQuant	56.54	39.60	50.07	35.62	30.20
Ours	88.47	75.73	83.87	71.22	64.13

Table 14: PredRec content and recovery controls on MedExpQA, averaged over the four evaluated models. Std. denotes the standard PredRec setting. Ext. uses Gemma-4-12B-IT as a fixed external reconstructor under the standard input setting.

Dataset	Model	FP	OSTQuant	Ours
MedExpQA	OpenBio	39.79	33.48±1.93	35.57±1.45
MedExpQA	Llama	45.43	38.94±0.49	40.21±0.95
MedExpQA	BioMistral	50.96	50.18±1.40	50.78±3.57
MedExpQA	Mistral	62.98	56.67±2.03	56.91±2.40
MedExpQA	OpenBio	25.60	23.20±0.80	24.27±0.92
MedExpQA	Llama	24.80	27.20±2.40	26.40±1.39
MedExpQA	BioMistral	28.80	31.47±4.41	30.13±3.95
MedExpQA	Mistral	44.00	40.00±2.12	43.47±1.85
ChallengeClin.QA	OpenBio	25.32	23.59±2.09	23.16±0.82
ChallengeClin.QA	Llama	25.00	23.16±0.68	23.59±1.63
ChallengeClin.QA	BioMistral	25.00	25.54±1.60	26.62±0.56
ChallengeClin.QA	Mistral	33.77	30.74±1.79	29.76±2.91

Table 15: Gold-answer reconstruction under W4A4KV4 quantization. Ours uses the $\lambda = 0.1$ setting.

and rationale carry recovery signal. The external-reconstructor control gives the same directional comparison, with ours higher than OSTQuant under Gemma-4-12B-IT recovery. We therefore use PredRec as a protocol-aligned consistency metric and interpret it together with FER, UCR, and pairwise judgments.

G Gold-Answer Reconstruction

Gold-answer reconstruction asks whether a rationale supports the correct answer. We treat GoldRec as complementary because it is coupled to task correctness, whereas PredRec isolates answer-rationale consistency.

H Diagnostic Analysis Details

This section describes how Table 1 is computed. This diagnostic analysis is used only to motivate the method.

Setting We run the diagnostic on Llama-3-8B-Instruct with MedExpQA. Let $\mathcal{I}_{\text{diag}}$ be the set of examples where the full-precision model M_0 and the OSTQuant model M_{OST} select the same answer. We use full-precision answer-conditioned rationales as fixed teacher-forced sequences, so the analysis measures rationale-token distribution drift rather than answer-switch cases.

Token-level signal For each example $i \in \mathcal{I}_{\text{diag}}$, let $r_i = (r_{i,1}, \dots, r_{i,T_i})$ be the full-precision rationale and let c_i be the rationale-generation prompt. At each rationale-token position t , we compute

$$d_{i,t} = \text{KL}(p_{M_0}(\cdot \mid c_i, r_{i,<t}) \parallel p_{M_{\text{OST}}}(\cdot \mid c_i, r_{i,<t})). \quad (9)$$

This teacher-forced KL measures how much quantization changes the next-token distribution at that rationale position.

Token grouping Each rationale token is assigned to one group using the following priority: numeric cues, answer/option tokens, clinical cues, and generic/discourse tokens. Numeric cues include numbers, lab values, dosages, ages, measurements, and other numeric quantities and selected-option text. Clinical cues are question-grounded context phrases that also appear in the full-precision rationale after normalization, excluding stopwords, generic words, numeric spans, and option strings. All remaining tokens are labeled generic/discourse.

Aggregation Let $z_{i,t}$ be the group assigned to token $r_{i,t}$, and let $\mathbf{1}[\cdot]$ denote the indicator function. For token group g , we report

$$T(g) = \frac{\sum_{i \in \mathcal{I}_{\text{diag}}} \sum_{t=1}^{T_i} \mathbf{1}[z_{i,t} = g]}{\sum_{i \in \mathcal{I}_{\text{diag}}} T_i}, \quad (10)$$

$$S(g) = \frac{\sum_{i \in \mathcal{I}_{\text{diag}}} \sum_{t=1}^{T_i} \mathbf{1}[z_{i,t} = g] d_{i,t}}{\sum_{i \in \mathcal{I}_{\text{diag}}} \sum_{t=1}^{T_i} d_{i,t}}. \quad (11)$$

We report $T(g)$ and $S(g)$ as TokenPct and SignalPct percentages in Table 1. TokenPct measures how often each group appears in full-precision rationales, while SignalPct measures how much of the total full-precision-to-OSTQuant KL mass is assigned to that group.

