Abstract:

In the domain of official statistics, public institutions face the dual challenge of managing increasing volumes of user inquiries while maintaining rigorous standards of factual accuracy and formal communication. The Central Bank of the Republic of Türkiye (CBRT) manages the Electronic Data Delivery System (EVDS) which is a critical platform serving researchers, financial analysts, academia and media. To address the inefficiencies of manual support processes, this study investigates the automation of email responses using Large Language Models (LLMs), specifically focusing on the trade-off between semantic precision and institutional voice.

We conducted a comparative analysis of two distinct generative AI architectures implemented within the CBRT infrastructure:

1. Retrieval-Augmented Generation (RAG): This approach integrates a generative model with a dynamic knowledge base derived from historical support logs and documentation, prioritizing factual grounding and up-to-date information.

2. Low-Rank Adaptation (LoRA): This parameter-efficient fine-tuning method adapts the open-source Gemma-3 model to the specific domain by aiming to internalize the stylistic norms and phrasing of the EVDS support team without extensive retraining.

The performance of both systems was assessed using a hybrid framework combining automated metrics (BERTScore and ROUGE) and human evaluation by domain experts across dimensions of accuracy, usefulness and trustworthiness.

Results reveal significant insights in terms of accuracy and style balance. The RAG architecture demonstrated superior performance in semantic reliability deriving from high BERTScore and factual consistency, significantly reducing the risk of hallucinations by anchoring responses in retrieved context. While LoRA proved effective in generating linguistically fluent drafts that aligned to the institution's communication style, it was limited by the temporal rigidity of its training data and lacked the real-time precision of the retrieval-based system.

The study concludes that while fine-tuning captures the form of institutional tone, retrieval mechanisms are essential for the substance required in official statistics. We propose a hybrid framework that leverages LoRA for stylistic alignment and RAG for factual grounding as the optimal strategy for deploying generative AI in sensitive public sector environments.

Keywords:

Automation, Central-banking, Data Dissemination, Data Hub, Generative AI, Large Language Models, LLMs, Retrieval-augmented-generation, Fine-tuning, Low-Rank Adaption

1. Introduction

Organizations providing large-scale data services face an increasing volume of inquiries from end-users. The Central Bank of the Republic of Türkiye (CBRT) manages the Electronic Data Dissemination System (EVDS), which serves as a critical platform by delivering economic and financial data to researchers, analysts and the general public. Currently, user inquiries ranging from query construction to API troubleshooting are processed manually by the EVDS support team. While ensuring quality control, this manual response system creates inefficiencies and resource constraints, particularly during periods of high demand, leading to slower response times and user dissatisfaction.

Early approaches to automating support relied on rule or template-driven systems that lacked flexibility in handling complex requests. Transformer-based models, beginning with BERT (Devlin et al., 2019) and continuing with GPT-style autoregressive models (Brown et al., 2020), have renovated this position by enabling systems to generate fluent and contextually relevant responses without explicit templates. However, general-purpose large language models (LLMs) often lack domain-specific accuracy, present risks of producing hallucinations and fail to deliver the expected formal communication style required in specialized settings such as central banking.

To address the limitations of generative models, Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) has been widely studied as a hybrid architecture. It improves factual consistency and reduces hallucinations by gathering outputs from external sources and concatenating retrieved texts with the query. Separately, parameter-efficient fine-tuning methods such as LoRA (Hu et al., 2022) allow large models to adapt to new domains without retraining all parameters. For central banking applications, LoRA is particularly attractive because it allows a model to internalize specialized institutional styles while remaining lightweight enough to deploy securely on internal infrastructure.

Central banks and financial regulators have already started experimenting with LLMs in supervisory contexts, emphasizing that AI tools must rigorously balance innovation with data governance. Notable examples include the Bank for International Settlements (BIS) Innovation Hub's Project Gaia and Project AISE (Bank for International Settlements, 2025). Aligning with this broader institutional context, this project investigates two complementary approaches for automating EVDS support: integrating document retrieval with generative modeling (RAG) and fine-tuning a pre-trained open-source model to institutional norms (LoRA). By conducting a comparative evaluation using human judgments and automated metrics such as BERTScore (Zhang et al., 2019), this study contributes to the practical debate on how public institutions can responsibly leverage generative AI in user-facing services.

