

Article

Memory-Efficient Adaptation and Unanswerable-Aware Supervision for Reliable Private Email Question Answering

Fesih Keskin 

Department of Computer Engineering, Siirt University, Siirt 56100, Türkiye; fesih.keskin@siirt.edu.tr

Abstract

Private document question answering (QA) requires language models to answer from local evidence while abstaining when the supplied context is insufficient. This study evaluates quantized parameter-efficient adaptation for private email QA using EnronQA. A Qwen2.5-7B-Instruct backbone is adapted with QLoRA and QPiSSA under matched rank, target modules, quantization, and training budget. QPiSSA combines PiSSA initialization with residual quantization, whereas QLoRA quantizes the original frozen weight matrix and trains standard low-rank adapters. In addition to answerable-only fine-tuning, an unanswerable-aware supervision variant augments training with synthetic missing-evidence examples whose target response is “I don’t know.” Experiments are conducted on a fixed EnronQA subset with 20,000 training, 2000 development, and 2000 test examples. Fine-tuning substantially improves oracle-email answer quality over zero-shot prompting, but answerable-only tuning sharply reduces abstention under missing evidence. Adding unanswerable examples restores abstention accuracy to approximately 0.998 for both QLoRA and QPiSSA while preserving most answer-generation performance. Bootstrap confidence intervals over the fixed test set, multi-seed evaluation, BERTScore evaluation, leakage diagnostics, strict filtered unanswerable evaluation, retrieval-augmented testing, quantization-error diagnostics, and a small diagnostic ablation support a cautious interpretation: QLoRA and QPiSSA perform similarly under this matched configuration, and the main reliability gain comes from explicit missing-evidence supervision rather than from the adapter choice alone.

Keywords: abstention; EnronQA; large language models; LoRA; missing-evidence reliability; private document question answering; QLoRA; QPiSSA; retrieval-augmented generation



Academic Editors: Francisco De Arriba-Pérez, Silvia García-Méndez and Enrique Costa-Montenegro

Received: 10 June 2026

Revised: 23 June 2026

Accepted: 25 June 2026

Published: 3 July 2026

Copyright: © 2026 by the author.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

The deployment of large language models (LLMs) has transformed natural language processing, yet using these models to reason over private domain-specific data is still difficult [1]. Models trained on large public corpora do not have access to proprietary information such as internal corporate communication or private email repositories, and many organizations cannot send sensitive documents to external providers because of privacy and compliance constraints [2]. Retrieval-augmented generation (RAG) has become the most common way to close this gap by grounding generation in retrieved local context, which can reduce hallucinations and allow the system to use up-to-date private knowledge stores [3]. In practice, many RAG pipelines also need clear abstention behavior so the model answers from evidence and declines when evidence is missing, which is often expressed

through explicit instructions to say it does not know when the answer cannot be found in the retrieved content [4].

Email question answering (QA) is a particularly challenging form of private document QA. Emails include long threads, inconsistent formatting, dense named entities, and context that may be distributed across replies and forwarded messages. These characteristics make private email QA a suitable setting for evaluating whether an adapted model can generate answers from supplied evidence while abstaining when the evidence is insufficient. This study uses EnronQA as the evaluation benchmark; dataset structure, sampling, and preprocessing details are given in Section 3.

Adapting large models to such downstream tasks through full fine-tuning is often out of reach because of compute and memory requirements, especially as model size grows [5]. Parameter-efficient fine-tuning (PEFT) methods offer a practical alternative by freezing most pretrained weights and updating only a small subset of parameters, which reduces the memory burden of optimizer states and gradients [6]. Among these methods, low-rank adaptation (LoRA) has become a dominant approach by injecting trainable low-rank matrices into existing linear layers, approximating weight updates without changing the base architecture and without introducing meaningful inference latency in common deployment settings [7].

Building on LoRA, quantized low-rank adaptation (QLoRA) has become a strong baseline when GPU memory is limited. QLoRA enables fine-tuning through a frozen 4-bit quantized base model while training low-rank adapters, and it introduces practical techniques such as 4-bit NormalFloat (NF4), double quantization, and paged optimizers to reduce memory use and prevent out-of-memory errors during training [8]. Even so, quantization and initialization can affect the starting point of optimization. Recent work highlights that the combination of a quantized base model with standard adapter initialization can create a mismatch from the original pretrained weights, which can slow convergence or reduce downstream performance in tasks where knowledge retention and stable adaptation matter [6,9].

To address limitations linked to random initialization and quantization error, Principal Singular Values and Singular Vectors Adaptation (PiSSA) and its quantized variant QPiSSA have been proposed. PiSSA uses singular value decomposition to factorize a pretrained weight matrix into principal components and a residual component, then initializes the trainable adapters using the principal components while keeping the residual frozen [9]. This SVD-based initialization is intended to update important directions of the weight space immediately, leading to faster convergence and improved performance compared with standard LoRA initialization [9]. QPiSSA extends this idea to the quantized setting by quantizing the residual rather than quantizing the full weight matrix, which can reduce quantization error because the residual has a narrower distribution than the original weights and is better suited to NF4-style quantization [9,10].

In this work, we investigate how SVD-based adaptation strategies affect oracle-email answer generation over private documents. Using a controlled oracle-email evaluation protocol, we study how initialization, quantization, and unanswerable-aware supervision influence convergence, answer quality, and missing-evidence reliability when adapting LLMs to private email corpora.

The contributions of this paper are summarized as follows.

- We provide a controlled empirical comparison of QLoRA and QPiSSA on EnronQA under matched backbone, adapter rank, target modules, quantization configuration, prompt format, and training budget.

- We evaluate how unanswerable-aware supervision, defined as training on answerable examples plus synthetic missing-evidence examples, affects the trade-off between oracle-email answer quality and missing-evidence abstention reliability.
- This study reports bootstrap, multi-seed, semantic, leakage, retrieval-augmented, quantization-error, diagnostic-ablation, efficiency, and error-analysis evidence to characterize the strengths and limitations of quantized adapter tuning for private email QA.

2. Background and Related Work

This section reviews private document question answering and the efficient adaptation of LLMs that supports it. We begin with benchmarks for private and email-based retrieval-augmented generation (RAG), then summarize parameter-efficient fine-tuning with an emphasis on low-rank adaptation. Next, we discuss quantized fine-tuning through QLoRA and why quantization and initialization can create a difficult starting point for optimization. We then describe PiSSA and QPiSSA, which use singular value decomposition to improve initialization and reduce quantization error. Finally, we summarize several recent alternatives and clarify the research gap that motivates our study.

2.1. Private Document and Email QA Benchmarks

Retrieval-augmented generation (RAG) is commonly used to ground model outputs in external documents rather than relying on parametric memorization, which is especially important when the target information is private and not present in public pretraining data [2]. A recurring problem with many public RAG benchmarks is that their underlying sources are often public corpora such as Wikipedia or general web pages, which modern models may partially memorize. This makes it harder to separate retrieval quality from what the model already knows. Ryan et al. highlight this concern and note that, on popular RAG benchmarks, it is possible to score above 60% without any retrieval, which weakens the diagnostic value of those datasets for retrieval research [2].

To better reflect enterprise-like scenarios, Ryan et al. introduce EnronQA, a benchmark built from the Enron email corpus that targets personalized retrieval over private documents [2]. Because the content is private, user-specific, and email-centered, EnronQA reduces the likelihood that a model can succeed by relying only on memorized public knowledge. This makes it a suitable benchmark for studying grounding, retrieval dependence, and private document question answering [2]. The dataset details and sampling protocol used in this study are described in Section 3.

2.2. Parameter-Efficient Fine-Tuning and Low-Rank Adaptation

As model sizes grow, full fine-tuning becomes difficult for many users due to GPU memory constraints. PEFT methods reduce training cost by freezing the base model and training a small set of additional parameters, which reduces GPU memory needs largely by shrinking optimizer states [5]. Low-rank adaptation is one of the most widely used PEFT methods. It represents weight updates with a low-rank decomposition so that only the adapter parameters are trained while the pretrained weights remain fixed [7,8].

Standard LoRA initialization is designed to preserve the pretrained model behavior at step zero. A common setup initializes one low-rank factor with Gaussian noise and the other to zero so that the adapter contribution is initially zero [11]. While this is stable, PiSSA-related analyses argue that random initialization can be poorly aligned with important directions in the pretrained weight space, which can contribute to slower convergence in practice [9].

Another practical issue is that LoRA does not reduce the memory footprint of the frozen base model, so users still face hardware limits as model size increases [8].

PEFT choices can also affect learning dynamics and retention. Biderman et al. analyze a learning and forgetting trade-off and report that LoRA often learns less and forgets less than full fine-tuning, and that adapter rank can act as a control knob for this behavior [12]. Although this study is not specific to email QA, it motivates careful comparisons between methods when the target setting values both accuracy and stability.

2.3. Quantized Fine-Tuning and QLoRA

Quantization-based fine-tuning is a practical approach for training large models under tight memory budgets. QLoRA stores the base model in 4-bit precision and trains only low-rank adapters in higher precision, enabling fine-tuning of large backbones that would otherwise be impractical on a single GPU [8]. QLoRA introduces NF4, double quantization to reduce memory overhead from quantization constants, and paged optimizers to manage memory spikes by paging optimizer states between GPU and CPU memory [8]. Dettmers et al. also point out an important system detail that most of the memory footprint in PEFT fine-tuning can come from activation gradients rather than the LoRA parameters themselves, so focusing only on shrinking adapter size does not always deliver large end-to-end memory savings [8].

Quantization can also shift the optimization starting point. With a quantized base model, the effective initial weights reflect the quantized version of the pretrained weights, while the adapters start near zero, so the model begins training from a state that can deviate from the original high-precision initialization [6,9]. Liao et al. describe this as a key reason memory-efficient fine-tuning can show inconsistent performance across bit widths and tasks because quantization can harm preserved knowledge and amplify error propagation during adaptation [6]. This motivates methods that explicitly reduce quantization error at initialization rather than expecting the adapters to correct it later.

2.4. PiSSA and QPiSSA with SVD-Based Initialization

PiSSA was proposed to address the limitations of random LoRA initialization by using singular value decomposition (SVD) to initialize low-rank adapters from principal components of pretrained weight matrices [9]. Its quantized variant, QPiSSA, represents the principal low-rank component through the adapter path and quantizes the residual component rather than the full weight matrix. Prior work argues that this residual-quantization strategy can reduce quantization error and improve optimization because the residual has a narrower distribution than the original weights [9]. Related follow-up work further develops PiSSA-style ideas, including variants that focus on higher-rank adaptation and distributed settings [10]. The present article gives the full operational formulation only in Section 3.8.2 to avoid repeating the same decomposition in the Background and Related Work and Materials and Methods Sections.