I Faithfulness Cache Details

This section provides the cache-construction details omitted from Section 4.1. The cache is built offline from the shared non-test calibration pool and is used only during calibration-time transformation optimization. For each calibration example $x_i = (q_i, \mathcal{O}_i, y_i)$, the full-precision teacher M_0 first selects an answer $\hat{y}_i^0$ using Eq. 2 and generates an answer-conditioned rationale $r_i = (r_{i,1}, \dots, r_{i,T_i})$. We keep only examples where $\hat{y}_i^0 = y_i$. This is a reliability filter. A prompted Qwen2.5-72B-Instruct model then proposes candidate evidence spans $\mathcal{C}_i$ from r_i . The proposer is used only for high-recall

candidate generation. Evidence support is determined by the behavioral scoring below. Direct answer-label cues are removed before evidence extraction.

teacher-answer-supporting evidence Let $a_i = \hat{y}_i^0$ be the teacher target answer. Since retained examples satisfy $\hat{y}_i^0 = y_i$, this target equals the gold option on cached examples. Let $u_i(z)$ denote a recovery prompt containing the question, answer options, and rationale content z . Let r_i^{-e} denote the rationale with span e removed, and let $\emptyset$ denote a prompt with no rationale evidence. For any model M , define

$$m_i^M(z) = m_{M,i}(u_i(z); a_i), \quad (12)$$

where $m_{M,i}$ is the answer margin in Eq. 3. For each candidate span $e \in \mathcal{C}_i$, we compute necessity- and sufficiency-style support scores:

$$\begin{aligned} N_M(e) &= [m_i^M(r_i) - m_i^M(r_i^{-e})]_+, \\ S_M(e) &= [m_i^M(e) - m_i^M(\emptyset)]_+, \end{aligned} \quad (13)$$

where $[x]_+ = \max(x, 0)$. We aggregate the two views with a maximum,

$$I_M(e) = \max\{N_M(e), \alpha S_M(e)\}, \quad (14)$$

so a span can be retained if it strongly supports the teacher answer. The teacher-answer-supporting evidence set is

$$\mathcal{F}_i = \{e \in \mathcal{C}_i : I_{M_0}(e) > \eta_F\}. \quad (15)$$

We discard examples with empty $\mathcal{F}_i$ or whose concatenated evidence does not recover a_i under M_0 .

Quantization-affected evidence Let M_p be the fixed same-calibration quantized probe. We apply the same support scoring to M_p and keep teacher-answer-supporting spans whose answer-supporting effect is weakened after quantization. For each $e \in \mathcal{F}_i$, define

$$\begin{aligned} D_i^{\text{imp}}(e) &= [I_{M_0}(e) - I_{M_p}(e)]_+, \\ D_i^{\text{mar}}(e) &= [m_i^{M_0}(e) - m_i^{M_p}(e)]_+, \\ D_i(e) &= D_i^{\text{imp}}(e) + \beta D_i^{\text{mar}}(e). \end{aligned} \quad (16)$$

The quantization-affected evidence set is

$$\mathcal{Q}_i = \{e \in \mathcal{F}_i : D_i(e) > \eta_Q\}. \quad (17)$$

Cache entries with $\mathcal{Q}_i = \emptyset$ are discarded.

Token weights We convert span-level scores into token-level weights over the teacher rationale. Let $\text{cov}_i(t)$ be the set of spans in $\mathcal{Q}_i$ that cover token $r_{i,t}$. We set

$$\tilde{\omega}_{i,t} = \sum_{e \in \text{cov}_i(t)} D_i(e), \quad \omega_{i,t} = \frac{\tilde{\omega}_{i,t}}{\sum_{s=1}^{T_i} \tilde{\omega}_{i,s}}. \quad (18)$$

Tokens not covered by $\mathcal{Q}_i$ receive zero weight. Each retained cache entry is

$$\xi_i = (x_i, r_i, \mathcal{F}_i, \mathcal{Q}_i, \omega_i) \in \mathcal{A}.$$

We use $\mathcal{Q}_i$ and ω_i for the rationale-token loss, and $\mathcal{F}_i$ for the rationale-conditioned recovery loss. The cache parameters $\alpha, \beta, \eta_F, \eta_Q$ are fixed before evaluation and reported in Appendix C.

J Prompts

This appendix lists the prompts used for answer selection, rationale generation, recovery evaluation, faithfulness-cache construction, full-precision Evidence Retention (FER), Unsupported-Claim Rate (UCR), pairwise judging, and qualitative analysis. Unless otherwise stated, generation calls use deterministic decoding.

J.1 Common System Prompts

Answer and rationale system prompt

You are a careful medical exam assistant. Always follow the requested response format exactly.