2. Methodology

The methodology of this project aligns with the key steps of a research process, from data collection to model evaluation, aiming to ensure both reproducibility and robustness in a real institutional context.

2.1. Dataset Preparation and Preprocessing

Historical email correspondence provided by the EVDS support team constitutes the keystone of this project. The dataset includes a wide range of inquiries, such as data access and API usage, paired with their corresponding human-written responses. To ensure compliance with privacy constraints, incoming messages were sanitized by removing residual HTML brackets and collapsing whitespace. URLs and email

addresses were masked using canonical placeholders such as <mail> to preserve domain information while mitigating the leakage of personally identifiable data.

2.2. Text Segmentation and Vectorization

A lightweight regex splitter was utilized to segment messages into sentences, filtering out short, semantically unhelpful fragments like greetings. Each sentence was then converted into overlapping chunks using a sliding window to maximize semantic coverage, effectively bridging the length mismatch between full emails and the short context preferred by dense retrievers (Gupta et al., 2024). Each chunk was embedded utilizing a multilingual MiniLM model via llama.cpp. These dense vectors, alongside their metadata, were persisted in a Hierarchical Navigable Small World (HNSW) index and a PostgreSQL database, allowing for memory-efficient approximate nearest-neighbor search with logarithmic complexity.

3. Retrieval-Augmented Generation (RAG)

The RAG system embodies the core principle of grounded generation by leveraging the retrieved EVDS-specific context to enhance factual accuracy and institutional consistency.

3.1. Context Retrieval

The knowledge base was composed entirely of anonymized historical answers. When a new user query arrived, it was subjected to the identical normalization and chunking pipeline. The system queried the HNSW index to obtain the top-k nearest neighbors below a configurable distance threshold. Retrieved IDs were mapped back to their parent records, producing a compact set of semantically similar examples that accurately captured prior institutional practice for related issues.

3.2. Prompt Construction and Generation

A structured Turkish prompt was constructed using a few-shot approach. This prompt integrated the retrieved examples alongside specific instructions dictating tone, formatting constraints and a strict refusal policy. The composed prompt was subsequently sent to an internal endpoint utilizing the Gemma-3 model. This architecture enabled the model to generate a draft email response grounded in factual data, thereby providing traceability and significantly reducing the risk of hallucinations.

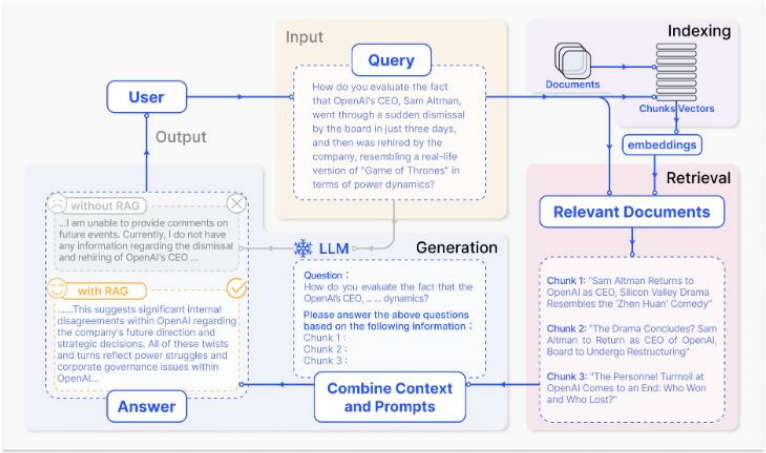

Figure 1. A basic flow of the RAG approach with the components. Reproduced from Gao et al. (2023).

4. Low-Rank Adaptation (LoRA)

While retrieval-based methods excelled at factual grounding, fine-tuning via LoRA provided a parameter-efficient approach to internalize the specific stylistic and communicative norms of the EVDS support staff.