2.5. Other PEFT and Quantization Alternatives

Several related PEFT and quantization methods also target initialization, quantization error, inference efficiency, or richer adapter parameterizations. LoftQ-style approaches initialize adapters to compensate for quantization error [6], LQ-LoRA decomposes fine-tuning into low-rank and quantized components [13], QuAILoRA uses calibration data for quantization-aware initialization [14], and L4Q combines quantization-aware training with adapter tuning for efficient inference [15]. Other methods, including LoRETTA, DoTA, and Kronecker-structured adaptation, explore tensorized or structured adapter parameterizations as alternatives to simple matrix-based low-rank updates [16–18]. These

approaches are complementary to the present focus on comparing QLoRA and QPiSSA in private email QA.

2.6. Grounded Answering and Abstention in Private Settings

Private document QA also requires reliability behaviors that go beyond average accuracy. In RAG systems, it is common to instruct the model to answer using only retrieved context and to abstain when the evidence is missing, which reduces unsupported claims and helps to prevent accidental disclosure [4]. Recent work on secure RAG further emphasizes that unanswerable outcomes can be enforced through input and output defenses, which makes abstention a design requirement rather than only a user-interface preference [4]. These concerns are especially relevant for email QA, where the content can be sensitive and the model may otherwise generate plausible sounding but incorrect answers.

2.7. Research Gap

Although QLoRA and PiSSA-style methods have been studied on popular public benchmarks, it remains unclear how much their advantages transfer to private email QA, where success depends on correctly using supplied context that is unlikely to be memorized. EnronQA was designed to benchmark this private setting and highlights that retrieval and grounding behavior matter more than parametric recall of public facts [2]. At the same time, PiSSA and QPiSSA claim both faster convergence and reduced quantization error through SVD-based initialization and residual quantization [9]. What is still missing is a focused evaluation that asks whether these benefits translate into improved oracle-email answer generation and missing-evidence reliability in private email question answering under realistic memory constraints. This work targets that gap by comparing QPiSSA and QLoRA on EnronQA, with an evaluation that emphasizes both answer quality and abstention reliability.

3. Materials and Methods

3.1. Dataset Overview

Experiments are conducted on EnronQA, a private document question answering benchmark derived from the Enron email corpus [2]. EnronQA contains 103,638 cleaned emails and 528,304 question–answer pairs across 150 user inboxes, providing a realistic testbed for personalized and private retrieval-augmented question answering over email collections [2]. Unlike many public-domain RAG benchmarks, EnronQA is grounded in private-document-style email content, making it suitable for evaluating whether a model can answer from supplied evidence rather than relying on parametric memorization.

This work focuses on the oracle-email setting, where the model receives the gold email associated with the question. This setting isolates answer generation under supplied gold-email evidence from retrieval quality. A separate BM25 diagnostic is also reported to characterize lexical retrievability in the sampled corpus, but the main model-comparison results are based on oracle-email answer generation.

3.2. Task Definition

Each example is represented as $x_i = (e_i, q_i, a_i, u_i, d_i)$, where e_i denotes the email context, q_i denotes the question, a_i denotes the gold answer, u_i denotes the user or mailbox identifier, and d_i denotes the document identifier or email path. The objective is to generate an answer $\hat{a}_i = f_\theta(e_i, q_i)$ that is supported by the supplied email. When the supplied evidence does not contain the answer, the desired behavior is abstention using the response “I don’t know.”

For supervised fine-tuning and metric computation, the full tuple is projected onto the email–question–answer fields as $\pi_{\text{eqa}}(x_i) = (e_i, q_i, a_i)$. The metadata fields u_i and

d_i are retained for traceability, retrieval diagnostics, and possible audit checks, but they are not included in the generated answer target. In the remainder of the methodology, supervised splits are written in projected form after applying π_{eqa} unless metadata fields are explicitly discussed.

3.3. Sampling and Data Snapshots

The original EnronQA records are flattened from email-level records into question-level examples. If one email is associated with m question–answer pairs, then it is converted into m independent training or evaluation instances. A fixed random seed of 42 is used for sampling, shuffling, unanswerable construction, and retrieval diagnostics. The final medium-scale experimental subset consists of 20,000 training examples, 2000 development examples, and 2000 test examples:

$$\mathcal{S}_{\text{train}} = \text{Sample}_{20,000}(\mathcal{D}_{\text{train}}; 42), \quad (1)$$

$$\mathcal{S}_{\text{dev}} = \text{Sample}_{2000}(\mathcal{D}_{\text{dev}}; 42), \quad (2)$$

$$\mathcal{S}_{\text{test}} = \text{Sample}_{2000}(\mathcal{D}_{\text{test}}; 42). \quad (3)$$

The same development and test subsets are used across all methods, enabling paired comparison. All sampled splits are saved as JSONL snapshots to support reproducibility.

3.4. Proposed Framework

This study investigates quantized parameter-efficient adaptation for private document question answering under an oracle-email setting. The experimental design compares QLoRA with the combined SVD-initialized and residual-quantized QPiSSA strategy and evaluates whether unanswerable-aware supervision improves missing-evidence reliability without substantially degrading answer generation. Figure 1 summarizes the study design, including EnronQA sampling, prompt construction, answerable-only and unanswerable-augmented training, quantized adapter tuning, and evaluation. The algebraic details of QLoRA and QPiSSA are given once in Sections 3.8.1 and 3.8.2, respectively.

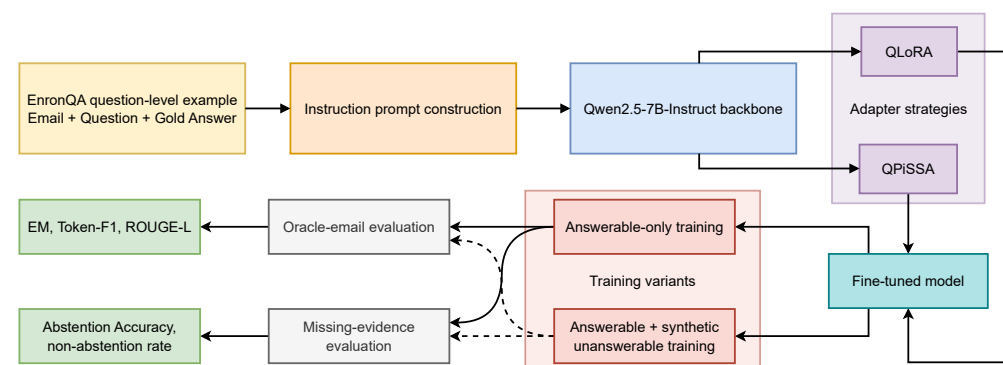


Figure 1. Overall experimental pipeline. EnronQA is flattened into question-level examples and sampled into fixed training, development, and test subsets. Answerable-only and unanswerable-augmented training sets are used to fine-tune QLoRA and QPiSSA adapters, which are then evaluated under oracle-email and missing-evidence settings.

The answerable training subset is denoted by $\mathcal{S}_{\text{ans}}$, and the unanswerable-augmented training subset is denoted by $\mathcal{S}_{\text{ans}}^+$.

3.5. Prompt Construction and Supervised Fine-Tuning Objective

Each example is converted into an instruction-style causal language modeling sequence for supervised fine-tuning (SFT). The prompt contains a system instruction, the

email context, the question, and an answer prefix. The system instruction requires the model to answer only from the supplied email and to produce “I don’t know.” when the answer is absent.

The prompt is represented as $p_i = \text{Prompt}(e_i, q_i) = [\text{System}; \text{Email} : e_i; \text{Question} : q_i; \text{Answer} :]$. Let y_i denote the tokenized target answer, and let $z_i = [p_i; y_i]$ denote the full causal language modeling sequence. Only target-answer tokens contribute to the supervised loss; prompt tokens are masked using the ignore index. The loss for example i is

$$\mathcal{L}_i = - \sum_{t=1}^{|y_i|} \log P_{\theta}(y_{i,t} \mid p_i, y_{i,<t}). \quad (4)$$

The maximum sequence length is set to 2048 tokens. When the email and question exceed this limit, the email is truncated from the beginning while preserving the question and answer target. This strategy keeps the most recent portion of long email threads, which often contains replies and forwarded context. During evaluation, generated answers are limited to 64 new tokens, and sampling is disabled to make decoding deterministic.

3.6. Synthetic Unanswerable Example Construction

To evaluate missing-evidence reliability, synthetic unanswerable examples are generated by replacing the correct email with an unrelated email while keeping the original question. The target response for such examples is fixed as $a_i^{\text{unans}} = \text{“I don’t know.”}$.

Let $\rho = 0.15$ be the unanswerable augmentation ratio. This value is used as a practical compromise between exposing the model to missing-evidence supervision and keeping answerable examples dominant in the training objective; Section 4 reports a sensitivity analysis supporting this choice. A subset $\mathcal{U} \subset \mathcal{S}_{\text{train}}$ is sampled from the answerable training set. For each $(e_i, q_i, a_i) \in \mathcal{U}$, a replacement email $\tilde{e}_i$ is sampled from the training email pool such that $\tilde{e}_i \neq e_i$. The resulting unanswerable example is $(\tilde{e}_i, q_i, \text{“I don’t know.”})$. The augmented training set is defined as

$$\mathcal{S}_{\text{train}}^+ = \mathcal{S}_{\text{train}} \cup \{(\tilde{e}_i, q_i, \text{“I don’t know.”}) : (e_i, q_i, a_i) \in \mathcal{U}\}. \quad (5)$$

This procedure creates a controlled missing-evidence supervision signal without introducing external documents. Because different-email replacement cannot fully exclude indirect supporting evidence or accidental answer overlap, we quantify this risk using the leakage diagnostics described in Section 3.17 and report abstention both before and after strict leakage filtering.

3.7. Experimental Conditions

Five experimental conditions are evaluated under the same prompt format, test set, decoding procedure, and metric implementation:

1. Zero-shot: the base instruction model is evaluated without fine-tuning.
2. QLoRA: standard LoRA adapters are trained on answerable examples, while the base model is stored in 4-bit precision [8].
3. QPiSSA: PiSSA-initialized adapters are trained using a decomposition in which the principal low-rank component initializes the adapter path and the residual base is stored in 4-bit precision [9].
4. QLoRA + unanswerable training: QLoRA is trained on $\mathcal{S}_{\text{train}}^+$.
5. QPiSSA + unanswerable training: QPiSSA is trained on $\mathcal{S}_{\text{train}}^+$.