Recovery system prompt

You are a careful evaluator of medical multiple-choice question explanations.

Your task is to infer which option is best supported by the provided explanation.

Important rules:
Base your decision only on the question, options, and explanation.
Do not use outside medical knowledge beyond what is stated or directly implied by the explanation.
If the explanation is weak or incomplete, still choose the option best supported by the explanation.

Judge system prompt

You are an expert medical evaluator for clinical multiple-choice QA. Follow the rubric exactly. Output only the requested JSON object.

J.2 Answer Selection Prompt

For answer selection, we construct the prompt below and score each option completion after

Final: using the length-normalized conditional log-likelihood in Eq. 2. The model is not evaluated by free-form answer parsing.

Answer selection prompt

Answer the medical multiple-choice question.
Select the single best answer.

Return the option label followed by the option text.

Question:
{question}

{optional_contexts}

Options:
{label}. {option text}
...

Final:

J.3 Rationale Generation Prompt

After answer selection, we generate a concise rationale conditioned on the selected answer. The selected answer is provided explicitly, and the model is asked to explain why that answer is supported.

Answer-conditioned rationale generation prompt

Answer the medical multiple-choice question.
The final answer has already been selected.
Write a concise medical rationale that supports this selected answer.

Output exactly in this format:
Rationale: <2-3 concise medical sentences explaining the key evidence>

Question:
{question}

Options:
{label}. {option text}
...

Selected answer:
{selected_answer_label}. {selected_answer_text}

J.4 Candidate Evidence-Span Proposal Prompt

We use Qwen2.5-72B-Instruct only as a high-recall candidate span proposer for cache construction. The model is used only as a candidate span proposer; proposed spans are subsequently filtered by the behavioral option-scoring tests described in Section 4.1 and Appendix I.

Candidate evidence-span proposal prompt

You are a key-evidence span extractor.

Given a question, options, selected answer, and rationale, extract minimal evidence spans.

Definition:
An evidence span is a short phrase or clause that supports the selected answer without directly naming the selected answer.

Rules:

1. Extract up to {MAX_EVIDENCE_UNITS} candidate evidence spans.
2. Each evidence span must be an exact substring of the rationale.
3. Do NOT include the selected answer option text: "{SELECTED_ANSWER_TEXT}".
4. Do NOT include the option label: "{SELECTED_ANSWER_LABEL}".
5. Do NOT include phrases such as "the correct answer is", "therefore", "thus", "the most appropriate choice", or "the selected answer".
6. If a sentence contains the selected answer term, extract only the clue portion that does not name it.
7. Prefer clinical/statistical/mechanistic clues, contrastive facts, symptoms, signs, lab or imaging findings, and causal links.
8. Avoid generic conclusions and meta-text such as "End of response".
9. Each span should usually be 3-15 words.
10. If a candidate is longer than {MAX_EVIDENCE_WORDS} words, split it into smaller evidence spans.
11. Return JSON only, exactly with this schema: {"evidence_units":[{"text":"exact substring"}]}

Question:
{QUESTION}

Options:
{label}. {option text}
...

Selected answer:
{SELECTED_ANSWER_LABEL}. {SELECTED_ANSWER_TEXT}

Rationale:
{RATIONALE}

J.5 Recovery Prompt for PredRec and GoldRec

For predicted-answer reconstruction and gold-answer reconstruction, the recovery model receives the question, answer options, and generated rationale. We use deterministic candidate scoring rather than free-form generation. Each candidate option is appended after Final:, and the option with the highest length-normalized conditional log-likelihood is selected.

Rationale-based recovery prompt

Infer the answer to the following medical multiple-choice question based only on the provided explanation.
Choose exactly one option.

Return the option label followed by the option text.

Question type: {question_type}

Question:
{question}

Options:
{label}. {option text}
...

Explanation:
{generated_rationale}

Final:

J.6 Full-precision Evidence Retention Claim Extraction Prompt

For full-precision Evidence Retention (FER), we first decompose the full-precision rationale into atomic clinical claims.

FER atomic clinical-claim extraction prompt

You are an expert medical evaluator for clinical multiple-choice QA.

Extract atomic clinical evidence claims from the full-precision rationale.

Each claim must contain exactly one clinical assertion that is relevant to supporting the selected answer.

Include clinical findings, lab findings, diagnoses, treatments, risk factors, pathophysiology, or other discriminative clinical evidence.

Do not include direct answer statements, option letters, or phrases such as "the answer is".

Do not add information that is not stated or clearly implied by the rationale.