4.1. Training Data and Parameter-Efficient Adaptation

The training corpus was serialized into an instruction-completion format, where the user's query functioned as the instruction and the staff response as the target. Tokenization was handled using the 12-billion parameter Gemma-3 model with a maximum sequence length of 1024 tokens and labels were appropriately masked to prevent prompt leakage into the loss calculation. A Quantized Low-Rank Adaptation (QLoRA) strategy was utilized to load the base model in 4-bit NF4 precision, drastically reducing GPU memory requirements. Low-rank adapter matrices (rank 16, scaling factor 32, dropout 0.05) were injected into the attention and feed-forward projection layers, allowing the model to adapt while keeping the vast majority of its pretrained parameters frozen.

4.2. Cross-Validation, Merging and Export

To test the model's ability to generalize and avoid overfitting, training employed a 5-fold cross-validation setup. The model was trained using the AdamW optimizer with bfloat16 precision, utilizing a learning rate of 2e-4 with cosine decay, a warm-up ratio of 0.03 and a batch size of 16 via gradient accumulation for two epochs per fold. Following cross-validation, adapters from the best-performing fold were merged into the base model. The finalized, consolidated model was then exported to the GPT-Generated Unified Format (GGUF), yielding a self-contained artifact that ensured seamless integration and uniformity across the corporate deployment infrastructure.

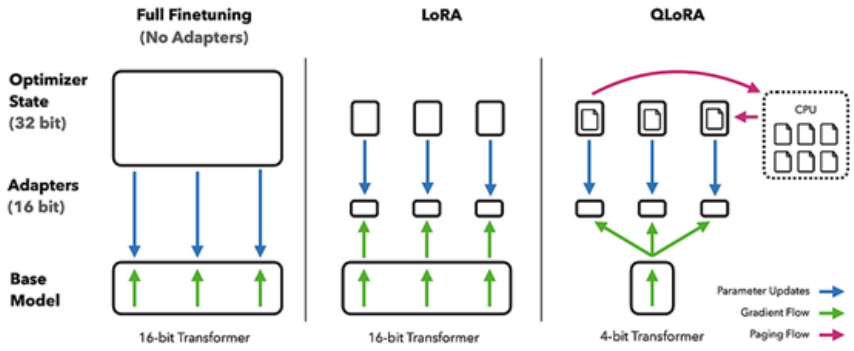

Figure 2. Comparison of different finetuning methods and their memory requirements. QLoRA improves over LoRA by quantizing the transformer. Reproduced from Dettmers et al. (2023).

5. Evaluation Framework

The evaluation of the two proposed architectures was conducted using a mixed framework that combined automated metrics with human assessment. This dual approach ensured both scalability and statistical consistency while capturing subtle nuances—such as institutional tone, readability and user trust—that are inherently difficult to measure using solely automated quantitative scores. The primary objective was to fairly compare the retrieval-augmented and fine-tuned systems under strictly consistent experimental conditions.

5.1. Experimental Setup and Data Integrity

To ensure the integrity of the evaluation and prevent any potential data leakage or bias, all assessments were performed on a dedicated test set comprised of previously unseen email and response pairs. For the RAG architecture, test responses were deliberately excluded from the vector knowledge base to avoid the trivial retrieval of exact matches, ensuring that the model's generation capabilities were genuinely tested. Similarly, the LoRA model was fine-tuned exclusively on the training partition of the dataset and was never exposed to the test set during the training phase. This rigorous separation guaranteed that the performance measurements reflected the models' true generalization capabilities rather than mere memorization of historical data.

5.2. Automated Evaluation Metrics

The automated evaluation phase utilized BERTScore and ROUGE-L, which provided complementary analytical perspectives on model performance. BERTScore (F1) was employed to measure semantic similarity within the embedding space. Because it is highly robust to paraphrasing, BERTScore served as an ideal metric to assess whether the model-generated drafts accurately captured the intended meaning of the original EVDS staff replies. Conversely, ROUGE-L was utilized to evaluate the longest common subsequence overlap, effectively rewarding lexical closeness and sequence consistency. The combination of these two metrics facilitated a clear distinction between the semantic adequacy prioritized by the RAG approach and the stylistic alignment emphasized by the LoRA fine-tuning.