The answerable-only and unanswerable-augmented variants are stored in separate run directories to prevent checkpoint or adapter overwriting.

3.8. Backbone Model and Adapter Placement

The backbone model is Qwen/Qwen2.5-7B-Instruct, an instruction-tuned decoder-only transformer. The base parameters remain frozen during fine-tuning; only adapter parameters are updated. All methods use the same tokenizer, maximum sequence length, generation length, and prompt template.

LoRA modifies a frozen linear transformation using a low-rank update [11]. For a pretrained weight matrix $W_0 \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$, the adapted transformation is

$$h = W_0 x + \Delta W x, \quad \Delta W = \frac{\alpha}{r} B A, \quad (6)$$

where $A \in \mathbb{R}^{r \times d_{\text{in}}}$ and $B \in \mathbb{R}^{d_{\text{out}} \times r}$. The adapter rank is $r = 16$, the scaling coefficient is $\alpha = 32$, and LoRA dropout is 0.05. Adapters are inserted into the attention and feed-forward projection modules $\mathcal{M} = \{\text{q_proj}, \text{k_proj}, \text{v_proj}, \text{o_proj}, \text{gate_proj}, \text{up_proj}, \text{down_proj}\}$.

3.8.1. QLoRA Adaptation

QLoRA stores the frozen base model in 4-bit precision and trains LoRA adapters in higher precision [8]. Dettmers et al. introduced QLoRA with 4-bit NormalFloat (NF4), double quantization, and paged optimizers to reduce memory usage while retaining strong fine-tuning performance [8]. Let $Q(\cdot)$ denote the 4-bit quantization operator. The effective QLoRA transformation is

$$h = Q(W_0)x + \frac{\alpha}{r} B A x. \quad (7)$$

NF4 quantization and double quantization are used for base model storage. Computation is performed in bfloat16 when supported by the hardware. Optimization is performed with AdamW using high-precision 32-bit paging, which provides stable optimizer precision while reducing memory-spike risk during quantized fine-tuning.

3.8.2. QPiSSA Adaptation

PiSSA initializes low-rank adapters using principal singular directions of pretrained weights [9]. Unlike standard LoRA, which commonly starts from a near-zero adapter update, PiSSA first decomposes the pretrained matrix as $W_0 = U \Sigma V^\top$ and assigns the dominant low-rank component to the adapter path while keeping the remaining residual component frozen. Let U_r , Σ_r , and V_r denote the top- r singular components. The principal and residual components are

$$W_r = U_r \Sigma_r V_r^\top, \quad (8)$$

$$W_{\text{res}} = W_0 - W_r. \quad (9)$$

This gives $W_0 = W_{\text{res}} + W_r$. The principal component is represented through the trainable adapter path as $W_r \approx \gamma B_p A_p$, where A_p and B_p correspond to the rank- r singular-vector factors under the orientation used in (6), and γ denotes the adapter scaling convention used by the implementation. In the present configuration, this scaling follows the LoRA convention $\gamma = \alpha/r$. The approximation sign reflects implementation-dependent scaling and numerical approximation in fast PiSSA initialization.

In QPiSSA, PiSSA decomposition is performed before quantization. The residual model W_{res} is saved after full-precision decomposition and then reloaded in 4-bit NF4 form. At initialization, the scaled adapter path approximates the removed principal component:

$$\gamma B_p^{(0)} A_p^{(0)} \approx W_r. \quad (10)$$

During fine-tuning, A_p and B_p remain trainable, whereas the residual base remains frozen and quantized. The effective transformation at step t is therefore

$$h_t = Q(W_{\text{res}})x + \gamma B_p^{(t)} A_p^{(t)} x. \quad (11)$$

This differs from the QLoRA transformation in Equation (7), which quantizes W_0 directly and starts from a standard LoRA adapter initialization. QPiSSA changes both the initialization and the matrix being quantized: the principal low-rank structure is assigned to the adapter path, and only the residual component is quantized. The PiSSA study formulates the corresponding quantization errors as $\|W - \text{NF4}(W)\|_*$ for QLoRA and $\|W_{\text{res}} - \text{NF4}(W_{\text{res}})\|_*$ for QPiSSA, motivating residual quantization as a way to reduce quantization error [9]. Accordingly, the experiment compares QLoRA with the combined QPiSSA adaptation strategy rather than independently separating SVD initialization from residual quantization.

A proxy layer-wise quantization diagnostic is conducted over the LoRA target modules to empirically examine the quantization-error motivation for QPiSSA. For each target projection module, the original pretrained weight matrix used in the QLoRA path and the residual matrix used in the QPiSSA path are quantized using the same controlled 4-bit proxy quantizer. Frobenius reconstruction error, relative Frobenius error, mean squared error, maximum absolute error, and the weight Frobenius norm are then computed. This analysis serves as a diagnostic approximation rather than an exact reproduction of bitsandbytes NF4 reconstruction error. It assesses whether residual matrices are generally easier to quantize, but it is not used as a direct predictor of downstream QA performance.

Fast PiSSA initialization with 16 subspace iterations is used to reduce preprocessing cost. The residual base and initial PiSSA adapter are saved and reused across QPiSSA experiments. During adapter loading, the adapter configuration is adjusted to prevent PiSSA initialization from being re-executed after residual quantization. This ensures the intended sequence: full-precision decomposition, residual quantization, and adapter fine-tuning.

3.9. Training Configuration and Validation Monitoring

All fine-tuned models are trained for 1000 optimizer steps. The per-device batch size is 1, and gradient accumulation is set to 16, giving an effective batch size of $B_{\text{eff}} = 16$. The learning rate is fixed at $\eta = 10^{-4}$ with zero weight decay. A cosine learning rate schedule is used with a warmup ratio of 0.03. Checkpoints are saved every 100 steps, development loss is evaluated every 100 steps, and training loss is logged every 10 steps. The maximum number of retained checkpoints is three.

The main hyperparameters are $r = 16$, $\alpha = 32$, $p_{\text{drop}} = 0.05$, $L_{\text{max}} = 2048$, $T_{\text{max}} = 64$, $B_{\text{eff}} = 16$, and $\eta = 10^{-4}$.

The development set is used to monitor validation loss during training. The validation procedure is identical across all fine-tuned variants and is performed every 100 steps. The final adapter after 1000 steps is used for test evaluation; no best-checkpoint selection based on development loss is applied. For QPiSSA, all reported results are obtained by loading the quantized residual base together with the trained PiSSA adapter. Converted LoRA versions of QPiSSA adapters may be stored for inspection, but the reported QPiSSA results are obtained from the residual base and PiSSA adapter directly.

3.10. Answer Quality Metrics

Answer generation is evaluated on the fixed test set of 2000 oracle-email examples. Each generated answer $\hat{a}_i$ is compared with the gold answer a_i using Exact Match (EM), token-level F1, and ROUGE-L.

Before EM and F1 computation, answers are normalized by lowercasing, removing articles, removing punctuation, and normalizing whitespace. EM is defined as $EM(\hat{a}, a) = \mathbb{I}\{\text{norm}(\hat{a}) = \text{norm}(a)\}$, where $\mathbb{I}\{\cdot\}$ denotes the indicator function, which returns 1 when the condition inside braces is true and 0 otherwise.

Let $\hat{A}$ and A denote the multisets of normalized prediction and gold tokens. Token-level precision, recall, and F1 are

$$P = \frac{|\hat{A} \cap A|}{|\hat{A}|}, R = \frac{|\hat{A} \cap A|}{|A|}, F1 = \frac{2PR}{P + R}. \quad (12)$$

If the normalized prediction is empty, F1 is set to zero unless both the normalized prediction and the normalized reference are empty, in which case F1 is set to one. If $P + R = 0$, F1 is set to zero. ROUGE-L is computed from longest common subsequence overlap and is used as a complementary lexical similarity measure.

3.11. Abstention and Missing-Evidence Evaluation

Missing-evidence reliability is evaluated using a synthetic unanswerable test set. For each selected test question, the correct email is replaced with a different email. The expected behavior is abstention.

Different-email replacement may accidentally produce an answerable context if the replacement email contains direct or indirect evidence for the same answer. We therefore quantify this risk using the leakage diagnostics described in Section 3.17 and report abstention both before and after strict leakage filtering.

A generated response is counted as an abstention if its normalized form contains an abstention expression such as “I don’t know”, “not provided”, “not present”, “not in the email”, “cannot determine”, or “not mentioned”. Let $\text{MatchAbs}(\hat{a}_i, \mathcal{A}_{\text{abs}})$ denote this normalized containment predicate. Abstention accuracy is defined as

$$\text{AbstentionAcc} = \frac{1}{N_{\text{unans}}} \sum_{i=1}^{N_{\text{unans}}} \mathbb{I}\{\text{MatchAbs}(\hat{a}_i, \mathcal{A}_{\text{abs}})\}, \quad (13)$$

where $\mathcal{A}_{\text{abs}}$ denotes the set of accepted abstention expressions. Equation (13) is used to compute abstention accuracy on the synthetic missing-evidence test set. The missing-evidence non-abstention rate is defined as $\text{NonAbstentionRate} = 1 - \text{AbstentionAcc}$. This metric measures failure to abstain under intentionally insufficient evidence. It should not be interpreted as a general hallucination metric.

3.12. Retrieval Diagnostic

Although the main experiments use oracle emails, a BM25 lexical retrieval diagnostic is included to characterize retrievability in the sampled corpus. Unique emails are indexed using lowercased alphanumeric tokens. For each question, the top- k retrieved document identifiers are compared with the gold document identifier. Recall at k is

$$\text{Recall@}k = \frac{1}{N} \sum_{i=1}^N \mathbb{I}\{d_i \in \text{TopK}(q_i)\}. \quad (14)$$

This diagnostic is reported separately from oracle-email QA performance and is not used during model fine-tuning. This separation is important because EnronQA also evaluates retrieval pipelines, and BM25 has been reported as a strong lexical baseline for the benchmark [2].