Return only a JSON object with this schema:

```
{
  "claims": [
    {
      "claim_id": "c1",
      "claim": "<one atomic clinical claim>",
      "evidence_type":
        "<clinical_finding|lab_finding|diagnosis|treatment
        |risk_factor|pathophysiology|other>"
    }
  ],
  "reason": "<brief reason>"
}
```

Question:
{question}

Options:
{label}. {option text}
...

Full-precision selected answer:
{full_precision_selected_answer}

Full-precision rationale:
{full_precision_rationale}

J.7 Full-precision Evidence Retention Claim-Support Judge Prompt

Given atomic clinical claims extracted from the full-precision rationale, the judge checks whether each claim is retained by a candidate rationale. A claim is counted as retained when the candidate rationale states, entails, or clearly preserves the clinical content of the full-precision claim. Contradicted claims are treated as not retained.

FER claim-support judgment prompt

You are an expert medical evaluator for clinical multiple-choice QA.

Judge whether each full-precision clinical evidence claim is supported by the candidate rationale in the context of the same question and answer options.

Mark a claim as:

- supported: the candidate rationale states, entails, or clearly preserves the clinical content of the full-precision claim.
- unsupported: the candidate rationale does not contain enough information to support the claim.
- contradicted: the candidate rationale states something incompatible with the claim.

Do not judge whether the answer is correct.

Do not use outside medical knowledge except to interpret standard medical language.

Judge only whether the provided candidate rationale supports the full-precision claim in this question context.

Return JSON only, with one result for every claim_id:

```
{
  "claim_results": [
    {
      "claim_id": "c1", "support": "supported", "reason":
        "<short reason>"
    },
    {
      "claim_id": "c2", "support": "contradicted", "reason":
        "<short reason>"
    }
  ],
  "reason": "<brief overall note>"
}
```

Question:
{question}

Options:
{label}. {option text}
...

Full-precision selected answer:
{full_precision_selected_answer}

Full-precision rationale:
{full_precision_rationale}

Candidate system:
{candidate_system}

Candidate selected answer:
{candidate_selected_answer}

Candidate rationale:
{candidate_rationale}

Full-precision clinical evidence claims:
{claims}

J.8 Unsupported-Claim Rate Claim Extraction Prompt

To compute UCR, we first decompose each system rationale into atomic clinical claims. Unlike FER, which extracts claim from the full-precision rationale and evaluates whether they are retained by a system rationale, UCR extracts claims from each system rationale and evaluates whether they are supported by the full-precision reference.

UCR atomic clinical-claim extraction prompt

You are an expert medical evaluator for clinical multiple-choice QA. Follow the rubric exactly. Output only the requested JSON object.

Extract atomic factual claims from the candidate rationale.

Definition:

- A claim is one medically or clinically meaningful factual statement, diagnostic assertion, mechanism, treatment statement, risk factor, or answer-supporting reasoning step.
- Extract claims only from the candidate rationale.
- Use the question and options only for disambiguation.

- Do not extract the answer letter itself as a claim.
- Exclude generic filler, uncertainty hedges, and purely stylistic text.
- Prefer short, atomic claims.
- Split compound statements when needed.
- Return at most 20 claims.
- Set truncated=true if additional factual claims were omitted because of this cap.
- If there is no factual medical claim, return an empty list and truncated=false.

Question type:

{question_type}

Question:

{question}

Options:

{label}. {option text}

...

Candidate selected answer:

{candidate_selected_answer}

Candidate rationale:

{candidate_rationale}

Return JSON only:

```
{
  "claims": [
    {
      "claim_id": "c1",
      "claim": "<one atomic claim>",
      "claim_type":

      "<clinical_finding|lab_finding|diagnosis|treatment|
      risk_factor|pathophysiology|general_medical_fact|
      answer_mapping|other>"
    }
  ],
  "reason": "<brief note on extraction>",
  "truncated": false
}
```

J.9 Unsupported-Claim Rate Claim-Support Judge Prompt

Given atomic clinical claims extracted from each system rationale, the judge checks whether each claim is supported by the full-precision rationale. The question and answer options are provided only to disambiguate claim meaning. Claims labeled unsupported added or contradicted are counted as unsupported, while generally valid background claims are not penalized. Although the judge also provides a secondary lenient clinical judgment, only the primary teacher-source judgment is used to compute UCR.

UCR Claim-support Judgment prompt

You are an expert medical evaluator for clinical multiple-choice QA. Follow the rubric exactly. Output only the requested JSON object.

Judge each candidate rationale claim under two reference settings.