5.3. Human Assessment Protocol

Recognizing that automated metrics are often insufficient for capturing the semantic subtleties required in sensitive institutional communications, a comprehensive human evaluation was conducted with domain experts from the EVDS team. To systematically minimize selection and confirmation bias, the human raters were neither the original authors of the reference responses nor were they informed which system had generated the text. For each test case, the model outputs were strictly anonymized, randomized in order and presented alongside the corresponding user query.

The raters were tasked with assessing the quality of the generated answers using a structured format across four major categories: Accuracy and Content Quality (encompassing truthfulness, completeness, originality and consistency); Usefulness and Practicality (covering relevance, usefulness, actionability and conciseness); Language and Expression Quality (including clarity, style, grammatical fluency and formatting); and User Experience (focusing on trustworthiness, readability and length appropriateness). Each criterion was meticulously rated on a 5-point Likert scale, ranging from 1 (representing a factually incorrect or completely unhelpful response) to 5 (denoting a fully correct, actionable and stylistically flawless institutional reply).

6. Results

The comparative evaluation of the systems yielded insights into both the quantitative alignment and qualitative appropriateness of the generated responses. The results highlight the complementary capabilities of the two architectures.

6.1. Automated Evaluation

The initial evaluation stage utilized automatic metrics to quantify semantic and lexical similarity between the model-generated drafts and the official EVDS staff answers. The results are detailed in Table 1.

Model	***BERTScore (F1)***	***ROUGE-L***
RAG	0.859	0.115
LoRA	0.846	0.080

Table 1. Automated evaluation results.

The automated metrics indicated that the RAG model achieved superior semantic similarity, recording a BERTScore (F1) of 0.859 compared to LoRA's 0.846. This suggests that the retrieval grounding mechanism enabled RAG to generate responses more semantically aligned with the original human-written replies. Furthermore, RAG significantly outperformed LoRA in ROUGE-L scoring (0.115 versus 0.080), demonstrating that RAG's outputs possessed a greater lexical overlap with the reference answers. Conversely, LoRA's reliance on learned fine-tuning patterns resulted in stylistically consistent but more paraphrastic outputs, thereby reducing direct lexical overlap.

6.2. Human Evaluation

To complement the automated metrics, a comprehensive human evaluation was conducted across four main dimensions. Table 2 summarizes the average scores assessed on a 1-5 Likert scale.

	RAG	***LoRA***
Accuracy & Content Quality		
Truthfulness	3.47	3.28
Completeness	3.55	3.53
Originality	3.43	3.43
Consistency	3.73	3.55
Usefulness & Practicality		
Relevance	3.62	3.53
Usefulness	3.38	3.23
Actionability	3.42	3.27
Conciseness	3.62	3.35
Language and Expression Quality		
Clarity	3.87	3.75
Style	4.07	3.93
Grammar/Fluency	4.13	3.90
Formatting	3.98	3.65
User Experience		
Trustworthiness	3.55	3.52
Readability	4.02	3.88
Length Appropriateness	3.95	3.75

Table 2. Human evaluation results (mean scores on 1-5 Likert scale).

Across all four categories, RAG consistently outperformed LoRA, albeit by modest margins in certain sub-criteria. In the *Accuracy & Content Quality* dimension, RAG scored higher on truthfulness and consistency, suggesting that grounding in the indexed knowledge base yielded more reliable outputs. However, LoRA performed nearly identically in completeness and originality, indicating that fine-tuning effectively captured the structural coverage of staff replies.

The most pronounced performance gap emerged in *Language and Expression Quality*. RAG obtained the highest scores in clarity (3.87), style (4.07) and grammar/fluency (4.13). While LoRA was explicitly fine-tuned to internalize institutional style, the retrieval mechanism in RAG appeared to reinforce the use of precise, institutionally appropriate language, thereby narrowing LoRA's anticipated stylistic advantage. Ultimately, both models achieved strong results in *User Experience*, confirming their practical viability, though RAG maintained a slight edge in trustworthiness and readability.