3.13. BM25-Augmented QA Evaluation

To complement the oracle-email evaluation, the unanswerable-aware QLoRA and QPiSSA adapters are also evaluated with BM25-retrieved contexts. The retrieval corpus is constructed from the unique emails in the sampled train, development, and test snapshots using the same lowercased alphanumeric tokenization as the BM25 diagnostic. For each test question, BM25 retrieves the Top-1 or Top-5 candidate emails. In the Top-1 condition, the single highest-ranked email is supplied as the context. In the Top-5 condition, the five retrieved emails are concatenated in retrieval-rank order and supplied as one context. The same prompt template, maximum sequence length of 2048 tokens, deterministic decoding, and maximum generation length of 64 tokens are used as in the oracle-email evaluation. If the concatenated context exceeds the input budget, the prompt is truncated according to the same preprocessing rule used in the main evaluation. Retrieval hit rate is computed by checking whether the gold email appears in the supplied retrieved context. Answer quality and answerable abstention are then computed using the same metrics as the oracle-email test set.

3.14. Efficiency and Resource Measurements

Efficiency is measured using trainable parameter count, total parameter count, adapter storage size, peak training GPU memory, evaluation GPU memory, and wall-clock training time. All reported experiments and post hoc analyses were conducted on the TRUBA akya-cuda high-performance computing queue using single-GPU processes on NVIDIA V100 16 GB GPUs. The reported wall-clock training times and GPU memory measurements should therefore be interpreted with respect to this execution environment rather than as hardware-independent complexity estimates.

For all fine-tuned adapter methods, the trainable parameter count is $N_{\text{train}} = 40,370,176$, which corresponds to approximately 0.5273% of the Qwen2.5-7B backbone parameter count. Adapter size is measured from the final saved adapter directory, and GPU memory is recorded using CUDA peak-memory statistics. GPU memory is reported as video memory usage (VRAM). Wall-clock training time is reported as a practical runtime measure rather than a hardware-independent complexity metric.

3.15. Paired Evaluation and Reproducibility

All methods are evaluated on the same fixed test examples; therefore, comparisons are paired at the example level. Mean EM, token-level F1, and ROUGE-L are computed over the same 2000 oracle-email test examples. Abstention accuracy and missing-evidence non-abstention rate are computed over the same 2000 missing-evidence examples. This paired protocol supports direct comparison across methods without confounding from different evaluation samples.

Dataset sampling, split construction, synthetic unanswerable construction, and retrieval diagnostics use seed 42 so that all methods are evaluated on the same fixed examples. The primary single-run analysis uses the same fixed data snapshots. To estimate training-seed variability, the four fine-tuned adapter conditions are additionally evaluated across five independently fine-tuned training seeds while keeping the EnronQA subset, development set, test set, prompt format, decoding procedure, and metric implementation fixed. The sampled answerable training set, unanswerable-augmented training set, development set, and test set are saved as JSONL snapshots. Each experimental condition is stored in a separate directory containing the run configuration, run status, training log, checkpoints, final adapter, test predictions, abstention predictions, and summary metrics. The QPiSSA preprocessing directory additionally stores the residual base model and ini-

tial PiSSA adapter. Final tables, plots, convergence curves, prediction summaries, and error-analysis files are exported after training.

The experimental framework is organized into five distinct runs: zero-shot evaluation, answerable-only QLoRA, answerable-only QPiSSA, unanswerable-augmented QLoRA, and unanswerable-augmented QPiSSA. This structural separation ensures traceability by maintaining independent adapters, prediction logs, and configuration files for the answerable-only and unanswerable-augmented variants.

3.16. Statistical Uncertainty and Semantic Evaluation

Because all methods are evaluated on the same fixed test examples, statistical uncertainty is estimated in two complementary ways. First, paired bootstrap resampling is applied at the example level for the primary fixed-seed run. For each metric, 95% confidence intervals are computed from bootstrap resamples of the 2000-example test set. Bootstrap intervals are computed using the percentile method with 10,000 resamples and random seed 42. Pairwise comparisons between QLoRA and QPiSSA are also performed using paired bootstrap differences, where positive values indicate higher scores for QLoRA. The paired resampling unit is the test example, so each resample preserves the correspondence between model outputs for the same question. Two-sided paired bootstrap p values are computed as twice the smaller tail probability of the resampled difference distribution relative to zero. Second, the four fine-tuned adapter conditions are evaluated across five independently fine-tuned training seeds, and aggregate performance is reported as mean $\pm$ standard deviation. The bootstrap analysis estimates fixed-test-set uncertainty for one trained model per condition, whereas the multi-seed analysis estimates variability due to training stochasticity under the same fixed dataset and evaluation protocol.

In addition to Exact Match, Token-F1, and ROUGE-L, semantic answer similarity is evaluated using BERTScore on the 2000 oracle-email answerable test examples. This metric is included because lexical overlap metrics may penalize semantically correct paraphrases. BERTScore is computed with the English language setting using deterministic predictions from the same decoding protocol as the lexical metrics. BERTScore is reported as a supplementary metric and is not used for model selection.

3.17. Leakage Diagnostics for Synthetic Missing-Evidence Examples

The synthetic unanswerable examples are further audited to quantify the possibility that different-email replacement accidentally preserves answer evidence. Each replacement email is checked for exact normalized gold-answer occurrence, answer-token overlap greater than 0.50, question-token overlap greater than 0.80, same-document replacement, same-user/mailbox replacement, and BM25 rank less than or equal to 5 for the question. A strict filtered subset is then formed by excluding examples that trigger any of these leakage indicators. Abstention reliability is reported both on the original synthetic missing-evidence set and on this strict filtered subset.

3.18. Data Ethics and Reproducibility

The experiments are based on EnronQA, a publicly released benchmark derived from the Enron email corpus [2]. No new human participants are recruited, and no additional private data are collected. The study therefore constitutes secondary analysis of an existing public research dataset. The EnronQA paper notes that the benchmark is based on a cleaned release of the Enron corpus and applies filtering to reduce toxic or NSFW content [2]. Because the underlying domain contains real-world email communication, reported results are aggregated at the dataset level, and qualitative examples should be anonymized or omitted unless essential. The dataset is available through its public Hugging Face repository.

Reproducibility artifacts include sampled data snapshots, run configurations, training logs, final adapters, prediction files, abstention evaluation files, paper-ready tables, and figures.

4. Results

4.1. Main Oracle-Email QA Performance

Table 1 reports answer-generation performance on the fixed 2000-example EnronQA oracle-email test set.

The bootstrap intervals show that fine-tuning produces a large improvement over zero-shot prompting across lexical and semantic metrics. Zero-shot prompting obtains Token-F1 of 0.4690 and BERTScore-F1 of 0.9041, whereas the fine-tuned variants reach Token-F1 values around 0.807–0.813 and BERTScore-F1 values around 0.965. The confidence intervals for QLoRA and QPiSSA substantially overlap for EM and Token-F1 in the primary fixed-seed evaluation, indicating that the small point-estimate differences should not be interpreted as strong evidence of superiority for either adapter method.

The reliability contrast is much larger than the QA differences. Answerable-only fine-tuning nearly eliminates abstention under missing evidence: QLoRA reaches abstention accuracy of only 0.0060 and QPiSSA reaches 0.0005. Adding unanswerable-aware supervision restores abstention accuracy to 0.9985 for QLoRA and 0.9980 for QPiSSA while preserving most oracle-email answer quality. This indicates that missing-evidence supervision, rather than the choice between QLoRA and QPiSSA, is the dominant factor controlling abstention reliability in this setting.

Semantic answer similarity follows the same pattern as the lexical metrics, as shown in Table 2. Zero-shot prompting obtains substantially lower BERTScore-F1 than the fine-tuned variants, while all the fine-tuned QLoRA and QPiSSA variants remain closely matched. This indicates that their generated answers are semantically similar even when small lexical differences affect EM, Token-F1, or ROUGE-L.

Zero-shot prompting is therefore best interpreted as a conservative but low-utility baseline. It often avoids unsupported answers under missing evidence, but it does not reliably extract correct answers when the oracle email is supplied. Fine-tuning corrects this answer-generation weakness, whereas unanswerable-aware supervision is needed to preserve abstention behavior.

Table 1. QA, semantic similarity, and missing-evidence reliability results with 95% bootstrap confidence intervals.

Metric	Zero-Shot	QLoRA	QPiSSA	QLoRA + Unans.	QPiSSA + Unans.
EM	0.0020 [0.0005, 0.0040]	0.3400 [0.3190, 0.3610]	0.3330 [0.3120, 0.3535]	0.3370 [0.3165, 0.3580]	0.3360 [0.3155, 0.3570]
Token-F1	0.4690 [0.4611, 0.4769]	0.8133 [0.8038, 0.8226]	0.8070 [0.7972, 0.8165]	0.8084 [0.7987, 0.8179]	0.8100 [0.8005, 0.8195]
ROUGE-L	0.4312 [0.4232, 0.4390]	0.7933 [0.7831, 0.8031]	0.7838 [0.7736, 0.7942]	0.7884 [0.7781, 0.7984]	0.7909 [0.7806, 0.8009]
BERTScore-F1	0.9041	0.9660	0.9649	0.9650	0.9653
Abs. Acc.	0.9995 [0.9985, 1.0000]	0.0060 [0.0030, 0.0095]	0.0005 [0.0000, 0.0015]	0.9985 [0.9965, 1.0000]	0.9980 [0.9960, 0.9995]
Non-Abs. Rate	0.0005 [0.0000, 0.0015]	0.9940 [0.9905, 0.9970]	0.9995 [0.9985, 1.0000]	0.0015 [0.0000, 0.0035]	0.0020 [0.0005, 0.0040]

Table 2. Semantic answer similarity on the 2000-example oracle-email test set. BERTScore is reported as a supplementary metric to complement lexical overlap metrics.

Model	BERTScore-P	BERTScore-R	BERTScore-F1
Zero-shot	0.8756	0.9353	0.9041
QLoRA	0.9667	0.9658	0.9660
QPISSA	0.9663	0.9641	0.9649
QLoRA + unans.	0.9669	0.9636	0.9650
QPISSA + unans.	0.9672	0.9640	0.9653

Table 2 shows that semantic similarity improves substantially after fine-tuning and remains stable after adding unanswerable-aware supervision. This supports the conclusion that the abstention gains are not obtained by substantially degrading answer semantics in the oracle-email setting.