Primary teacher-source setting, used for the reported UCR:

- The active source is only the FP16 teacher rationale.
- The question and options are shown only to disambiguate claim meaning.
- Do not use the question or options to mark a claim as

- source_supported unless the same claim is stated or clearly entailed by the FP16 rationale.
- Do not use the gold answer, gold explanation, or standard medical knowledge to mark a claim as source_supported.
- The question is whether the quantized rationale introduced a new evidence claim that is not supported by the FP16 teacher rationale.

Secondary lenient clinical setting:

- The active source is the FP16 teacher rationale plus clearly valid standard medical knowledge.
- The gold answer and gold explanation are intentionally excluded.
- This setting asks whether a new claim can still be clinically justified without using the gold explanation as a reference.

Labels for both settings:

- source_supported:
The claim is stated or clearly entailed by the active source.
 - valid_background:
The claim is not stated in the active source, but is a generally valid medical background fact and does not introduce a new patient-specific finding, diagnosis, treatment effect, causal claim, or answer-specific evidence.
 - unsupported_added:
The claim introduces a new patient-specific condition, finding, diagnosis, treatment effect, causal or risk claim, or answer-supporting evidence that is not supported by the active source.
 - contradicted:
The claim contradicts the active source or standard medical knowledge.
 - not_a_claim:
The item is only an answer letter, formatting, generic filler, or is not a factual medical or clinical claim.
- For fp16_overlap:
- present_in_fp16:
The FP16 rationale states the same claim or a clear paraphrase.
 - not_in_fp16:
The claim is newly introduced relative to the FP16 rationale.
 - unclear:
The overlap cannot be determined reliably.

Important:

Standard medical knowledge can make a claim valid_background in the primary teacher-source setting, but it must not make the claim source_supported unless it is also stated or clearly entailed by the FP16 rationale.

Question type:

{question_type}

Question:

{question}

Options:

{label}. {option text}

...

FP16 teacher selected answer:

{full_precision_selected_answer}

FP16 teacher rationale:

```
{full_precision_rationale}

Candidate selected answer:
{candidate_selected_answer}

Candidate rationale:
{candidate_rationale}

Candidate claims to judge:
{candidate_claims}

Return JSON only, with one result for every claim_id:
{
  "claim_results": [
    {
      "claim_id": "c1",
      "fp16_overlap":
        "<present_in_fp16|not_in_fp16|unclear>",
      "teacher_support_status":

      "<source_supported|valid_background|unsupported_added|
        contradicted|not_a_claim>",
      "teacher_support_source":
        "<fp16_rationale|standard_medical_knowledge|none|
        not_applicable>",
      "clinical_support_status":

      "<source_supported|valid_background|unsupported_added|
        contradicted|not_a_claim>",
      "clinical_support_source":
        "<fp16_rationale|standard_medical_knowledge|none|
        not_applicable>",
      "reason": "<short reason>"
    }
  ],
  "reason": "<brief overall note>"
}
```

J.10 Pairwise Clinical-Rationale Judge Prompt

For pairwise judging, two systems are anonymized as Response A and Response B. A/B order is randomized outside the prompt, and the judge is given an explicit tie option. The comparison focuses on clinical correctness, support for the system's own selected answer, clinical sufficiency, and overall quality.

Pairwise clinical-rationale judgment prompt

You are an expert medical evaluator.

Task:
Two anonymized model responses were written for the same medical multiple-choice question. Each response includes a predicted answer and an explanation. Choose which response is clinically better overall.

Compare them on these dimensions:

- clinical_correctness: Which response is more medically correct overall?
- answer_support: Which explanation better supports that response's own predicted answer?
- clinical_sufficiency: Which response provides more sufficient clinical evidence to justify its own predicted answer?

Comparison constraints:

- Compare the overall clinical quality of the two responses, not just style.
- Each response includes a predicted answer and an explanation.
- If the two responses predict the same answer, focus mainly on which explanation better justifies that shared answer.
- If the two responses predict different answers, compare which full response is clinically better overall,

- including whether the predicted answer itself is clinically more appropriate and whether the explanation justifies it.
- Judge answer_support and clinical_sufficiency relative to each response's own predicted answer.
- A wrong predicted answer can still be better supported by one explanation than the other, but overall it should usually lose to a clearly more clinically appropriate response.
- If an explanation only restates the predicted option without meaningful clinical reasoning or evidence, it should not beat a more reasoned explanation on answer_support or clinical_sufficiency.
- If an explanation is generic and could apply to multiple options, it should not beat a more specific explanation on clinical_sufficiency.
- If an explanation contains a central medical error, it should not beat a more medically correct explanation on clinical_correctness or clinical_sufficiency.
- If an explanation contradicts that response's predicted answer, it should not beat a supporting explanation on answer_support.
- If a response has no explanation, treat that as a weakness.
- Prefer Tie when the two explanations are effectively equal.
- Do not reward verbosity or confidence.