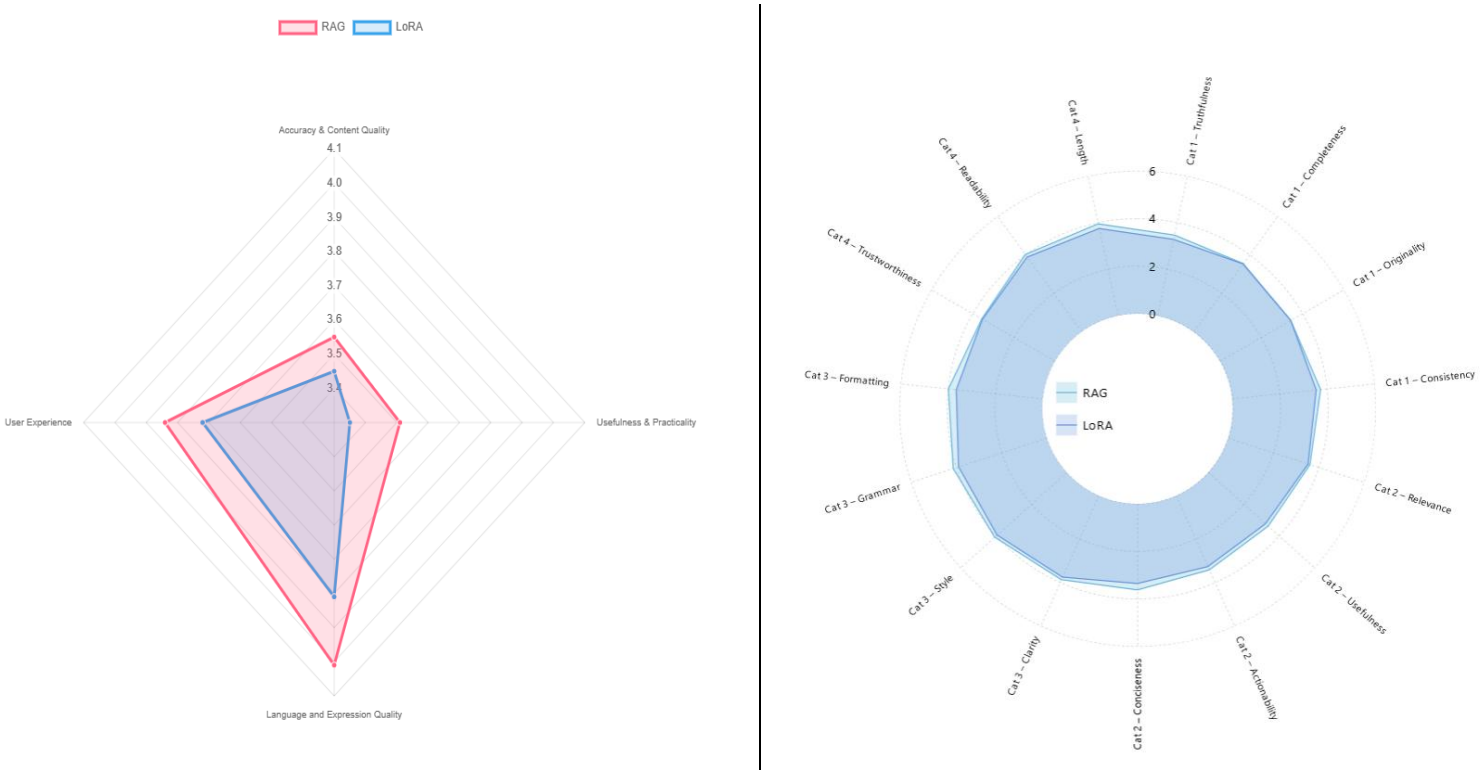

Figure 3. Radar chart visualization of human evaluation results. The left panel summarizes average scores across the four main categories (Accuracy & Content Quality, Usefulness & Practicality, Language & Expression Quality and User Experience), while the right panel presents detailed scores for all individual criteria. The plots illustrate that RAG consistently outperforms LoRA across most dimensions, with the largest margins observed in language quality and user experience. LoRA remains competitive on completeness and originality, highlighting the complementary nature of the two approaches.