Table 3 shows that most QLoRA–QPISSA differences are small relative to bootstrap uncertainty. In the answerable-only setting, QLoRA has a small advantage in ROUGE-L, but EM and Token-F1 differences are not statistically significant. The abstention difference between answerable-only QLoRA and QPISSA is statistically detectable, but both models fail to abstain in practical terms. In the unanswerable-aware setting, none of the paired QLoRA–QPISSA differences are statistically significant for EM, Token-F1, ROUGE-L, or abstention accuracy. These results indicate that QPISSA is competitive with QLoRA under the matched configuration, but the experiments do not support a claim of QPISSA superiority.

Table 3. Paired bootstrap comparison between QLoRA and QPISSA. Positive differences indicate higher scores for QLoRA.

Setting	Metric	Diff.	95% CI	<i>p</i>
Answerable-only	EM	0.0070	[−0.0090, 0.0235]	0.4042
Answerable-only	Token-F1	0.0063	[−0.0009, 0.0136]	0.0848
Answerable-only	ROUGE-L	0.0095	[0.0020, 0.0171]	0.0150
Answerable-only	Abstention Acc.	0.0055	[0.0020, 0.0090]	0.0008
Unanswerable-aware	EM	0.0010	[−0.0160, 0.0175]	0.9268
Unanswerable-aware	Token-F1	−0.0016	[−0.0091, 0.0059]	0.6832
Unanswerable-aware	ROUGE-L	−0.0025	[−0.0103, 0.0055]	0.5512
Unanswerable-aware	Abstention Acc.	0.0005	[−0.0010, 0.0025]	0.7972

Table 4 shows that the main conclusions are stable across five independently fine-tuned training seeds. In the answerable-only setting, QLoRA has slightly higher mean answer-generation scores than QPISSA, especially for Token-F1. However, both answerable-only models consistently fail to abstain under missing evidence, with mean abstention accuracy below 0.006. In the unanswerable-aware setting, the two adapter methods are nearly indistinguishable across seeds: QLoRA and QPISSA differ by 0.0001 EM, 0.0003 Token-F1, and 0.0001 abstention accuracy in the aggregate means. Both unanswerable-aware variants retain abstention accuracy near 0.998 with very small standard deviations. These multi-seed results reinforce the conclusion that missing-evidence supervision is the dominant driver of reliability, while the adapter choice has only a secondary effect under the matched configuration.

Table 4. Multi-seed aggregate performance on the fixed EnronQA oracle-email and missing-evidence test sets. Values report mean $\pm$ standard deviation across five independently fine-tuned training seeds.

Model	EM	Token-F1	ROUGE-L	Abstention Acc.
QLoRA	0.3392 $\pm$ 0.0069	0.8152 $\pm$ 0.0038	0.7905 $\pm$ 0.0035	0.0052 $\pm$ 0.0014
QPISSA	0.3336 $\pm$ 0.0061	0.8032 $\pm$ 0.0049	0.7865 $\pm$ 0.0036	0.0002 $\pm$ 0.0006
QLoRA + unans.	0.3358 $\pm$ 0.0056	0.8079 $\pm$ 0.0029	0.7893 $\pm$ 0.0034	0.9976 $\pm$ 0.0008
QPISSA + unans.	0.3357 $\pm$ 0.0064	0.8076 $\pm$ 0.0035	0.7916 $\pm$ 0.0031	0.9977 $\pm$ 0.0006

Adding unanswerable examples produces only a small change in oracle-email answer quality. Relative to answerable-only QLoRA, QLoRA with unanswerable-aware supervision decreases EM by 0.0030, Token-F1 by 0.0049, and ROUGE-L by 0.0049. In contrast, QPiSSA with unanswerable-aware supervision slightly improves over answerable-only QPiSSA on all three lexical metrics, reaching EM of 0.3360, Token-F1 of 0.8100, and ROUGE-L of 0.7909. These differences are small, but they are important because they show that missing-evidence supervision does not force a severe trade-off against answer generation when the oracle email is available. In other words, the model can learn abstention behavior for insufficient contexts while largely preserving its ability to answer when sufficient evidence is supplied.

Table 1 highlights two patterns. First, the improvement from zero-shot prompting to adapter tuning is much larger than the differences among the fine-tuned variants, indicating that task adaptation is the dominant factor for oracle-email answer generation. Second, QLoRA and QPiSSA remain close under the matched adapter budget. The multi-seed results in Table 4 further show that QLoRA has a modest advantage in the answerable-only setting, whereas the unanswerable-aware QLoRA and QPiSSA variants are nearly identical in aggregate QA and abstention performance. These results support competitive performance rather than clear superiority of QPiSSA over QLoRA.

4.2. Missing-Evidence Reliability

Table 1 and Figure 2 report missing-evidence reliability. The zero-shot model is highly conservative but has poor answer quality, whereas answerable-only fine-tuning sharply reduces abstention for both QLoRA and QPiSSA. This indicates that answerable-only supervised fine-tuning creates a strong response bias toward generating an answer, even when the supplied context is intentionally insufficient.

Unanswerable-aware supervision reverses this behavior. QLoRA with unanswerable examples reaches abstention accuracy of 0.9985 and a missing-evidence non-abstention rate of 0.0015, reducing the non-abstention rate by 0.9925 compared with answerable-only QLoRA. QPiSSA with unanswerable examples reaches abstention accuracy of 0.9980 and a missing-evidence non-abstention rate of 0.0020, reducing the non-abstention rate by 0.9975 compared with answerable-only QPiSSA. The five-seed results confirm the stability of this effect: mean abstention accuracy is 0.0052 for answerable-only QLoRA and 0.0002 for answerable-only QPiSSA but rises to 0.9976 and 0.9977 for the corresponding unanswerable-aware variants. These large reliability gains, combined with the small changes in answer-quality metrics, indicate that missing-evidence behavior is primarily controlled by the supervision signal rather than by the adapter initialization strategy alone.

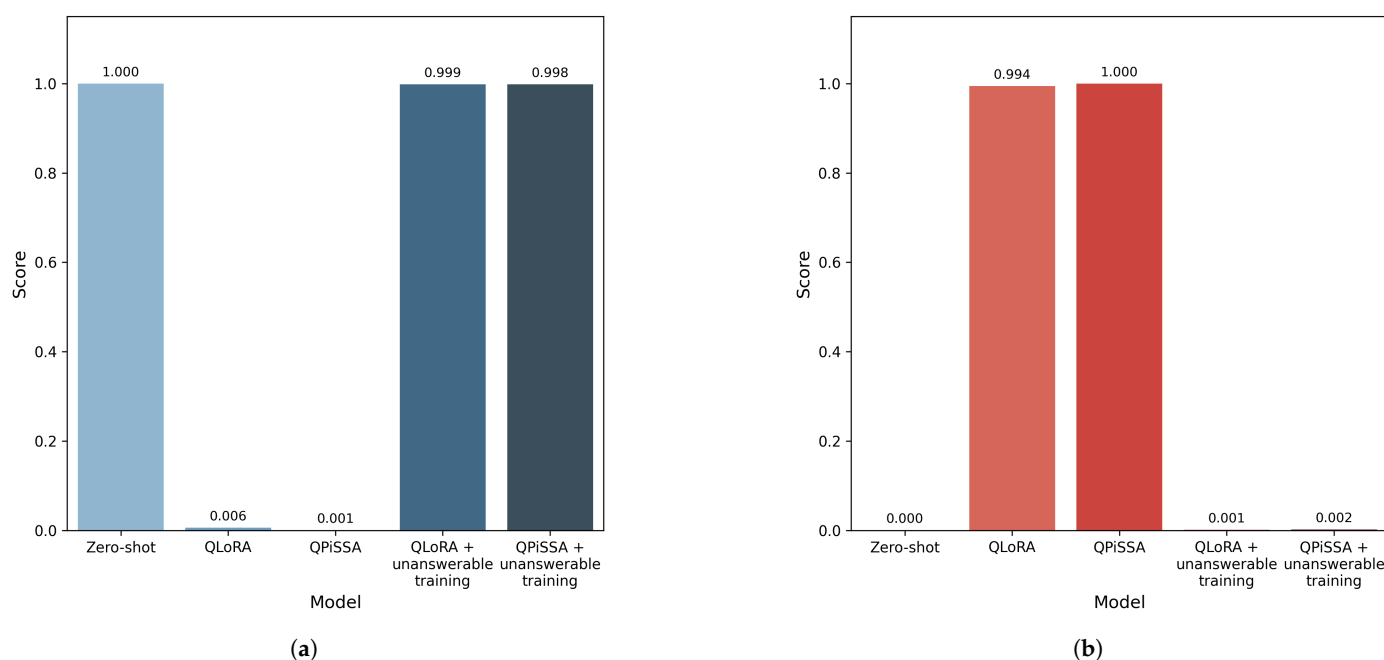


Figure 2. Reliability comparison under missing evidence. (a) Abstention accuracy on the synthetic missing-evidence test set. (b) Missing-evidence non-abstention rate on the same test set. Answerable-only fine-tuning nearly eliminates abstention and produces high missing-evidence non-abstention, whereas unanswerable-aware training restores reliable abstention behavior for both QLoRA and QPiSSA.

Table 5 shows a clear answer-quality and abstention trade-off controlled by ρ . Without unanswerable augmentation, QLoRA achieves the strongest oracle-email answer quality but almost never abstains under missing evidence. Increasing ρ from 0.00 to 0.05 raises abstention accuracy sharply from 0.0060 to 0.9245 with only minimal reductions in EM, Token-F1, and ROUGE-L. Abstention then approaches saturation by $\rho = 0.10$ and reaches 0.9985 at $\rho = 0.15$. Larger values provide little additional abstention benefit but reduce answer-generation quality more noticeably. Thus, $\rho = 0.15$ is selected as a practical knee point that provides near-saturated missing-evidence reliability while preserving most oracle-email answer quality.

Table 5. Sensitivity to the unanswerable augmentation ratio ρ using QLoRA under the same fixed data split and evaluation protocol.

ρ	EM	Token-F1	ROUGE-L	Abstention Acc.
0.00	0.3400	0.8133	0.7933	0.0060
0.05	0.3395	0.8121	0.7918	0.9245
0.10	0.3382	0.8102	0.7896	0.9870
0.15	0.3370	0.8084	0.7884	0.9985
0.25	0.3335	0.8038	0.7831	0.9990
0.40	0.3262	0.7951	0.7749	0.9995

Table 6 shows that the different-email replacement procedure introduces limited direct leakage. Only two of 2000 replacement emails contain the exact normalized gold answer, and 56 examples exceed the answer-token overlap threshold of 0.50. No replacement uses the same document or the same user/mailbox, and only one replacement email appears within the BM25 Top-5 results for the question. After filtering these potential leakage cases, 1943 examples remain, corresponding to 97.15% of the synthetic missing-evidence set.