Question type:
{question_type}

Question:
{question}

Options:
{label}. {option text}
...

Response A predicted answer:
{response_A_predicted_answer}

Response A explanation:
{response_A_explanation}

Response B predicted answer:
{response_B_predicted_answer}

Response B explanation:
{response_B_explanation}

Output exactly one JSON object and nothing else, using only A, B, or Tie as values:
{ "clinical_correctness": "A", "answer_support": "A", "clinical_sufficiency": "Tie", "overall": "A" }

J.11 Qualitative Analysis Prompt

This analysis is illustrative and is not used as a quantitative metric. The two quantized systems are anonymized as System A and System B before being shown to the assistant, and the resulting summaries are manually inspected and edited by the authors.

Qualitative rationale-comparison prompt

You are assisting a qualitative comparison of rationales generated by medical multiple-choice QA systems.

The reference system is the comparison target for preservation analysis, but do not assume it is clinically perfect.

The two quantized systems are anonymized as Quantized system A and Quantized system B.

Task:
Compare how each quantized system's rationale uses clinical evidence relative to the reference rationale.

Important rules:

- Return only the requested structured output.
- Do not write an overall comparison.
- Do not summarize or rewrite the generated rationales.
- Do not assign numerical scores.
- Do not decide which system is globally better.
- Focus only on evidence preservation, rationale drift, and clinically relevant differences.
- Keep each field concise.
- Ground the analysis in the provided question, options, predictions, and rationales.
- You may use general medical knowledge only to identify whether a rationale introduces an unsupported, irrelevant, or clinically questionable claim.
- If a claim is unsupported or less relevant, explain it briefly and tie it to the provided rationale or clinical item.

Input:

Question:
{question}

Options:
{options}

Gold answer:
{gold_answer}

Full-precision prediction:
{full_precision_prediction}

Full-precision rationale:
{full_precision_rationale}

Quantized system A prediction:
{system_A_prediction}

Quantized system A rationale:
{system_A_rationale}

Quantized system B prediction:
{system_B_prediction}

Quantized system B rationale:
{system_B_rationale}

Return exactly this JSON object:

```
{
  "system_A": {
    "preserved_evidence": ["<concise item>", "..."],
    "omitted_or_weakened_evidence": ["<concise item>", "..."],
    "unsupported_or_questionable_claims": ["<concise item>", "..."],
    "clinical_rationale_drift": "<one concise sentence>"
  },
  "system_B": {
    "preserved_evidence": ["<concise item>", "..."],
    "omitted_or_weakened_evidence": ["<concise item>", "..."],
    "unsupported_or_questionable_claims": ["<concise item>", "..."],
    "clinical_rationale_drift": "<one concise sentence>"
  },
  "brief_comparison": "<two concise sentences comparing evidence preservation only>"
}
```

K Human Validation of FER Judgments

Sample and task We conduct a human validation on 20 Llama-3-8B-Instruct MedExpQA cases from the same-prediction subset used for FER. The 20 full-precision rationales contain 126 automatically extracted atomic clinical claims. Each claim is evaluated against two anonymized candidate rationales, resulting in 252 candidate–claim judgments per annotator. The candidate rationales correspond

to OSTQuant and our method, but their identities are hidden during annotation.

Annotation protocol Two annotators complete the evaluation independently. For each case, annotators are shown the medical multiple-choice question, answer options, selected answer, full-precision reference rationale, two anonymized candidate rationales, and the extracted reference claims. For each claim, they judge whether each candidate rationale preserves the same clinical meaning. The claims are fixed in advance, and annotators do not evaluate the automatic claim-extraction stage. The examples are de-identified benchmark medical QA items, and annotations are reported only in aggregate.

Instructions given to annotators Annotators received the following instructions:

The goal of this evaluation is to judge whether each candidate explanation preserves the clinical meaning of the listed reference claim.

For each case, you will see a medical multiple-choice question, answer options, the selected answer, a full-precision reference rationale, two anonymized candidate rationales, and a list of atomic clinical claims extracted from the reference rationale.

For each listed claim, mark whether Candidate A and Candidate B preserve the same clinical meaning as the reference claim.

Use the following labels:

- *Supported*: the candidate rationale explicitly states the claim, clearly entails it, or provides an unambiguous paraphrase with the same clinical meaning.
- *Not supported*: the claim is absent, too vague, only partially preserved, or contradicted by the candidate rationale.

Judge semantic preservation rather than exact wording. Do not judge whether the selected answer is correct. Do not use this task to evaluate whether the automatically extracted claims are good or complete. The claims are fixed in advance. Focus only on whether each candidate rationale supports each listed reference claim.