7. Discussion and Conclusion

The comparative evaluation of the Retrieval-Augmented Generation (RAG) and Low-Rank Adaptation (LoRA) architectures highlights the distinct, yet complementary, strengths and limitations inherent in each approach when applied to institutional support scenarios. Automated evaluation results demonstrated a clear advantage for RAG, establishing that direct access to an indexed knowledge base naturally increases both semantic reliability and lexical overlap with historical phrasing. Human evaluations reinforced this pattern, noting that the stabilizing effect of retrieval consistently yielded text that evaluators found clearer, more trustworthy and more professionally formatted.

Despite RAG's quantitative and qualitative advantages, the differences between the two systems are not absolute. LoRA performed nearly on par with RAG concerning completeness and originality, indicating that parameter-efficient fine-tuning successfully captures the general structure and coverage of staff replies. Furthermore, LoRA's reliance on historical training data allows it to operate fully offline, presenting a distinct advantage under the strict data governance and privacy constraints often encountered in public institutions. However, the architectures exhibit contrasting limitations. RAG is highly sensitive to the quality, coverage and timeliness of its indexed material, whereas LoRA is constrained by the temporal rigidity of its training corpus, unable to address newly introduced datasets or policy shifts without periodic retraining. Table 3 details this comparative feature analysis.

Feature	***RAG***	***LoRA***
Knowledge & Updates	Retrieval index easily updated, no retraining required	Requires retraining to incorporate new knowledge
Preprocessing Complexity	Minimal preprocessing (chunking, indexing)	Requires curated, domain-specific training data
Adaptability	Adapts quickly to new queries if context exists	Strong domain style alignment, but retraining needed for new domains
Transparency	High - answers traceable to sources	Low - black-box model, harder to interpret
Resources	Needs ANN index, moderate compute at inference	High compute and storage required during fine-tuning
Scalability	Scales well with large corpora by expanding index	Scalability tied to retraining and GPU availability
Latency	Retrieval adds slight latency	Faster response after fine-tuning, no retrieval overhead
Robustness	Less prone to hallucinations (grounded in context)	May hallucinate when input is out-of-domain

Table 3. Comparative feature analysis of RAG and LoRA. The table highlights trade-offs across knowledge & updates, preprocessing complexity, adaptability, transparency, resources, scalability, latency and robustness.

Acknowledging the limitations of this study is crucial for contextualizing the results. The dataset was of modest size and restricted to historical EVDS support emails, potentially limiting broader generalizability. Additionally, the evaluation was constrained to a single base model family (Gemma-3) and expanding the human rater pool could further strengthen confidence in the qualitative metrics.

In conclusion, this study demonstrates that generative AI techniques can meaningfully support institutional staff by producing draft responses that are both factually accurate and stylistically appropriate. The findings strongly suggest that a hybrid design—where LoRA enforces institutional tone and structural communication norms while RAG supplies up-to-date factual grounding—offers the most promising pathway for deployment. Future work will focus on exploring hybrid pipelines that combine RAG and

LoRA at inference time, incorporating continual learning strategies to mitigate temporal limitations and expanding the evaluation framework to include cross-lingual datasets to further enhance central bank communication workflows.

8. References

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers) (pp. 4171-4186).

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in neural information processing systems, 33, 1877-1901.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in neural information processing systems, 33, 9459-9474.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... & Chen, W. (2022). Lora: Low-rank adaptation of large language models. ICLR, 1(2), 3.

Bank for International Settlements. (2025, August 25). Project Gaia: Enabling climate risk analysis using generative AI. https://www.bis.org/publ/othp84.htm

Bank for International Settlements. (2025, August 25). Project AISE: enhancing financial supervision with AI-powered tools. https://www.bis.org/about/bisih/topics/suptech_regtech/aise.htm

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2019). Bertscore: Evaluating text generation with bert. arXiv preprint arXiv:1904.09675.

Gupta, S., Ranjan, R., & Singh, S. N. (2024). A comprehensive survey of retrieval-augmented generation (rag): Evolution, current landscape and future directions. arXiv preprint arXiv:2410.12837.

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., ... & Wang, H. (2023). Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, *2*(1).

Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). Qlora: Efficient finetuning of quantized llms. Advances in neural information processing systems, 36, 10088-10115.