Table 6. Leakage diagnostics for synthetic missing-evidence examples. The strict subset removes examples triggering the listed leakage indicators.

Indicator	Count	Percent
Exact answer string present	2	0.10
Answer token overlap > 0.50	56	2.80
Question token overlap > 0.80	1	0.05
Same document	0	0.00
Same user/mailbox	0	0.00
Replacement BM25 rank ≤ 5	1	0.05
Strict retained examples	1943	97.15

The strict-filtered results in Table 7 preserve the main reliability pattern. Answerable-only QLoRA and QPiSSA still fail to abstain, with strict abstention accuracies of 0.0046 and 0.0015, respectively. In contrast, both unanswerable-aware variants retain high abstention accuracy of 0.9974 after leakage filtering. This indicates that the observed reliability improvement is not explained by accidental evidence leakage in the synthetic unanswerable set.

Table 7. Missing-evidence reliability before and after strict leakage filtering. Non-Abs. Rate denotes the missing-evidence non-abstention rate.

Model	N_{strict}	Original Abs. Acc.	Strict Abs. Acc.	Original Non-Abs. Rate	Strict Non-Abs. Rate
Zero-shot	1943	0.9995	0.9985	0.0005	0.0015
QLoRA	1943	0.0060	0.0046	0.9940	0.9954
QPiSSA	1943	0.0005	0.0015	0.9995	0.9985
QLoRA + unans.	1943	0.9985	0.9974	0.0015	0.0026
QPiSSA + unans.	1943	0.9980	0.9974	0.0020	0.0026

Figure 2a shows abstention accuracy, and Figure 2b shows the corresponding missing-evidence non-abstention rate. The answerable-only models almost completely fail to abstain, whereas both unanswerable-aware variants recover near-perfect abstention behavior and near-zero non-abstention rates under the definition used in this study. This contrast is central to the private QA setting: a model that performs well on answerable examples can still be unsafe if it confidently answers when evidence is absent. The figure therefore shows that reliability must be evaluated separately from answer-generation quality. The results also suggest that the synthetic unanswerable examples provide an effective supervision signal for learning to abstain under the missing-evidence setting evaluated in this study, even though they are generated through a simple different-email replacement procedure. These results motivate future work on manually audited unanswerable sets, stronger replacement strategies, and retrieval-conditioned abstention evaluation.

4.3. Convergence and Development Loss

Figure 3a shows training loss over 1000 optimization steps, and Figure 3b shows development loss over the same steps. The training curves show that all the fine-tuned models reduce the supervised objective rapidly during the early phase and then stabilize. The development curves show that this optimization also improves held-out objective values rather than merely fitting the training examples. The QLoRA variants reach lower training and development loss earlier than the QPiSSA variants in this run, suggesting that SVD-based residual initialization does not automatically translate into faster supervised convergence in this private email QA setting. Nevertheless, the final downstream differences in Token-F1

and ROUGE-L remain small, so the convergence results should be interpreted as showing competitive behavior rather than clear superiority of one adapter method.

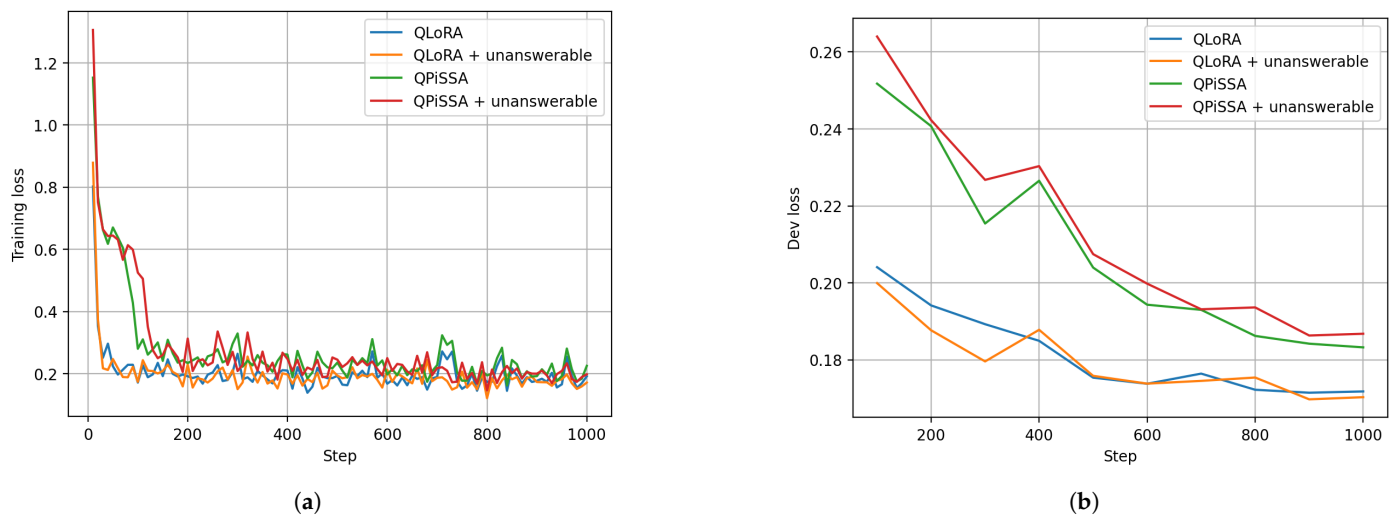


Figure 3. Convergence comparison over 1000 optimization steps. (a) Training loss across fine-tuning steps. (b) Development loss across fine-tuning steps. QLoRA variants reach lower loss earlier in this run, while QPiSSA variants remain competitive in final downstream QA metrics.

4.4. Efficiency and Resource Footprint

Table 8 reports efficiency and resource measurements. All the fine-tuned adapter methods use the same trainable parameter count of 40,370,176, corresponding to 0.5273% of the Qwen2.5-7B backbone. This confirms that the comparisons are matched in trainable capacity and that performance differences are not caused by different numbers of trainable parameters. The final adapter size is approximately 169.2 MB for all the adapter methods, which makes the resulting checkpoints substantially smaller than full-model fine-tuning checkpoints. Peak training memory remains close to 13 GB across all the fine-tuned variants, demonstrating that the experiments are feasible on a single high-memory GPU under a quantized adapter-tuning setup. Evaluation GPU memory is slightly lower for the QPiSSA and unanswerable-aware variants than for the zero-shot and answerable-only QLoRA runs, but the differences are small relative to the overall footprint. These differences may reflect the model-loading path, residual-base representation, adapter state, and runtime memory accounting rather than a fundamental memory advantage of one method. The zero-shot baseline does not require training and therefore has zero trainable parameters and no training time, but it still requires GPU memory during evaluation.

Wall-clock training time varies across runs. QLoRA takes 12,546 s, and QPiSSA takes 11,581 s. The unanswerable variants take longer because they include additional synthetic training cases, with QLoRA + unanswerable training taking 37,594 s and QPiSSA + unanswerable training taking 22,545 s. QPiSSA is faster than QLoRA in both the answerable-only and unanswerable-aware settings in these observed runs, but this result should be treated cautiously because wall-clock time is sensitive to runtime conditions, checkpointing, I/O, and session variability. Therefore, Table 8 should be read mainly as evidence of practical feasibility and resource footprint rather than as a hardware-independent speed benchmark.

Table 8. Efficiency and evaluation-footprint summary. All adapter methods use NF4 4-bit quantization with double quantization, bfloat16 compute, paged AdamW, rank $r = 16$, $\alpha = 32$, and LoRA dropout 0.05. The QLoRA rows count only the incremental adapter artifact because the Hugging Face base model is shared externally. The QPiSSA rows additionally count the saved residual base used for evaluation.

Model	Trainable Params	Peak Train VRAM (GB)	Eval VRAM (GB)	Train Time (s)	Adapter MB	Residual Base MB	Eval. Artifact MB
Zero-shot	0	–	6.584	0	0.000	0.000	0.000
QLoRA	40,370,176	12.964	6.584	12,546	169.199	0.000	169.199
QPiSSA	40,370,176	13.156	5.974	11,581	169.199	14,540.844	14,710.044
QLoRA + unans.	40,370,176	12.968	5.976	37,594	169.199	0.000	169.199
QPiSSA + unans.	40,370,176	13.155	5.975	22,545	169.199	14,540.844	14,710.044

Table 8 distinguishes trainable adapter size from the full evaluation artifact required by each method. All the adapter methods train the same number of parameters, 40,370,176, and have nearly identical incremental adapter sizes of approximately 169 MB. Peak training memory is also similar, remaining close to 13 GB across the fine-tuned variants. However, QPiSSA evaluation requires the residual base saved after PiSSA decomposition; counting this residual base increases the evaluation artifact footprint to approximately 14.7 GB. Thus, QPiSSA is comparable to QLoRA in trainable capacity and GPU memory but differs in stored evaluation footprint.

4.5. Quantization-Error Diagnostic

Table 9 provides empirical support for the residual-quantization motivation of QPiSSA but also shows that the effect is module-dependent. In absolute Frobenius error, the residual matrix has lower reconstruction error than the full matrix for every target module. Aggregated across modules, the residual/full Frobenius-error ratio is approximately 0.95, and the corresponding MSE ratio is approximately 0.87. The largest relative improvement appears in o_proj, where the residual/full relative-error ratio is 0.7822, followed by q_proj with a ratio of 0.9434. However, relative error does not improve uniformly: k_proj has a residual/full relative-error ratio above 1.0. These results support the plausibility of residual quantization reducing reconstruction error for most modules, but they do not by themselves explain the downstream QA results. This is consistent with the main experiments, where QPiSSA remains competitive with QLoRA but does not show a consistent final-performance advantage.

Table 9. Proxy 4-bit quantization-error diagnostic by target module. Full denotes the original pretrained matrix quantized in the QLoRA path, and residual denotes the QPiSSA residual matrix. The ratio is residual relative error divided by full relative error; values below 1 indicate lower relative error for residual quantization.