Annotator recruitment and compensation Two researchers from the authors’ research group assisted with this small validation study after being invited by the authors. No external crowdsourcing platform was used, and no monetary compensation was provided. We acknowledge their assistance in the Acknowledgements.

Result analysis Both annotators completed all 252 candidate–claim judgments, with no missing labels. After de-anonymization, Candidate A corresponds to OSTQuant and Candidate B corresponds

to our method. We compute agreement at the binary candidate–claim level, treating each judgment as *Supported* or *Not Supported*. For the GPT comparison, we compare each human label with the corresponding GPT-based label for the same candidate–claim pair.

Comparison	Labels	Agree.	κ
Human–Human	252	79.37	0.58
GPT vs. Annotator 1	252	83.73	0.68
GPT vs. Annotator 2	252	78.17	0.58
GPT vs. pooled human	504	80.95	0.63
GPT vs. human consensus	200	89.00	0.78

Table 16: Human–human and human–GPT agreement for FER judgments on the 20-case Llama-3/MedExpQA sample. Agreement is computed over binary candidate–claim support labels.

The two annotators agree on 79.37% of the binary support judgments, with Cohen’s $\kappa = 0.58$, indicating moderate agreement. Across all human–GPT label comparisons, the GPT-based judge agrees with the human labels in 80.95% of cases, with Cohen’s $\kappa = 0.63$. On the subset where both human annotators agree, GPT agrees with the human consensus in 89.00% of cases, with Cohen’s $\kappa = 0.78$. This suggests that the automatic judge is reasonably aligned with human annotations, especially on less ambiguous judgments.

Evaluator	OSTQuant	Ours	Δ
Annotator 1	47.62	57.14	+9.52
Annotator 2	57.94	67.46	+9.52
Pooled human	52.78	62.30	+9.52
GPT-based FER	38.10	50.00	+11.90

Table 17: Human validation of FER on the 20-case Llama-3/MedExpQA sample. Values are micro FER percentages. Δ denotes Ours minus OSTQuant.

The human judgments show the same directional trend as the automatic FER judge. Pooling both annotators gives 52.78% micro FER for OSTQuant and 62.30% for our method, a +9.52 point difference. Thus, while the automatic judge is stricter in absolute support rate, both human and automatic evaluations agree directionally that our method better preserves full-precision clinical evidence on this sample.

L Pairwise Rationale Judging Details

Table 18 reports the full pairwise preference counts on MedExpQA same-answer cases. The pairwise

Model	Same pred.	OST win	Ours win	Tie	Ours%
OpenBioLLM	44/125	9	18	17	66.67
Llama-3	38/125	12	16	10	57.14
BioMistral	64/125	25	26	13	50.98
Mistral	71/125	17	29	25	63.04
Avg./pooled	217/500	63	89	65	58.55

Table 18: Pairwise rationale preference counts on MedExpQA. Ours% excludes ties.

prompt and rubric are provided in Appendix J.10.

M Qualitative Examples

We provide two MedExpQA examples from Llama-3-8B-Instruct to examine rationale drift when the final answer is preserved. In both experiments, the full-precision model, OSTQuant, and our method select the same option, allowing comparison of rationale content independent of answer differences. We used GPT-5.4 API to assist a structured qualitative comparison of generated rationales. The quantized systems were anonymized during comparison, and the resulting summaries were manually inspected and edited by the authors. Table 19 presents the comparisons, and the qualitative-analysis prompt is provided in Appendix J.

Example 1 OSTQuant preserves the broad neuropathic-pain interpretation and the selected answer, but introduces an unsupported medication-causation claim involving morphine and lactulose. In contrast, our method more closely follows the full-precision evidence structure and uses the normal neurological examination as a cue against focal nerve compression. This example illustrates how answer-preserving quantization can still alter the clinical justification.

Example 2 OSTQuant preserves the broad worsening-heart-failure interpretation, but introduces a clinically relevant medication-class error by describing enalapril as a diuretic, which blurs the roles of ACE-inhibitor up titration and intravenous diuresis. In contrast, our method retains the main full-precision evidence, including orthopnea, jugular venous distension, edema, and intravenous furosemide for fluid overload, although it omits some mechanistic detail about enalapril and after-load reduction. This example shows that even when the selected answer is unchanged, quantization can alter clinically important justification details.

Example 1: Neuropathic pain management. Question: A 66-year-old patient with stage IV pancreatic adenocarcinoma is taking delayed-release morphine. He develops paresthesias and occasional lancinating pain in the right lumbar and periumbilical area that does not subside with rescue morphine. Neurological examination is normal.
Options: A: CT and evaluate nerve compression; B: neurolytic treatment; C: decrease morphine; D: administer amitriptyline or gabapentin, dexamethasone, and increase morphine. **Gold: D.**

Aspect	FP	OSTQuant	Ours
Prediction	D	D	D
Generated rationale	The patient’s symptoms of paresthesias and lancinating pain in the right lumbar and periumbilical area, despite adequate morphine treatment, suggest the presence of neuropathic pain. Amitriptyline or gabapentin, dexamethasone, and increasing the dose of morphine are effective in managing neuropathic pain, which is a common complication of pancreatic cancer.	The patient’s symptoms are consistent with neuropathic pain, which is a common complication of pancreatic cancer. The patient’s history of delayed release morphine and lactulose may contribute to the development of neuropathic pain. The selected answer addresses the patient’s symptoms by adding adjuvant therapy with amitriptyline or gabapentin, dexamethasone, and increasing the dose of morphine, which is a combination of treatments commonly used to manage neuropathic pain.	The patient’s symptoms of paresthesias and lancinating pain in the right lumbar and periumbilical area are consistent with neuropathic pain, which is a common complication of pancreatic cancer. The patient’s normal neurological examination suggests that the pain is not due to compression of a specific nerve. Therefore, the most appropriate management is to try to control the neuropathic pain with medications such as amitriptyline or gabapentin, and dexamethasone, while also increasing the dose of morphine. This approach is supported by evidence that these medications can be effective in managing neuropathic pain in patients with cancer.
Clinical evidence comparison	Uses the key evidence: paresthesias, lancinating pain, inadequate response to morphine, pancreatic cancer context, and option D therapies.	Preserves the neuropathic-pain interpretation and option D support, but adds an unsupported claim that morphine and lactulose may contribute to neuropathic pain.	Preserves the reference evidence and adds relevant use of the normal neurological examination to argue against focal nerve compression.
Takeaway	Full-precision reference reasoning, though it assumes treatment effectiveness broadly.	Mostly aligned with the FP answer, but shows rationale drift from an unsupported medication-causation claim.	Closely aligned with the FP rationale and adds a relevant discriminative cue without major drift.

Example 2: Decompensated heart failure management. Question: A 73-year-old woman is admitted with progressive dyspnea at rest, orthopnea, and 4 kg weight gain. Physical examination shows BP 150/84 mmHg, HR 100/min, increased jugular venous pressure, bibasal crackles, and malleolar edema. Usual treatment: enalapril 5 mg every 12 hours, furosemide 80 mg/day.
Options: A: administer intravenous furosemide; B: increase enalapril dose according to tolerance and administer intravenous furosemide; C: start a beta-blocker; D: add amlodipine. **Gold: B.**

Aspect	FP	OSTQuant	Ours
Prediction	B	B	B
Generated rationale	The patient presents with signs of heart failure, including orthopnea, weight gain, and jugular venous distension, indicating a need for increased diuretic therapy. Increasing the dose of enalapril, a converting enzyme inhibitor, can help reduce afterload and improve cardiac function, while adding intravenous furosemide can further enhance diuresis and alleviate symptoms.	The patient’s symptoms of progressive dyspnea, orthopnea, and weight gain suggest worsening heart failure. Increasing the dose of enalapril, a diuretic, and administering intravenous furosemide will help to further reduce fluid overload and alleviate symptoms. This approach is supported by the patient’s current treatment regimen, which includes a diuretic, and the need to further reduce fluid accumulation.	The patient is presenting with signs of decompensated heart failure, including orthopnea, jugular venous distension, and pedal edema, indicating a need for increased diuretic therapy. Increasing the dose of enalapril and administering intravenous furosemide will help to reduce fluid overload and alleviate symptoms.
Clinical evidence comparison	Uses the key heart-failure evidence: orthopnea, weight gain, jugular venous distension, and need for intravenous diuresis; also gives an ACE-inhibitor uptitration rationale.	Preserves worsening heart failure and fluid-overload evidence, but incorrectly calls enalapril a diuretic, blurring ACE-inhibitor and diuretic roles.	Closely preserves the reference evidence, including decompensated heart-failure signs, jugular venous distension, edema, and intravenous furosemide for diuresis, but omits the afterload/cardiac-function rationale for enalapril.
Takeaway	Preserves the core clinical basis for option B.	Mostly preserves the evidence but introduces a clinically relevant medication-class error.	Strong evidence preservation with minor loss of ACE-inhibitor mechanism detail.

Table 19: Representative qualitative examples from the all-answer-match category. All systems predict the same answer, but OSTQuant introduces rationale drift, while our method better preserves the full-precision model’s answer-supporting clinical evidence.