Module	Full Frob. Err.	Residual Frob. Err.	Full Rel. Err.	Residual Rel. Err.	Residual/Full Rel. Ratio
down_proj	130.118	124.934	0.9886	0.9585	0.9696
gate_proj	124.894	122.238	0.9071	0.9059	0.9987
k_proj	12.766	12.228	0.5249	0.5378	1.0247
o_proj	52.605	39.860	0.9009	0.7046	0.7822
q_proj	41.171	37.410	0.6785	0.6402	0.9434
up_proj	123.986	122.750	0.9265	0.9260	0.9995
v_proj	9.025	8.645	0.3718	0.3662	0.9850

4.6. Diagnostic QPiSSA Factor Analysis

Table 10 provides a diagnostic factor analysis rather than a fully clean causal ablation. Under the reduced 5000-example and 300-step setting, QLoRA standard and combined QPiSSA obtain very similar answer-generation scores. QLoRA has a lower development loss of 0.1777 compared with 0.1933 for QPiSSA, while Token-F1 differs only slightly: 0.8005 for QLoRA and 0.7993 for QPiSSA. This supports the main result that QPiSSA is competitive with QLoRA but does not provide a clear downstream advantage in this setting.

Table 10. Diagnostic QPiSSA factor analysis on a smaller 5000-example training subset with 300 training steps. The third condition is a diagnostic stress test rather than a clean SVD-only ablation: the PiSSA adapter is attached to the full quantized base, which may double-count the principal component.

Condition	SVD Init.	Residual Quant.	Base Type	Dev Loss	EM	Token-F1	ROUGE-L	Abs. Acc.
QLoRA standard	No	No	Full W_0	0.1777	0.3160	0.8005	0.7776	0.0000
QPISSA combined	Yes	Yes	Residual base + adapter	0.1933	0.3080	0.7993	0.7750	0.0020
PiSSA adapter + full quant base	Yes	No	Full W_0 + PiSSA adapter	7.2573	0.0000	0.0010	0.0682	0.0000

The PiSSA-adapter-plus-full-quantized-base stress test fails, with development loss of 7.2573 and near-zero answer quality. This condition is a diagnostic stress test rather than a valid SVD-only ablation because the PiSSA adapter is designed to complement the residual base after removal of the principal component. Attaching the same adapter to the full quantized base may double-count the principal component. This result is consistent with evaluating QPiSSA as the combined residual-base plus PiSSA-adapter construction.

4.7. Retrieval Diagnostic and BM25-Augmented QA

Table 11 reports the BM25 retrieval diagnostic. BM25 is evaluated with $k = 5$ on 1000 queries over 1655 unique documents in the sampled corpus and obtains Recall@5 of 0.976. This high lexical retrieval score suggests that the sampled subset has strong retrievability under sparse lexical matching, likely because many generated questions retain lexical cues from their associated emails. However, this diagnostic is not used during fine-tuning and should be distinguished from the retrieval-augmented generation results in Table 12.

Table 11. BM25 retrieval diagnostic on the sampled EnronQA corpus.

Retriever	k	Queries	Docs	Recall@ k
BM25	5	1000	1655	0.976

Table 12 and Figure 4 extend the oracle-email evaluation to a retrieved-context setting. With BM25 Top-1 retrieval, the gold email is retrieved for most test questions, and answer quality remains close to the oracle setting for both unanswerable-aware adapters. Answerable abstention also remains low under Top-1 retrieval. This suggests that the oracle-email results are reasonably informative when retrieval returns a single highly relevant email.

However, naive BM25 Top-5 concatenation behaves differently. Although the retrieval hit rate increases, answer-generation performance drops sharply and answerable abstention increases substantially. This pattern is consistent with context dilution, truncation, or evidence-selection difficulty when multiple full emails are concatenated. Therefore, the Top-5 result suggests that practical multi-document email QA requires passage selection, reranking, or context compression rather than naive full-email concatenation.

Table 12. Oracle-email and BM25 retrieval-augmented evaluation for the unanswerable-aware models. Retrieval hit rate indicates whether the gold email is present in the supplied retrieved context.

Model	Context	Hit Rate	EM	Token-F1	ROUGE-L	Answerable Abstention
QLoRA + unans.	Oracle email	1.0000	0.3370	0.8084	0.7884	0.0035
QPISSA + unans.	Oracle email	1.0000	0.3360	0.8100	0.7909	0.0020
QLoRA + unans.	BM25 Top-1	0.9140	0.3205	0.7743	0.7553	0.0390
QPISSA + unans.	BM25 Top-1	0.9140	0.3165	0.7757	0.7582	0.0375
QLoRA + unans.	BM25 Top-5	0.9685	0.0375	0.2343	0.2299	0.6010
QPISSA + unans.	BM25 Top-5	0.9685	0.0345	0.2017	0.1979	0.6600

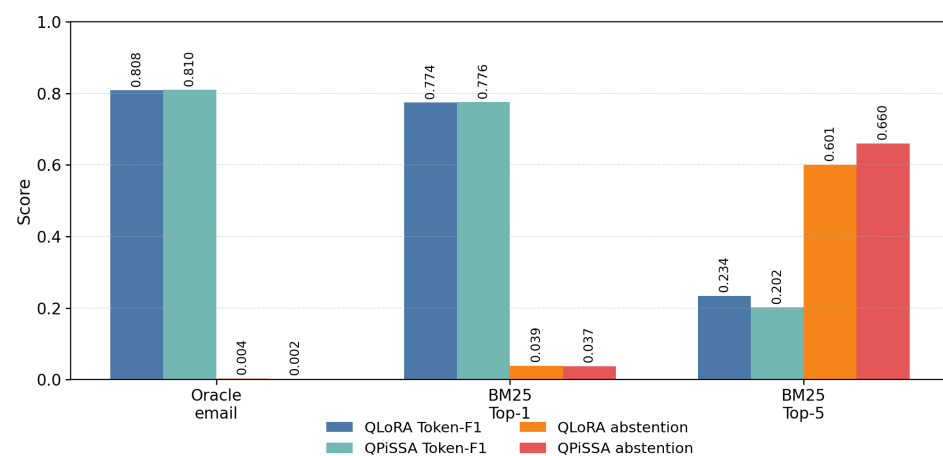


Figure 4. Oracle-email and BM25-retrieved context evaluation for unanswerable-aware QLoRA and QPiSSA. Token-F1 remains close to the oracle-email setting under BM25 Top-1 retrieval, while answerable abstinence stays low. Under naive BM25 Top-5 full-email concatenation, Token-F1 drops sharply and answerable abstinence increases substantially.

4.8. Heuristic Error Analysis

Table 13 shows that zero-shot prompting fails primarily through answerable over-abstention and poor extraction. The zero-shot model abstains on 64.55% of answerable oracle-email examples and has a low-F1 rate of 56.75%. Fine-tuning sharply reduces these failure modes: answerable abstinence falls below 0.35% for all the fine-tuned variants, and low-F1 cases fall to about 9.6–10.4%.

The error profile is similar across the fine-tuned methods. Possible entity-related mismatches remain the largest heuristic category, affecting about 38.5–39.7% of examples. Possible date or number mismatches occur in about 6.55–7.05% of examples. The unanswerable-aware variants show only a small increase in low-F1 and answerable-abstention rates relative to answerable-only fine-tuning: low-F1 increases from 9.60% to 10.30% for QLoRA and from 9.95% to 10.40% for QPiSSA, while answerable abstinence remains below 0.35%. This is consistent with the main result that missing-evidence supervision substantially improves reliability while preserving most answer-generation performance.

Because these categories are generated automatically and are not mutually exclusive, the analysis should be interpreted as a broad diagnostic rather than a definitive human error taxonomy. It nevertheless helps to explain why lexical Exact Match remains modest even when Token-F1, ROUGE-L, and BERTScore indicate strong answer similarity.

Table 13. Automatic heuristic error taxonomy on the 2000-example oracle-email test set. The categories are not mutually exclusive. Exact-match failure is high because many semantically similar answers differ lexically from the gold answer; therefore, this table is intended to characterize broad failure patterns rather than replace semantic or human evaluation.

Model	EM Failure	Low F1	Partial Answer	Possible Entity Error	Possible Date/Number Error	Answerable Abstention
Zero-shot	99.80	56.75	56.25	95.80	11.60	64.55
QLoRA	66.00	9.60	9.35	38.50	6.90	0.10
QPISSA	66.70	9.95	9.70	39.65	6.55	0.05
QLoRA + unans.	66.30	10.30	9.90	39.60	7.05	0.35
QPISSA + unans.	66.40	10.40	10.20	39.55	6.85	0.20

5. Discussion

The results show that quantized adapter tuning substantially improves private email QA performance over zero-shot prompting. This confirms that the base instruction model benefits from task-specific adaptation even when the correct email is supplied as context. The zero-shot model is highly conservative under missing evidence, but this conservatism comes with poor answer quality. Therefore, zero-shot prompting alone is not sufficient for private email QA: it may avoid unsupported answers, but it does not reliably extract the requested information from the supplied email.

The most important finding is the separation between answer quality and reliability. Answerable-only QLoRA and QPiSSA greatly improve EM, Token-F1, and ROUGE-L, but they nearly eliminate abstention under missing evidence. This indicates that supervised fine-tuning on answerable examples alone can teach the model a harmful behavioral prior: when given an email and a question, the model learns to answer rather than to verify whether the answer is actually supported by the context. In private document QA, this behavior is undesirable because unsupported answers may be interpreted as evidence-grounded responses. Consequently, answer-generation metrics alone are insufficient for evaluating private QA systems.

Unanswerable-aware supervision directly addresses this failure mode. Adding synthetic missing-evidence examples restores abstention accuracy to approximately 0.998 for both QLoRA and QPiSSA while preserving most of the oracle-email QA performance. This result suggests that reliable private QA requires explicit supervision for the boundary between answerable and unanswerable contexts. The reliability improvement is much larger than the differences between QLoRA and QPiSSA, indicating that the presence of missing-evidence supervision is the dominant factor for abstention reliability in this setting.

QPISSA is competitive with QLoRA under the matched quantized adapter budget, but the diagnostic and multi-seed analyses indicate that its expected technical advantages do not translate into a consistent downstream advantage in this private email QA setting. Across five training seeds, answerable-only QLoRA obtains higher mean Token-F1 than answerable-only QPiSSA, whereas the unanswerable-aware QLoRA and QPiSSA variants are nearly tied on EM, Token-F1, and abstention accuracy. The proxy quantization-error analysis supports the motivation for residual quantization: residual matrices generally show lower absolute reconstruction error than full matrices, and several target modules have lower relative error. However, the reduction is not uniform across all the modules, and the downstream QA metrics do not show a consistent advantage for QPiSSA over QLoRA.

The diagnostic factor analysis is consistent with the main comparison and the convergence pattern in Figure 3. Under a smaller 5000-example and 300-step setting, QLoRA and combined QPiSSA obtain nearly identical Token-F1 and ROUGE-L, while QLoRA has

slightly lower development loss. A PiSSA-adapter-plus-full-base stress test fails, which indicates that the PiSSA adapter should not be treated as an independent SVD-only component that can be attached to the original full quantized base. Therefore, the present study should be interpreted as an empirical comparison between QLoRA and the combined QPiSSA construction, not as a proof that SVD initialization or residual quantization independently improves private email QA.

The efficiency results show that all the adapter-based methods have the same trainable parameter count and approximately the same adapter size, confirming that the comparison is controlled with respect to trainable capacity. Peak training memory remains close to 13 GB, which supports the practical feasibility of the approach on a single high-memory GPU. Training-time differences should be interpreted more cautiously because they may reflect runtime variability, I/O, checkpointing, and session effects. Nevertheless, the resource profile shows that both QLoRA and QPiSSA can support medium-scale private QA adaptation without full-model fine-tuning.

The BM25 diagnostic shows high lexical retrievability in the sampled corpus, and the retrieval-augmented evaluation further qualifies the oracle-email findings. As shown in Figure 4, BM25 Top-1 retrieval gives a gold-email hit rate of 0.9140 and preserves much of the oracle-email answer quality, suggesting that the oracle-email results are informative when retrieval returns a single highly relevant email. However, naive BM25 Top-5 concatenation increases retrieval hit rate but sharply reduces answer quality and increases answerable abstention. This pattern suggests that practical multi-email RAG requires passage selection, reranking, or context compression rather than simply concatenating full retrieved emails.

These findings should be interpreted with several limitations in mind. First, the experiments use a fixed medium-scale EnronQA subset and a single backbone model rather than the full benchmark or multiple model families. The five-seed evaluation reduces uncertainty due to training stochasticity for the fine-tuned variants, but it does not replace validation on the full dataset or across additional backbones. Second, although BERTScore supplements EM, Token-F1, and ROUGE-L, the study still lacks human evaluation or grounded LLM-as-judge assessment of factual correctness. Third, the missing-evidence non-abstention metric measures failure to abstain under synthetic wrong-email contexts; it does not capture all possible forms of hallucination or unsupported reasoning. Fourth, the synthetic unanswerable construction relies on different-email replacement, and leakage diagnostics can reduce but not fully eliminate the possibility of accidental overlap or indirect evidence. Fifth, wall-clock training time may be affected by runtime conditions and should be interpreted as a practical measurement rather than a theoretical complexity result. Finally, the retrieval-augmented evaluation uses BM25 and simple full-email context construction; more realistic RAG systems should evaluate passage-level retrieval, reranking, compression, and robustness to noisy retrieval.

The QPiSSA-specific diagnostics also require caution. The quantization-error experiment uses a controlled proxy 4-bit quantizer rather than exact bitsandbytes NF4 reconstruction, so it should be interpreted as a diagnostic of relative quantization sensitivity rather than a direct measurement of the deployed quantization kernel. Similarly, the QPiSSA factor analysis is not a clean ablation of SVD initialization and residual quantization. Because PiSSA adapters are designed to complement the residual base, attaching a PiSSA adapter to the full quantized base can double-count the principal component. This condition is therefore reported only as a diagnostic stress test, and the results should not be used to make causal claims about SVD initialization alone.

The error taxonomy is also heuristic. Categories such as possible entity error and possible date or number error are based on automatic string and token patterns, and the

categories are not mutually exclusive. A future study should complement this analysis with human annotation or grounded LLM-based judging.

6. Conclusions

This work evaluated quantized adapter tuning for private email QA using EnronQA. QLoRA and QPiSSA were compared under matched prompt format, model backbone, adapter rank, quantization configuration, and training budget. Fine-tuning improved oracle-email QA quality substantially over zero-shot prompting. However, answerable-only fine-tuning sharply reduced abstention reliability under missing evidence. Adding synthetic unanswerable training examples restored abstention accuracy to approximately 0.998 for both QLoRA and QPiSSA while preserving most QA performance.

Taken together, the analyses indicate that QLoRA and QPiSSA are closely matched after unanswerable-aware supervision is added, although QLoRA shows a modest answerable-only advantage in the five-seed aggregate results. Leakage diagnostics and strict-filtered unanswerable evaluation indicate that the reliability gains are not explained by obvious answer leakage in the synthetic missing-evidence set. BERTScore confirms that the fine-tuned models improve semantic answer similarity, and retrieval-augmented evaluation shows that BM25 Top-1 context preserves much of the oracle-email performance, whereas naive Top-5 concatenation introduces substantial context-selection difficulty. Overall, the main empirical finding is that explicit missing-evidence supervision is essential for reliable abstention in this EnronQA private email setting.

Future work should extend the evaluation to the full EnronQA benchmark, additional model families, human or judge-based groundedness evaluation, stronger passage-level RAG pipelines, and broader multi-seed or multi-subset replication.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data used in this study are publicly available. The experiments are based on the EnronQA benchmark, which is derived from the Enron email corpus, and use the Hugging Face dataset release MichaelR207/enron_qa_0922 with the provided training, development, and test splits [2].

Acknowledgments: The numerical calculations reported in this paper were partially performed at TUBITAK ULAKBIM High Performance and Grid Computing Center (TRUBA resources).

Conflicts of Interest: The author declares no conflicts of interest.

References

1. Wang, S.; Liu, J.; Song, S.; Cheng, J.; Fu, Y.; Guo, P.; Fang, K.; Zhu, Y.; Dou, Z. DomainRAG: A Chinese Benchmark for Evaluating Domain-specific Retrieval-Augmented Generation. *arXiv* **2024**. [CrossRef]
2. Ryan, M.J.; Xu, D.; Nivera, C.; Campos, D. EnronQA: Towards Personalized RAG over Private Documents. *arXiv* **2025**. [CrossRef]
3. Li, J.; Yuan, Y.; Zhang, Z. Enhancing LLM Factual Accuracy with RAG to Counter Hallucinations: A Case Study on Domain-Specific Queries in Private Knowledge-Bases. *arXiv* **2024**. [CrossRef]
4. Wang, Y.; Qu, W.; Zhai, S.; Jiang, Y.; Liu, Z.; Liu, Y.; Dong, Y.; Zhang, J. Silent Leaks: Implicit Knowledge Extraction Attack on RAG Systems through Benign Queries. *arXiv* **2025**. [CrossRef]
5. Hounie, I.; Kanatsoulis, C.; Tandon, A.; Ribeiro, A. LoRTA: Low Rank Tensor Adaptation of Large Language Models. *arXiv* **2025**. [CrossRef]
6. Liao, B.; Herold, C.; Khadivi, S.; Monz, C. ApiQ: Finetuning of 2-Bit Quantized Large Language Model. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*; Al-Onaizan, Y., Bansal, M., Chen, Y.N., Eds.; Association for Computational Linguistics: Miami, FL, USA, 2024; pp. 20996–21020. [CrossRef]

7. Han, Z.; Gao, C.; Liu, J.; Zhang, J.; Zhang, S.Q. Parameter-Efficient Fine-Tuning for Large Models: A Comprehensive Survey. *arXiv* **2024**. [[CrossRef](#)]
8. Dettmers, T.; Pagnoni, A.; Holtzman, A.; Zettlemoyer, L. QLoRA: Efficient finetuning of quantized LLMs. In Proceedings of the 37th International Conference on Neural Information Processing Systems, Red Hook, NY, USA, 10–16 December 2023.
9. Meng, F.; Wang, Z.; Zhang, M. PiSSA: Principal Singular Values and Singular Vectors Adaptation of Large Language Models. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 121038–121072. [[CrossRef](#)]
10. Wang, Y.; Meng, F.; Zhang, X.; Jiang, F.; Tang, P.; Zhang, M. HD-PiSSA: High-Rank Distributed Orthogonal Adaptation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*; Christodoulopoulos, C., Chakraborty, T., Rose, C., Peng, V., Eds.; Association for Computational Linguistics: Suzhou, China, 2025; pp. 6515–6528. [[CrossRef](#)]
11. Hu, E.J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. LoRA: Low-Rank Adaptation of Large Language Models. *arXiv* **2021**. [[CrossRef](#)]
12. Biderman, D.; Portes, J.; Ortiz, J.J.G.; Paul, M.; Greengard, P.; Jennings, C.; King, D.; Havens, S.; Chiley, V.; Frankle, J.; et al. LoRA Learns Less and Forgets Less. *arXiv* **2024**. [[CrossRef](#)]
13. Guo, H.; Greengard, P.; Xing, E.; Kim, Y. LQ-LoRA: Low-rank plus Quantized Matrix Decomposition for Efficient Language Model Finetuning. In Proceedings of the Twelfth International Conference on Learning Representations, Vienna, Austria, 7–11 May 2024.
14. Lawton, N.; Padmakumar, A.; Gaspers, J.; FitzGerald, J.; Kumar, A.; Steeg, G.V.; Galstyan, A. QuAILoRA: Quantization-Aware Initialization for LoRA. *arXiv* **2024**. [[CrossRef](#)]
15. Jeon, H.; Kim, Y.; Kim, J.J. L4Q: Parameter Efficient Quantization-Aware Fine-Tuning on Large Language Models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*; Che, W., Nabende, J., Shutova, E., Pilehvar, M.T., Eds.; Association for Computational Linguistics: Vienna, Austria, 2025; pp. 2002–2024. [[CrossRef](#)]
16. Yang, Y.; Zhou, J.; Wong, N.; Zhang, Z. LoRETTA: Low-Rank Economic Tensor-Train Adaptation for Ultra-Low-Parameter Fine-Tuning of Large Language Models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*; Duh, K., Gomez, H., Bethard, S., Eds.; Association for Computational Linguistics: Mexico City, Mexico, 2024; pp. 3161–3176. [[CrossRef](#)]
17. Hu, X.; Cheng, X.; Liu, P.; Liu, W.; Luan, J.; Wang, B.; Liu, Y. DoTA: Weight-Decomposed Tensor Adaptation for Large Language Models. In *Advances in Knowledge Discovery and Data Mining*; Springer Nature: Singapore, 2025; pp. 16–27. [[CrossRef](#)]
18. Chong, Y.H.; Qu, P. Singular Value Decomposition on Kronecker Adaptation for Large Language Model. *arXiv* **2025**. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.